Speech Prosody 2024
2-5 July 2024, Leiden, The Netherlands

What makes a conversational agent sound trustworthy? Exploring the role of acoustic-prosodic factors

Yuwen Yu[1], *Sarah Ita Levitan*[1,2]

[1]Department of Computer Science, The Graduate Center, CUNY, USA
[2]Department of Computer Science, Hunter College, CUNY, USA
yyu4@gradcenter.cuny.edu, sarah.levitan@hunter.cuny.edu

Abstract

With advances in machine learning and speech technologies, conversational agents are becoming increasingly capable of engaging in human-like conversations. However, trust is crucial for effective communication and collaboration, and understanding the signals of trustworthy speech is essential for successful interactions. While researchers across disciplines have sought to discover the signals of trustworthy speech, mostly in human speech, in this paper, we explore the human perception of trustworthy synthesized speech. We present the results of a large-scale crowdsourced perception study, designed to investigate the acoustic-prosodic properties of trustworthy synthesized speech. Highly controlled parameters are manipulated to test the effects of acoustic-prosodic features including pitch, intensity, and speaking rate. We also extend the work to examine individual differences in the perception and production of trustworthy in speech. To evaluate trust perception in contexts that require vulnerability and trust, a real-world application of emotional support dialogues is used. The findings of this work contribute valuable insights to improve the perceived trustworthiness of conversational agents.

Index Terms: speech perception, speaking style, trustworthiness, human-computer interaction, computational paralinguistics

1. Introduction

Identifying verbal indicators of trustworthiness in conversational agents is an important problem with many far-reaching implications for human-computer interaction. We are rapidly approaching a future in which conversational agents, including chatbots, virtual assistants, and robots with dialogue capabilities, are becoming increasingly integrated into our daily lives. These technologies have gained significant traction across various domains such as customer service, healthcare, and education. As these agents aim to simulate human-like conversation, a crucial factor that profoundly influences user engagement and satisfaction is trust. Users must feel confident in the reliability, credibility, and trustworthiness of the information and responses provided by these agents. Researchers from diverse disciplines have sought to identify specific signals of trust and have examined nonverbal and verbal cues to trustworthiness, mostly in human-human interactions. However, there is little work that examines verbal indicators of trustworthiness in conversational agents. Further, we currently have a limited understanding of individual differences in human trust perception.

In this work, we address these gaps and conduct a large-scale crowdsourced perception study to understand how humans perceive the trustworthiness of synthesized speech. In the study, participants rate the trustworthiness of several samples of synthesized speech with varied acoustic-prosodic parameters. Acoustic-prosodic features such as pitch, speaking rate, and intensity, can convey emotional expressiveness, emphasis, and engagement, and can affect user perception of agent trustworthiness. Therefore, we examine the effects of these features on user trust perception. To evaluate trust perception in contexts that require vulnerability and trust, a real-world application of emotional support dialogues is used. In addition, demographic factors, such as speaker or listener gender may contribute to perceived credibility and reliability. We also explore these demographic factors in this paper. By uncovering the specific influences of acoustic-prosodic as well as demographic factors, we can advance our understanding of how trust is formed and maintained in human-machine interactions. Furthermore, this knowledge can directly inform the design and development of conversational agents that effectively build and maintain trust with users, ultimately improving user engagement and satisfaction.

1.1. Related work

There have been many studies examining speech factors that affect trust in human-human communication. [1, 2] conducted extensive studies of trust in human-human communication in the context of perceived deception in interview dialogues. They identified several verbal and nonverbal cues to trustworthiness. For example, faster speaking rate and increased pitch and intensity values were all associated with greater trust. They also found that speaker traits such as native language, gender, and personality traits affect trust perception. Although we understand a great deal about speech factors that affect trust in human-human communication [1, 2, 3, 4, 5], we have limited knowledge of how the speech and lexical patterns of conversational agents affect trust in human-computer communication. There have been few studies on the effects of synthesized speech prosody on trust perception. One body of work has identified a positive relationship between acoustic-prosodic entrainment and trust, in both human-human and human-computer dialogues. A study of entrainment and trust in human-computer dialogues across three languages showed a positive association between entrainment and trust in English but not in Slovak or Spanish [6]. [7] implemented a prosodically entraining spoken dialogue system and found that acoustic-prosodic entrainment is associated with user trust, measured by how often the user follows the advice of an avatar. Some studies have examined the effect of human vs. synthesized voices of conversational agents on trust [8, 9, 10] with inconsistent results; some found that human voices were trusted more, while others found that the voice interacted with other system features such as appearance and behavior to affect trust perception. [11] examined how the prosodic characteristics of a virtual player in an investment

10.21437/SpeechProsody.2024-250

game affect partner trust in the context of investment decisions. They found that the accent, mean pitch, and articulation of the virtual player all influence partner investment decisions.

This work expands on these prior studies and focuses on analyzing the acoustic-prosodic characteristics that are associated with trustworthy synthesized speech.

2. Data Collection

2.1. Speech Stimuli

We prepare synthesized speech samples to be used in the perception study. For the purposes of this study, we use sentences from the Emotional Support Conversations Dataset [12], a dataset of 1,300 conversations centered between a help-seeker and a supporter. The conversations center around 10 topic problems (relationship, employment, etc.) and turns are annotated by strategies for providing emotional support, including questions, self-disclosures, and suggestions. This corpus is chosen because emotional support conversations are a potential application where a conversational agent can be used to provide support. They require user trust and vulnerability from the human help-seeker in order to fully utilize and benefit from the interaction. Specifically, we select sentences that are labeled as supporter questions, as questions require the listener to trust the speaker in order to feel comfortable sharing personal information. Some examples of text utterances are "I am here to listen, how are you feeling?", "How is your life at the moment? Do you want to talk about anything?".

To understand how acoustic-prosodic aspects of speech relate to trust, we focus on 3 fundamental aspects of prosody: intensity, pitch, and speaking rate. These features have been identified in previous work to affect trust perception and they are straightforward to manipulate and measure objectively. We synthesize audio stimuli in various prosodic styles using the Amazon Polly Neural Text-to-Speech (TTS) system. The motivation for using this system is that it is a commercial state-of-the-art TTS which is integrated with multiple dialogue systems and conversational robots, making it easy to use the findings of this research in an existing dialogue system. It supports voice alterations using Speech Synthesis Markup Language (SSML) [13], an XML-based markup language which provides a platform-independent interface standard for controlling aspects of synthesized speech. Speech samples are synthesized using one of three settings for each of the prosodic features (pitch, intensity, and speaking rate): low, medium, or high. High pitch ranges from 212 to 117 Hz, medium pitch ranges from 170 to 95 Hz, low pitch ranges from 160 to 82 Hz. Fast speaking rate ranges from 383 to 235 words per minute, medium speaking rate ranges from 240 to 162 words per minute, slow speaking rate ranges from 311 to 195 words per minute. However, the volume manipulation using SSML loud and soft default settings did not result in a wide range of intensity values across all synthesized speech samples, so we added x-loud and x-soft settings into our experiments. X-loud intensity ranges from 73.7 to 13 dB, medium intensity ranges from 70 to 9 dB, X-soft intensity ranges from 65.6 to 8 dB. Because the literature on speaker gender effects on trust is mixed, and prosodic features differ across genders, we include gender as a demographic factor to explore in this work. A pre-trained male (Matthew) and female (Joanna) with standard American English voices are used to synthesize all speech samples.

In total, there are 45 possible prosodic combinations. Pitch and rate have 3 settings and intensity has 5 settings (3*3*5). Each of the prosodic combinations is synthesized using a male voice and a female voice, resulting in 90 unique combinations. 10 question utterances for each of the possible prosodic combinations are synthesized. Thus, resulted in 900 speech samples. The average duration of audio clips is 2 seconds. We extracted mean values of the three acoustic-prosodic features from all audio clips using Praat [14], a popular open-source software for speech analysis. The mean value of each acoustic-prosodic feature across low, medium, and high settings is shown in table 1. We ensured that all synthesized samples sound natural and not overmanipulated through pilot listening tests in our lab.

Table 1: *Mean value of different levels (x-loud, medium, x-soft) of intensity (dB), (high, medium, low) of pitch (Hz), and speaking rate (words/min) across speaker gender (F is female, M is male).*

Level	Intensity		Pitch		Speaking rate	
	F	M	F	M	F	M
High	57.2	55	203	131	294	321
Med	52	51	163	110	236	256
Low	49	47	131	103	192	205

2.2. Crowdsourcing Experiment

After preparing the 900 speech stimuli with various prosodic styles, we next conducted a crowdsourced perception study.[1] Participants were instructed to listen to 20 audio clips and provide their judgments of their perception of various speaker traits using a 5-point Likert scale. In addition to providing judgments of speaker trustworthiness, subjects rated other traits including whether the speaker sounded lively, empathetic, respectful, cold, boring, and engaging, all of which have been proposed to be positively or negatively associated with trustworthiness. To ensure that subjects pay attention to the task, a short transcription task was included for a subset of the audio samples; only submissions with correct transcriptions were accepted. They also completed the Ten Item Personality Inventory (TIPI) [15] to measure their Big-Five personality dimensions [16] and provided their self-identified gender.

Subjects were recruited using Amazon Mechanical Turk, a widely used crowdsourcing platform. Subjects were eligible to participate in the study if they: (1) are native speakers of Standard American English; and (2) have a task acceptance rate of at least 95%, to ensure that they are high quality crowdworkers. 210 subjects participated in this study and each audio sample was rated by 5 unique raters, resulting in a total of 4500 judgments of 900 speech stimuli. 96 reported themselves as female and 114 reported themselves as male. All collected ratings were normalized by rater ($z = (x - \mu)/\sigma$, where x is the rating of a speech sample, μ is the mean of the ratings provided by the rater, and σ is the standard deviation of the ratings provided by the rater. This was done to allow for a more meaningful comparison of responses across raters and to improve consistency.

3. Analysis and Results

We aim to answer the following questions: (1) How do raters define trustworthiness in terms of other speaker traits? (2) What are the acoustic-prosodic characteristics of trustworthy speech

[1]This study received approval from our university Institutional Review Board (IRB) and all human subjects protection guidelines were followed.

and other speaker traits? (3) How do listener characteristics (e.g. gender, personality) affect their perception of speaker attributes?

3.1. Inter-Annotator Agreement

We computed Krippendorff's alpha [17], a statistical measure of inter-annotator agreement, to understand how the raters tended to agree or disagree on their judgment of the speaker traits. We computed this measure across all raters, and also separately for male and female raters in table 2.

Table 2: *Inter-annotator agreement scores (Krippendorff's α) for each speaker trait, across all raters and separated by male and female raters.*

	trustworthy	lively	natural	boring	empathetic	respectful	cold	engaging
all raters	0.21	0.18	0.17	0.2	0.2	0.18	0.22	0.2
F raters	0.13	0.08	0.08	0.07	0.1	0.1	0.12	0.1
M raters	0.17	0.13	0.14	0.17	0.14	0.15	0.16	0.13

Overall, agreement was not strong for any of the traits, suggesting that the perception of these traits based on synthesized speech samples is subjective. Agreement was not stronger when considering only male or only female raters. The highest scores were found for the perception of *cold* (α=.22) and *trustworthy* (α=.21), indicating slight agreement for those traits.

We also computed the average ratings for each of the traits across all responses, the results are shown in Table 3. As shown in the table, audio clips had the highest average score for *respectful* ($\mu = 3.7$) and the lowest average score for *boring* ($\mu = 2.6$).

Table 3: *Average ratings for each speaker trait, across all responses and separated by male and female raters.*

	trustworthy	lively	natural	boring	empathetic	respectful	cold	engaging
all raters	3.66	3.66	3.48	2.6	3.34	3.7	3	3.6
F raters	3.69	3.7	3.6	2.5	3.3	3.6	3	3.6
M raters	3.66	3.6	3.4	2.6	3.3	3.7	3	3.5
F speaker	3.61	3.6	3.4	2.59	3.3	3.67	3	3.55
M speaker	3.62	3.62	3.41	2.6	3.31	3.69	3	3.54

3.2. Correlation Analysis of Speaker Attributes

We describe how raters define trustworthy speech by analyzing its positive or negative relationship with other speaker traits, including natural, empathetic, boring, cold, respectful, lively, and engaging. We calculated Pearson's correlation between ratings of trustworthiness and ratings of the other speaker traits. The results are shown in Figure 1. We find that the trait of *trustworthy* is positively correlated with the other positive traits, most strongly for *natural* (r=.5), *respectful* (r=.45), *lively* (r=.45), and *engaging* ($r = .45$). This suggests that voices that are perceived as more trustworthy are also perceived as respectful, sound natural and lively, and listeners are more likely to want to engage with the speaker. As expected, the speaker traits *boring* and *cold* are negatively correlated with all other positive traits. We also calculated the correlation for these traits separately for male and female voices but we did not observe any statistically significant differences. This suggests that trustworthiness is consistently perceived in relation to the speaker's other traits for both male and female voices.

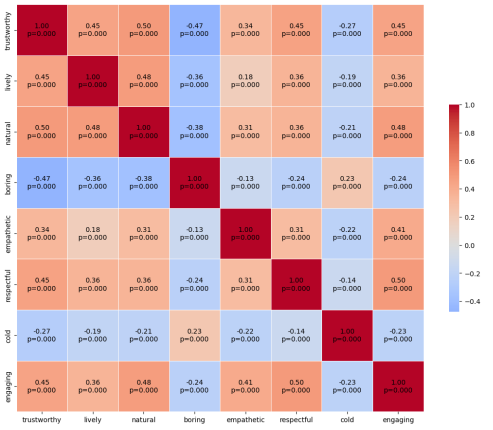

Figure 1: *Pearson's correlations between perceived speaker traits.*

3.3. Acoustic-prosodic Characteristics of Trustworthy Speech

We rank the 45 prosodic styles for both male and female voices by mean trustworthy scores and display the five combinations that received the highest trustworthy ratings and the five combinations that received the lowest trustworthy ratings in Table 4. As shown in the table, all five of prosodic styles associated with the lowest trustworthy scores were synthesized with a fast speaking rate. This was true for both male and female voices. The least trusted female voices also had a low pitch, while the least trusted male voices had a medium pitch. The value difference between loud, medium and soft intensity are small, thus the most trustworthy voices tended to have a medium pitch and speaking rates, and slightly higher intensity.

Table 4: *5 most trustworthy and 5 least trustworthy prosodic styles of synthesized speech.*

Intensity	Pitch	Rate	Gender	Avg Rating
loud	medium	medium	M	4.13
medium	medium	slow	F	4.09
loud	high	medium	M	4.05
soft	medium	medium	M	4.03
soft	high	medium	F	4.02
soft	medium	Slow	M	4
medium	medium	fast	M	3.14
x-loud	low	fast	F	3.14
loud	medium	fast	M	3.12
loud	low	fast	F	3
x-loud	medium	fast	M	2.9

To further understand how each of the acoustic-prosodic features correlates to the perception of speaker attributes, we apply a Generalized Least Squares (GLS) regression analysis. This analysis estimates the relationship between the acoustic-prosodic features (independent variables) and the Likert-scale attribute ratings (dependent variable). We consider results to be significant if the p-value obtained from the regression analysis is smaller than 0.05. As shown in Table 5, the results indicated that medium intensity is positively correlated with the perception of voices as positive traits like trustworthy, lively, natural, and engaging. Low pitch is negatively associated with all speaker attributes except for cold and boring, indicating that it contributes to a negative perception of a voice. On the other

Table 5: *Generalized Least Squares regression analysis results, estimating the relationship between acoustic-prosodic features and perceived speaker attributes. r is the correlation coefficient and p is the p-value. Only results with p ≤.05 are shown in the table. Positive relationships are indicated with green text and negative with red.*

Features	trustworthy		lively		natural		boring		empathetic		respectful		cold		engaging	
	r	p	r	p	r	p	r	p	r	p	r	p	r	p	r	p
intensity x-soft	-0.13	0	-0.08	0.04	-0.12	0	0.25	0			0.11	0	0.21	0	-0.07	0.04
intensity medium	0.31	0	0.29	0	0.13	0	-0.8	0			0.26	0	-0.36	0	0.25	0
intensity x-loud	-0.17	0	-0.07	0.05	-0.17	0	0.29	0	-0.08	0.03			0.22	0	-0.12	0
pitch low	-0.17	0	-0.23	0	-0.18	0	0.11	0	-0.1	0.006	-0.18	0.001	0.16	0.007	-0.2	0
pitch medium																
pitch high	0.29	0	0.3	0	0.13	0	-0.8	0			0.36	0	-0.32	0	0.28	0
speaking rate slow	0.3	0			0.3	0			0.32	0	0.25	0.014			0.23	0
speaking rate medium	0.4	0	0.21	0	0.5	0	-0.16	0	0.39	0	0.3	0			0.37	0
speaking rate fast			0.11	0	-0.27	0	-0.69	0	-0.22	0	-0.1	0	-0.26	0		

hand, high pitch is positively correlated with the perception of voices as trustworthy, lively, respectful, natural, and engaging, and negatively correlated with boring and cold. Finally, slow and medium speaking rates were positively correlated with the positive attributes of trustworthy, natural, empathetic, respectful, and engaging, while fast speaking rate was negatively correlated with natural, boring, empathetic, respectful, and cold attributes but positively correlated with lively.

We also estimate the relationship between acoustic-prosodic features and speaker attributes by modeling the features as numeric variables rather than categorical variables. We compute Pearson's correlations between the numeric features and the speaker attributes and find that the results are very similar so we do not display the results due to lack of space.

3.4. How do listener characteristics affect their perception of speaker attributes?

We also examined the listener traits of gender and personality to see whether these factors influence how they perceive speaker attributes. We use GLS regression analysis to estimate the relationship between the listener's self-reported gender and their ratings of the speaker attributes. We found it is a significant factor in relationship between listener gender and perceived natural (r= -0.18), boring (r= 0.11), and empathetic (r= -0.07). Female listeners were more likely to perceive speakers as natural and empathetic. Male listeners were more likely to perceive speakers as boring compared to female listeners.

Next, we calculated Pearson's correlation between the listeners' TIPI personality scores and their ratings of speaker attributes. The results are shown in Figure 2. All personality dimensions, notably with emotional stability and conscientiousness, were negatively correlated with boring ratings. Extraversion is positively correlated with positive traits such as trustworthy ratings. Moreover, listener traits of emotional stability is positively correlated with ratings of trustworthy and negatively correlated with boring and cold ratings. Finally, the listener trait of agreeableness is negatively correlated with ratings of boring, cold, natural, and engaging.

We also examined whether the listener's gender influenced how they rated speakers of different genders by calculating the correlation between the listener's gender and ratings of male and female speakers separately. The results indicate that the interaction of the listener's and speaker's gender is a statistically significant factor affecting the perception of respectful and engaging speaker attributes. Male listeners tend to rate female speakers as more respectful than female listeners do. And listeners rated speakers with a different gender as more engaging.

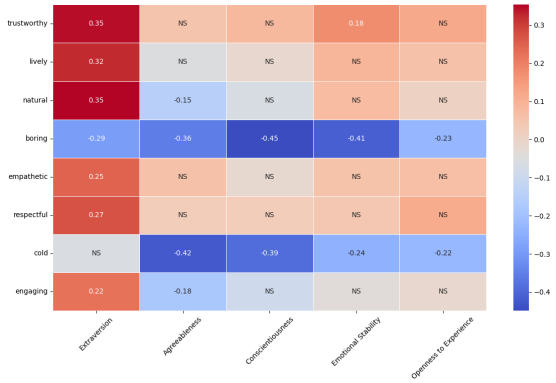

Figure 2: *Pearson's correlation between listener TIPI score and perceived speaker attributes. The correlation coefficient values are displayed for these statistically significant correlations (p-value ≤ 0.05). 'NS' is Not Significant (p-value > 0.05).*

4. Conclusions

In this work, we studied acoustic-prosodic factors that affect the perception of multiple attributes of synthesized speech. We synthesized speech samples with manipulated acoustic-prosodic parameters to study how variations in pitch, intensity, and speaking rate affect the perception of speaker trustworthiness. We also examined the role of speaker gender as well as listener gender and personality traits. We conducted a crowdsourced perception study to collect 4500 judgments of speech stimuli and identify acoustic-prosodic correlates of trustworthy speech. Our results identify specific prosodic patterns of synthesized speech that are associated with perceived trustworthiness. We also find that listener gender and personality traits may affect their perception of trustworthiness and other speaker attributes. This work contributes important insights for building conversational agents that can maximize the perception of trustworthiness, which can ultimately lead to increased usage and adoption of technologies that will benefit society.

5. Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant NSF 23-561 Computer and Information Science and Engineering (CISE).

6. References

[1] X. 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.

[2] S. I. Levitan, *Deception in spoken dialogue: Classification and individual differences*. Columbia University, 2019.

[3] H.-C. Chou and C.-C. Lee, ""your behavior makes me think it is a lie": Recognizing perceived deception using multimodal data in dialog games," in *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2020, pp. 393–402.

[4] S. I. Levitan and J. Hirschberg, "Believe It or Not: Acoustic-Prosodic Cues to Trusting and Untrusting Speech in Interview Dialogues," in *Proc. Speech Prosody 2022*, 2022, pp. 610–614.

[5] L. Gauder, L. Pepino, P. Riera, S. Brussino, J. Vidal, A. Gravano, and L. Ferrer, "Towards detecting the level of trust in the skills of a virtual assistant from the user's speech," *Computer Speech & Language*, vol. 80, p. 101487, 2023.

[6] F. Savoretti, M. Trnka, A. Weise, and J. Hirschberg, "Implementing acoustic-prosodic entrainment in a conversational avatar," 2016.

[7] R. H. Gálvez, A. Gravano, Š. Beňuš, R. Levitan, M. Trnka, and J. Hirschberg, "An empirical study of the effect of acoustic-prosodic entrainment on the perceived trustworthiness of conversational avatars," *Speech Communication*, vol. 124, pp. 46–67, 2020.

[8] L. Gong and C. Nass, "When a talking-face computer agent is half-human and half-humanoid: Human identity and consistency preference," *Human communication research*, vol. 33, no. 2, pp. 163–193, 2007.

[9] L. Qiu and I. Benbasat, "Evaluating anthropomorphic product recommendation agents: A social relationship perspective to designing information systems," *Journal of management information systems*, vol. 25, no. 4, pp. 145–182, 2009.

[10] I. Torre, J. Goslin, L. White, and D. Zanatto, "Trust in artificial voices: A" congruency effect" of first impressions and behavioural experience," in *Proceedings of the Technology, Mind, and Society*, 2018, pp. 1–6.

[11] I. Torre, L. White, and J. Goslin, "Behavioural mediation of prosodic cues to implicit judgements of trustworthiness," in *Speech Prosody 2016*. ISCA, 2016.

[12] S. Liu, C. Zheng, O. Demasi, S. Sabour, Y. Li, Z. Yu, Y. Jiang, and M. Huang, "Towards emotional support dialog systems," *arXiv preprint arXiv:2106.01144*, 2021.

[13] P. Taylor and A. Isard, "Ssml: A speech synthesis markup language," *Speech communication*, vol. 21, no. 1-2, pp. 123–133, 1997.

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

[15] S. D. Gosling, P. J. Rentfrow, and W. B. Swann Jr, "A very brief measure of the big-five personality domains," *Journal of Research in personality*, vol. 37, no. 6, pp. 504–528, 2003.

[16] R. R. McCrae and P. T. Costa, "Validation of the five-factor model of personality across instruments and observers." *Journal of personality and social psychology*, vol. 52, no. 1, p. 81, 1987.

[17] K. Krippendorff, "Computing krippendorff's alpha-reliability," 2011.