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

Acoustic-Prosodic Cues to Trust and Mistrust in Spanish and English Dialogues

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

Trust between conversational partners is critical for effective communication and collaboration. While numerous studies have examined spoken cues to deceptive speech to understand how untrustworthy speech is produced and perceived, little work has studied the characteristics of *trusting* speech, i.e. cues that indicate whether a speaker trusts their conversational partner. This is crucial for monitoring a speaker's perception of their interlocutor, which has implications for conversational outcomes. In this work, we examine trusting speech in both human-human and human-machine dialogues. We study this phenomenon across native speakers of three languages (American English, Mandarin Chinese, and Argentine Spanish) in order to examine how one's native language affects their production of trusting and mistrusting speech. We identify several stable acoustic-prosodic signals of trusting speech across speakers of different native languages and identify some notable differences. This work sheds light on the nature of trusting speech across settings such as culture, language, and domain. We build predictive models of trusting speech using acoustic-prosodic features, in both within- and cross-cultural settings. We study the interpretability of these models and use those insights to improve classification performance.

Index Terms: speech perception, human-human interaction, human-computer interaction, computational paralinguistics, dialogue, trust

1. Introduction

When people trust their conversational partners, they are able to communicate and collaborate effectively. In this work we study trusting and mistrusting speech, i.e. speech produced by a speaker who either trusts or mistrusts their conversational partner. We aim to answer the following research questions: **(1) What are the acoustic-prosodic characteristics of trusting speech and mistrusting speech in Spanish? (2) Are there similarities in acoustic-prosodic features of trusting speech and mistrusting speech across culture and language? (3) Can we build predictive models of trusting vs. mistrusting speech by leveraging acoustic-prosodic features?**

To answer these questions, we use two corpora: a Spanish corpus of human-computer dialogues and an English corpus of human-human dialogues, which we further divided into two subsets: one consisting of speech from native Mandarin speakers and one comprised of speech from native English speakers. All corpora include audio recordings, textual transcripts, and trust ratings provided by each human speaker rating their degree of trust in their conversational partner. We first identify significant acoustic-prosodic features of trusting and mistrusting speech in the Spanish corpus of trusting and mistrusting dialogues. We then compare our findings with cues to trusting and mistrusting speech in the corpora of English deceptive and truthful dialogues (from both native English speakers and native Chinese speakers), to identify similarities across different languages, cultures, and experimental conditions. After identifying trends across corpora, we then build classification models of trusting and mistrusting speech, in monolingual, cross-lingual, and cross-cultural settings. Finally, we use SHAP (SHapley Additive exPlanations) [1] to explain our machine learning model output and select useful features to improve model performance.

There are several potential applications of this work. For spoken dialogue systems, it is useful to model the level of trust that the user feels toward the system, in order to adapt to the user and build user trust. In addition, understanding the nature of trusting speech is useful for determining how people perceive their conversational partners, for example in interview dialogues. This can be helpful for detecting misinformation by capturing an interviewer's level of trust in their conversational partner. An important contribution of this work is its comparison of findings across corpora with speech in different languages and from different cultures. Most prior studies have focused on English speech, where there is a wealth of data and resources. We are particularly interested in examining how cues to trusting speech may be similar across languages and cultures, and possibly experimental conditions, in order to leverage resources from English to model other languages and cultures. In our work building machine learning models to automatically classify trusting and mistrusting speech, we aim to increase interpretability of the models by applying an explainable AI approach to understand how different features contribute to the output. In the remainder of the paper, we first review related work in this field and describe the corpora used in this work. We then detail the features and methods used, and present the results of our detailed analysis of features. Finally, we present classification models to automatically predict trusting vs. mistrusting speech, as well as explanations of the model. We conclude with a discussion of the findings and ideas for future work.

1.1. Related work

A related problem that has been widely studied is understanding and detecting deceptive and truthful speech. The Interspeech 2016 ComParE Deception Sub Challenge [2] presented the problem of automatic detection of deception using acoustic features, and several researchers built predictive models of deceptive speech using the Deceptive Speech Database (DSD) and a baseline acoustic feature set [3, 4]. Similarly, [5, 6] studied deceptive speech in a corpus of truthful and deceptive interview dialogues. They analyzed a set of acoustic-prosodic features in both truthful and deceptive responses and then built predictive models of deception using those features.

Other work has focused on the perception of deception, or

10.21437/SpeechProsody.2024-195

trustworthiness, of sythesized as well as human speech. [7] conducted an experiment in which subjects had to listen to a series of veracity of statements synthesized with varying prosody and indicate if they believed them to be true or false. They identified prosodic characteristics of perceived credibility, persuasion, deception and trustworthiness. [8] analyzed acoustic-prosodic and linguistic characteristics of perceived deception in human conversations, using a corpus of English speech that was judged as trustworthy or not. They also trained machine learning classifiers to distinguish trustworthy from untrustworthy speech.

While most previous work in this area has focused on understanding deceptive or trustworthy speech, some recent work has taken another perspective and studied trusting and mistrusting speech – speech produced by a speaker who either trusts or mistrusts their conversational partner. [9] examined trusting and mistrusting speech in a corpus of conversations in Spanish between humans and machine interlocutors and trained predictive models of trusting speech in Spanish. [10] also studied trusting and mistrusting speech, analyzing a set of acoustic-prosodic features in both trusting and mistrusting speech in human-human conversations in English.

While most previous studies have focused on monolingual data, some have explored cross-cultural [11] and cross-lingual [12] cues to deception and trust. For example, [12] analyzed deceptive and truthful essays in Spanish and English. They trained and evaluated classifiers across cultures and languages. Our work builds on these previous studies and fills in some gaps in the literature. We conduct a comprehensive analysis of acoustic-prosodic features of trusting and mistrusting speech in Spanish and compare the results with cues in English. We also build predictive models of trusting speech in both monolingual and cross-lingual settings. This work is important for understanding how findings in one language or culture may or may not generalize to other languages and cultures.

2. Data and features

2.1. Data

We performed this study in two corpora: The Trust-UBA Database [13] in Argentine Spanish, and the Columbia X-Cultural Deception (CXD) Corpus [11] in English.

Trust-UBA Database is a collection of spoken dialogues between human subjects and a virtual assistant (VA), conducted in Argentine Spanish. Subjects were asked to answer trivia questions with the help of a VA. The subjects were first informed that the VA was previously rated by other users with a high or low trust score, effectively priming them to trust or mistrust the assistant. After interacting with the VA, subjects provided ratings of their trust level in the VA on a scale from 1 (not at all trust) to 5 (completely trust). From the Trust-UBA corpus, we selected subject turns that immediately followed a VA's response to a question. These turns represent an interviewer's reaction to the VA's response and are ideal for studying cues to trust or mistrust. We converted the continuous trust rating to a binary rating (trust or mistrust) in order to be consistent with the labels in the CXD corpus described below. We compared different binning strategies and found that binning scores of 1 and 2 to MT (mistrust) and 3, 4, 5 to T (Trust) resulted in the most similar acoustic cues to the original continuous labels.

CXD Corpus is a collection of within-subject deceptive and non-deceptive speech from native speakers of Standard American English and Mandarin Chinese, all speaking in English. Because we are interested in cultural differences in trusting speech, we divide the corpus into two sub-corpora: CXD-MC, containing all speech from speakers of Mandarin Chinese; and CXD-SAE, containing all speech from speakers of Standard American English. The CXD data was collected using a fake resume paradigm, where participants played the roles of interviewer and interviewee. Interviews were structured around a 24-item biographical questionnaire. In each dialogue, one speaker played the role of interviewer and asked a set of biographical questions while interviewee answered them truthfully or deceptively. Interviewers recorded their judgments for each of the questions (a binary truth or deception label), which provide implicit measures of interviewer trust. From the CXD corpus, we selected interviewer turns that immediately followed an interviewee's response to a question. These turns represent an interviewer's reaction to an interviewee response, and are ideal for studying cues to trust or mistrust.

Table 1 provides a comparison of multiple features of these datasets, it shows that the three corpora are quite similar. The main differences lie in the language/culture of the speakers, as well as the interlocutor (human vs. machine).

Table 1: *Comparison of features of 3 corpora.*

Feature	CXD-MC	CXD-SAE	Trust-UBA
Language spoken	English	English	Spanish
Native language	Mandarin	English	Spanish
Speaker role	interviewer	interviewer	interviewer
Setting	lab	lab	lab
Interlocutor	human	human	machine
Avg. turn duration	3s	2.67s	4.4s
Avg. turn word count	7	7	6.8
Num turns	3669	4340	2950
Count of Trust	2252	2468	1886
Count of Mistrust	1417	1872	1064

2.2. Feature extraction

We extracted the following set of 17 acoustic-prosodic features using Praat [14], a popular open-source software for speech analysis: (1) Duration; (2-6) Pitch minimum, maximum, mean, median, and standard deviation; (7-11) Intensity minimum, maximum, mean, median, and standard deviation; (12) Voiced to total frames; (VCD2TOT, The fraction of pitch frames that are analysed as voiced in the analysed audio) (13) Shimmer; (14) Harmonics to noise ratio; (HNR) (15) Mean absolute slope; (MAS) F0; (16) Jitter; and (17) Speaking rate. All features were Z-normalized by gender (z = (x $-$ μ)/σ; x = value, μ = gender mean, σ = gender standard deviation).

3. Acoustic-Prosodic Cues to Trust and Mistrust

We extract the acoustic-prosodic features described above from all interviewer turns in both corpora. We analyze acoustic-prosodic cues to trust and mistrust in each corpus and then compare findings across corpora. The trust annotations in the Spanish data were originally provided on a continuous scale, from 1-5, while the trust annotations in the CXD corpus were provided on a binary scale (trust or mistrust). We analyze the Spanish data using two methods: (1) continuous; computing the Spearman's correlation between various acoustic features and the trust ratings, and (2) binary; first converting the continuous ratings in the Spanish data to binary ratings and then applying a paired t-test analysis to identify differences in feature distributions between trusting and mistrusting speech. The binary analysis of the Spanish data enables a more direct comparison with

the CXD corpus. However, it does require manipulation of the original ratings by binning trust scores of 1 and 2 to MT (mistrust) and 3, 4, 5 to T (Trust). We report both continuous and binary analysis results for completeness. We report significant results with a p-value < 0.05. Our dataset is large enough to assume all data is normally distributed according to the central limit theorem.

3.1. Acoustic Cues to Trust and Mistrust in Spanish

Table 2 shows the results of the feature analysis of the Spanish Trust-UBA corpus in the first two columns. For the continuous analysis, four features were significantly increased in mistrusting speech: shimmer, max/mean intensity, and jitter. On the other hand, four features were significantly increased in trusting speech: min/mean/median pitch and HNR. Interviewers spoke with significantly higher pitch and better quality and clarity of the voice when they trusted the preceding interviewee turn. When interviewers mistrusted the preceding interviewee's response, they tended to exhibit more diverse pitch patterns, a less stable or rough-sounding voice, and an overall higher volume level. The binary analysis results are shown in the second column of the table, labeled with Trust-UBA and Binary headers. Overall, we observe very similar results using binned vs. continuous trust ratings, with slightly more significant features using the binary labels.

Table 2: *Significant Acoustic-prosodic characteristics of trusting and mistrusting speech in the Spanish Trust-UBA corpus and in the English CXD corpus. T indicates that a feature was significantly increased in trusting speech, and MT indicates 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 ().*

Feature	Trust-UBA Continuous	Trust-UBA Binary	CXD_SAE Binary	CXD_MC Binary
duration		T	T	(T)
min pitch	T	T		
mean pitch	T	T		(MT)
median pitch	T	T		
SD pitch		MT		
MAS F0		MT		(MT)
shimmer	MT	MT		
HNR	T	T		
min intensity		MT		
max intensity	MT			
median intensity		MT	(MT)	
mean intensity	MT			
VCD2TOT			MT	
SD intensity		T		
jitter	MT			
Speaking Rate				(MT)

Comparing Acoustic Cues to Trusting Speech Across Corpora We compare the cues to trusting speech in Spanish with cues in both CXD-SAE and CXD-MC, in order to discover robust signals of trusting speech across language, culture, and experimental conditions. Table 2 shows the results for CXD-SAE and CXD-MC in the last two columns. Under identical experimental conditions (these are subsets of a larger CXD corpus), increased mean pitch, MAS F0, and speaking rate are cues to mistrust for MC speakers and not SAE speakers. When mistrust was observed, native Mandarin speakers tended to exhibit significantly higher pitch with a wider frequency range and faster speech rate. On the other hand, native English speakers displayed significantly louder speech with a more continuous and melodic pattern. This suggests that a speaker's native language or culture may affect their production of trusting or mistrusting speech. Moreover, we observe many more significant indicators of trusting and mistrusting speech in the Spanish corpus compared with the English corpus. This is despite the fact that the Spanish corpus has much fewer data points than the English corpus. It is possible that the Spanish experimental design, which explicitly primed subjects to trust or mistrust the virtual assistant, resulted in a subject speech that was more strongly trusting or mistrusting and therefore had more salient linguistic characteristics. In contrast, the trust ratings English CXD corpus were provided as a measure of believed deception, and were more implicit in nature, resulting in fewer salient cues to trusting and mistrusting speech.

Despite the differences in language, culture, and experimental design, we observed three consistent findings across English (CXD or MC) and Spanish corpora: (1) turn duration is significantly longer in trusting speech, (2) median intensity and (3) MAS (mean absolute slope) F0 are significantly greater in mistrusting speech. Interviewers tended to use shorter turns, speak at a louder volume, and employ voices with higher pitch or a wider frequency range when they did not trust their conversational partners. This suggests that there are some characteristics of trusting and mistrusting speech that are robust across languages, cultures, and experimental conditions.

4. Classification of Trusting and Mistrusting Speech

After identifying trends across corpor, we trained predictive models of trusting and mistrusting speech using linguistic features. We explored three experimental settings for building and evaluating models: (1) monolingual, training and evaluating models with only English or Spanish data, (2) cross-cultural, training models in English from native speakers of English or Mandarin and testing them on English data from native speakers of the other language; and (3) cross-lingual, training models in one language and evaluating them on another language.

We used the Python scikit-learn package [15] to compare several machine learning models, including a baseline classifier which predicts the class label based on the prior probabilities of the classes in the training data, random forest (RF), support vector machine (SVM), and logistic regression (LR). We evaluate all models using the macro average f1 score, which is computed by taking the unweighted mean of all the per-class F1 scores to measure the model performance. This metric is selected due to the unbalanced classes in our data. All our machine learning models outperform the baseline models, and we show the results of the RF model since it outperforms other classifiers. An RF model is an estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting [15]. The parameters such as the number of estimators used in models are the default setting in scikit-learn package.

Monolingual Trusting Speech Classification Table 3 shows the results of the RF model for monolingual classification of trusting speech. We first trained models using a combination of all acoustic features, as well as only a subset of the significant features that were identified previously in our statistical analysis. The best performance was obtained using a combination of all acoustic features for the Spanish Trust-UBA corpus, achieving an F1 of .73. The performance for English corpora was much lower than the performance on Spanish corpus; using a combination of significant acoustic features achieved an F1 of

0.5 for CXD-SAE and an F1 of 0.48 for CXD-MC. A possible reason for this large gap in performance is that the Spanish subjects were first primed to trust or mistrust the assistant, which may have yielded more pronounced cues to trust and mistrust in the subject speech. Across all monolingual experiments, we found that using only significant features resulted in either the same or better performance than using all features. In addition, all of our trained models outperformed the majority class baseline model by large margins.

We further explore methods to improve models' performance. We interpret the model output using SHAP (SHapley Additive exPlanations) [1], a game theoretic approach used to explain the output of each prediction as the sum of the contributions from each predictor. The SHAP Plot sorts features by the sum of SHAP value magnitudes over all samples and uses SHAP values to show the distribution of the impacts each feature has on the model output. The color represents the feature value (red high, blue low). For example, features like median intensity, are more equally distributed and are not clearly skewed toward trust or mistrust prediction. We remove those ambiguous features and the results are shown in the last row of Table 3. As shown in the table, using SHAP values to reduce the feature set resulted in improved classification performance for all three corpora. Overall, the strongest results were obtained for the Spanish corpus (.74 F1) while English-SAE and English-MC were more challenging to model (.54 F1 for both).

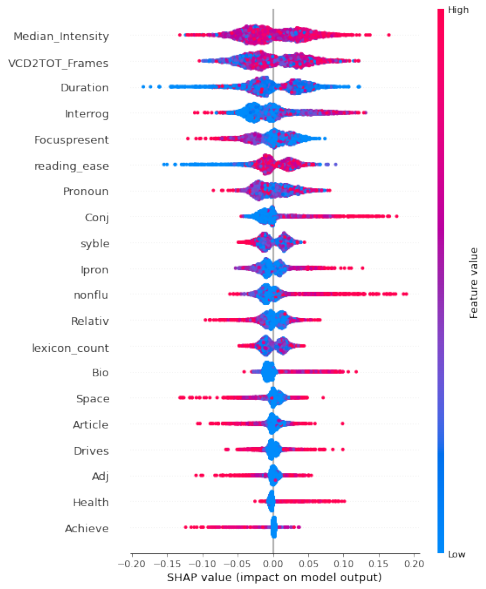

Figure 1: *SHAP plot for English (SAE)*

Table 3: *Monolingual trust classification result of RF and Baseline Models.*

Feature	Language	F1	Std	Baseline_F1	Baseline_Std
All Acoustic	English_SAE	0.48	1.3	0.36	0.7
	English_MC	0.46	4.3	0.38	0.8
	Spanish	0.73	1.6	0.4	0.5
Significant Acoustic	English_SAE	0.5	1.3	0.36	0.7
	English_MC	0.48	2.3	0.38	0.8
	Spanish	0.73	2.8	0.4	0.5
SHAP Acoustic	English_SAE	0.54	1.5	0.36	0.7
	English_MC	0.54	3.3	0.38	0.8
	Spanish	0.74	2.7	0.4	0.5

Cross-Lingual and Cross-Cultural Trust Classification Next, we explore cross-cultural trusting speech classification by training models in English (SAE) and evaluate in English (MC), and vice versa. We also explore cross-lingual classification, training models in one language (Spanish or English) and evaluating in another. This useful in case where labeled training data is available for a particular language or culture and we want to predict trusting speech in another language or culture without available labels. We trained RF models with two different features: (1) all acoustic-prosodic features, and (2) shared significant acoustic-prosodic features between two corpora. The cross-cultural and cross-lingual results are shown in Table 4. Overall, the best performance was achieved using shared significant features between two corpora (F1-Shared). In this setting, the RF model achieved an F1 of 0.5 when training in English SAE and testing in English MC, an F1 of 0.5 when training in English SAE and testing in Spanish. It seems that the English model generalized slightly better across languages than the Spanish model, despite the strong performance of the monolingual Spanish classifier. Overall, the cross-lingual and cross-cultural performance was quite low, although all models performed above the baseline. This is not surprising, as we previously observed substantial differences in cues to trust and mistrust across languages and cultures.

Table 4: *Cross-lingual and Cross-cultural trust classification using RF model. F1_all uses all acoustic-prosodic features and F1_shared uses sharing significant features between two corpora.*

Train	Test	F1_all	F1_shared	Baseline_F1
SAE	MC	0.49	0.5	0.38
MC	SAE	0.42	0.49	0.36
SAE	Spanish	.45	0.5	0.39
Spanish	SAE	0.41	0.47	0.36

To further understand the performance of the models, we computed the precision and recall of both trusting and mistrusting classes for each model. The majority class was the trusting class, and so the baseline model classified all test data points as trusting. In contrast, the RF models were able to correctly classify both trusting and mistrusting classes, and in particular, the models generally had higher recall scores for trusting than mistrusting speech. This suggests that the models tended to perform better at detecting trusting speech than mistrusting speech.

5. Conclusions

This paper presents a study of trusting and mistrusting speech in Spanish, compared with English produced by native speakers of English or Mandarin Chinese. We discovered significant similarities in acoustic-prosodic features across cultures and languages under different experiment settings. We also leveraged our findings on automated trusting vs mistrusting classification tasks in monolingual, cross-lingual, and cross-cultural settings. This work is important for understanding and modeling how speakers from different backgrounds perceive their human or machine conversational partners. We found that using significant acoustic-prosodic features performed the best in monolingual classification, and using shared significant acoustic-prosodic features between corpora performed the best in cross-lingual and cross-cultural settings. We use SHAP plots to improve model performance by removing less predictive features.

In future work, we plan to study cues to trusting and mistrusting speech in real-world interview settings using data from radio and podcast interviews. This will be useful for understanding how these findings compare with real-world speech.

6. References

[1] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[2] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, and K. Evanini, "The INTERSPEECH 2016 Computational Paralinguistics Challenge: Deception, Sincerity Native Language," in *Proc. Interspeech 2016*, 2016, pp. 2001–2005.

[3] B. Schuller, S. Steidl, A. Batliner, S. Hantke, F. Hönig, J. R. Orozco-Arroyave, E. Nöth, Y. Zhang, and F. Weninger, "The interspeech 2015 computational paralinguistics challenge: Nativeness, parkinson's & eating condition," 2015.

[4] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi *et al.*, "The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*, 2013.

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

[6] S. I. Levitan, G. An, M. Ma, R. Levitan, A. Rosenberg, and J. Hirschberg, "Combining acoustic-prosodic, lexical, and phonotactic features for automatic deception detection." in *Interspeech*, 2016, pp. 2006–2010.

[7] A. L. Sandham, C. J. Dando, R. Bull, and T. C. Ormerod, "Improving professional observers' veracity judgements by tactical interviewing," *Journal of Police and Criminal Psychology*, pp. 1–9, 2020.

[8] 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.

[9] 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.

[10] S. I. Levitan and J. Hirschberg, "Believe it or not: Acoustic-prosodic cues to trust and mistrust in spoken dialogue."

[11] V. Pérez-Rosas and R. Mihalcea, "Cross-cultural deception detection," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2014, pp. 440–445.

[12] R. M. Veronica Perez-Rosas, "Cross-cultural deception detection," vol. Volume 2: Short Papers. ACL, 2014, p. 440–445.

[13] L. Gauder, P. Riera, L. Pepino, S. Brussino, J. Vidal, L. Ferrer, and A. Gravano, "Trust-uba: A corpus for the study of the manifestation of trust in speech," *arXiv preprint arXiv:2006.05977*, 2020.

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

[15] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.