Acoustic-Prosodic Indicators of Deception and Trust in Interview Dialogues

Sarah Ita Levitan[1], *Angel Maredia*[1], *Julia Hirschberg*[1]

[1]Columbia University, USA

sarahita@cs.columbia.edu, asm2221@columbia.edu, julia@cs.columbia.edu

Abstract

We analyze a set of acoustic-prosodic features in both truthful and deceptive responses to interview questions, identifying differences between truthful and deceptive speech. We also study the perception of deception, identifying acoustic-prosodic characteristics of speech that is perceived as truthful or deceptive by interviewers. In addition to studying differences across all speakers, we identify individual variations in deception production and perception across gender and native language. We conduct machine learning classification experiments aimed at distinguishing between truthful and deceptive speech, using acoustic-prosodic features. We also explore methods of leveraging individual traits for deception classification. Our results show that acoustic-prosodic features are highly effective at classifying deceptive speech. Our best classifier achieved an F1-score of 72.77, well above both the random baseline and above human performance at this task. This work advances our understanding of deception production and perception, and has implications for automatic deception detection and the development of synthesized speech that is trustworthy.

Index Terms: deception, trust, prosody, computational paralinguistics

1. Introduction

Identifying verbal indicators of deception is an important problem with far-reaching implications. Researchers from several disciplines, including psychology, computational linguistics, and practitioners in law enforcement, military, and intelligence agencies, have attempted to solve this important problem. A less-studied complementary problem, is this task of identifying verbal indicators of trust. Trust is a fundamental component of human-human and human-computer interactions, and understanding the characteristics of trustworthy speech is useful for advancing human-computer interactions, such as spoken dialogue systems and social robots. If we can discover the characteristics of trustworthy speech, such information can be leveraged to create robots that inspire trust. Trust of computer agents is essential for successful interactions [1], whether it is for a robotic companion for elderly and disabled individuals or agents used for military applications.

Although there have been significant efforts aimed at automatic identification of deceptive language, both spoken and written, there has been little work done to characterize the prosodic characteristics of deception. And there has been almost no work that examines the characteristics of trusted speech. In this work we study deception as well as trust, in the context of *perceived deception*. In a related work, [2] manipulated the prosody of synthesized true and false factual statements, and studied the prosody of statements that were judged by humans as true and false. In this work we focus on natural speech in dialogues between human interlocutors, and the lies are not about universal facts, but rather about individual biographical information. There has also been interesting research examining the relationship between entrainment and trust in human-computer interaction [3, 4]. In this work we focus on acoustic-prosodic characteristics of deceptive and trusted speech, independent of the interlocutor's speech.

In this work we aim to answer the following questions:

1. What are the acoustic-prosodic characteristics of truthful and deceptive speech?
2. What are the acoustic-prosodic characteristics of speech that is perceived as truthful (trusted) and perceived as deceptive (mistrusted)?
3. Are there universal characteristics and/or *individual differences* in production and perception of deception?
4. Can we automatically classify deceptive and trusted speech using acoustic-prosodic features?

The remainder of this paper is organized as follows. Section 2 describes the corpus used for this work. In Section 3 we describe the methods used to analyze deception and trust. We present an analysis of acoustic-prosodic indicators of deception in Section 4, and characteristics of perceived deception in Section 5. For both sections, we include an analysis of individual differences in gender and native language. Finally, we present classification results in Section 6. We conclude in Section 7 with a discussion and ideas for future work.

2. Data

We examined the Columbia X-Cultural Deception (CXD) Corpus [5], 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 contains dialogues between 340 subjects. A variation of a fake resume paradigm was used to collect the data. Previously unacquainted pairs of subjects played a ”lying game” with each other. Each subject filled out a 24-item biographical questionnaire and were instructed to create false answers for a random half of the questions. They also reported demographic information including gender and native language, and completed the NEO-FFI personality inventory [6].

The experiment was recorded in a sound booth. For the first half of the game, one subject assumed the role of the interviewer, while the other answered the biographical questions, lying for half and telling the truth for the other; questions chosen in each category were balanced across the corpus. For the second half of the game, the subjects roles were reversed, and the interviewer became the interviewee. During the experiment, the interviewer was encouraged to ask follow-up questions to aid them in determining the truth of the interviewees answers. Interviewers recorded their judgments for each of the 24 questions, providing information about human perception of deception.

Previous work with the CXD corpus has focused on IPU-level and turn-level analysis and classification of local deception, mostly with acoustic-prosodic features [5, 7]. Here we are interested in exploring global deception at the dialogue-level. We explore two segmentation units: (1) first turn; the single interviewee turn directly following a biographical question, and (2) multiple turns; the entire segment of interviewee turns responding to the original interviewer question and subsequent follow-up questions. We used a question identification system [8] that uses word embeddings to match questions produced with lexical variations to a target question list. This was necessary because interviewers asked the 24 questions using different wording from the original list of questions. In our classification experiments, we explore whether a deceptive answer is better classified by examining the interviewee's initial response alone or by examining all of the follow-up conversation between interviewer and interviewee.

3. Method

In order to analyze the differences between deceptive and truthful speech, we extracted the acoustic-prosodic features from 1) the first turns of each question response-segment and 2) the entire question response-segment.

For our analysis of acoustic-prosodic characteristics of deception and perceived deception, we extracted 8 features that are commonly studied in speech research: intensity mean, intensity max, pitch mean, pitch max, jitter, shimmer, noise-to-harmonics ratio (NHR), and speaking rate. Intensity describes the degree of energy in a sound wave, pitch describes the fundamental frequency of a voice, and jitter, shimmer, and NHR are measures of voice quality. Jitter and shimmer are associated with vocal harshness, and NHR is associated with hoarseness. Speaking rate is estimated using the ratio of voiced to unvoiced frames. All acoustic features were extracted using Praat [9], an open-source audio processing toolkit. We then calculated a series of paired t-tests between the features of truthful speech and deceptive speech. All tests for significance correct for family-wise Type I error by 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 F to indicate that a feature was significantly increased in deceptive or perceived deceptive speech, and T to indicate a significant indicator of truth or perceived truth. We consider a result to approach significance if its uncorrected p value is less than 0.05 and indicate this with () in the tables.

4. Acoustic-prosodic Indicators of Deception

In this section we present the results of our analysis of acoustic-prosodic characteristics of truthful and deceptive interviewee responses. In our first analysis across all speakers, we aim to answer the following question: **What are the prosodic differences between truthful and deceptive speech in interviewee responses?** As shown in Table 1, we observed an increase in pitch max in deceptive speech, as well as an increase in intensity max in deceptive speech. This suggests that speakers on average tended to speak with an increase in maximum pitch and an increase in volume when lying.

We observed these trends across all speakers. In our next analysis, we aim to answer the following question: **Are there differences in characteristics of deceptive speech across gender or native language of the speaker?** To answer this question, we computed t-tests between truthful and deceptive interviewee responses for specific groups of speakers. Specifically, we ran four sets of experiments, considering 1) only male speakers, 2) only female speakers, 3) only native speakers of English, and 4) only native speakers of Mandarin Chines. Table 2 shows the results of these experiments, along with the results across all speakers for comparison.

Table 1: *T-tests for global deception: truthful vs. deceptive interviewee responses.*

Feature	***t***	***df***	***p***	***Sig.***
Pitch Max	4.37	7111	1.28E-05	F
Pitch Mean	0.56	7110	0.58	
Intensity Max	3.45	7118	0.0006	F
Intensity Mean	1.33	7131	0.18	
Speaking Rate	-1.69	7135	0.09	
Jitter	-1.31	6757	0.19	
Shimmer	-1.39	6731	0.17	
NHR	0.35	7074	0.73	

Table 2: *T-tests for individual differences in deception: truthful vs. deceptive interviewee responses.*

Feature	**Male**	**Female**	**English**	**Chinese**	**All**
Pitch Max	F			F	F
Pitch Mean					
Intensity Max	F	(F)	F		F
Intensity Mean			(F)		
Speaking Rate				T	
Jitter		(T)			
Shimmer					
NHR					

We previously observed that increased pitch maximum is an indicator of deception across all speakers. When testing differences between truthful and deceptive speech in subsets of our subject population, we find that pitch maximum is significantly increased in deceptive speech for male speakers and for native Chinese speakers, but not for female speakers or for native speakers of English. Another ”universal” indicator of deception was increased intensity max. In this analysis we found that this trend held in all groups except for native speakers of Chinese.

We also observed additional indicators of deception when analyzing individual groups. For example, we observed increased speaking rate in the truthful speech of native Chinese speakers. This is intuitive: non-native speakers tended to speak faster when telling the truth but spoke significantly slower while lying. This supports the theory that lying increases cognitive load [10]. It is interesting to note that this behavior was only exhibited by non-native speakers of English, suggesting that this effect of increased cognitive load is more apparent in speakers conversing in their L2 language. For female speakers, we also observed the trend that jitter was increased in truthful speech.

5. Acoustic-prosodic Indicators of Perceived Trust and Lack of Trust

In this section we present the results of our analysis of acoustic-prosodic characteristics of trustworthy speech. In our first analysis across all speakers, we aim to answer the following question: **What are the prosodic differences between speech that is *perceived* as truthful or deceptive – trustworthy vs. not**

trustworthy? Table 3 displays the results of paired t-tests comparing acoustic-prosodic features of interviewee responses that were believed by interviewers (i.e. judged as truthful – trusted) or not believed (i.e. judged as deceptive – not trusted).

Table 3: *T-tests for global trust: perceived truthful vs. perceived deceptive interviewee responses.*

Feature	*t*	*df*	*p*	*Sig.*
Pitch Max	2.35	6180	0.02	(F)
Pitch Mean	1.65	6183	0.1	
Intensity Max	2.62	6296	0.009	F
Intensity Mean	-0.78	6128	0.43	
Speaking Rate	-3.79	6281	0.0002	T
Jitter	-1.81	6108	0.07	
Shimmer	-1.90	5874	0.06	
NHR	0.58	6088	0.56	

Interviewee responses that were judged as truthful by interviewers had greater pitch max and intensity max. This is consistent with the results in Table 1 which shows that these two features were increased in deceptive speech. The increase in intensity max is also consistent with the findings of [2], where they observed a negative relationship between high intensity in synthesized speech and the probability of humans judging the speech as true.

In addition, utterances that were believed truthful by interviewers exhibited a faster speaking rate than utterances that were perceived as deceptive. This finding is intuitive: speech that is spoken quickly and fluently demonstrates the speaker's familiarity and ease with the topic and inspires trust in the listener. However, it is interesting to note that speaking rate was not in fact significantly different between all instances of truthful and deceptive speech. A difference in speaking rate was only observed in native speakers of Chinese (as shown in Table 2). Thus, it seems that the characteristics of deceptive speech are not always aligned with the characteristics of trusted speech. That is, interviewers were exploiting features that were not in fact useful in identifying deception.

To further understand the nature of speech that was perceived as truthful or deceptive, we conducted the same analysis, but this time examining subsets of the population with respect to gender and native language. We first analyzed the role of the interviewee traits: **Are there differences in characteristics of trusted speech across gender or native language of the *speaker?*** Table 4 compares acoustic-prosodic features in believed and disbelieved speech by gender and native language.

Table 4: *T-tests for individual differences in trust: perceived truthful vs. perceived deceptive interviewee responses, segmented by interviewee traits.*

Feature	Male	Female	English	Chinese	All
Pitch Max	(F)			(F)	(F)
Pitch Mean				F	
Intensity Max				F	F
Intensity Mean					
Speaking Rate	(T)	(T)		T	T
Jitter		(T)	(T)		
Shimmer		(T)	T		
NHR				F	

We see from this table that some characteristics were consistent across multiple groups. For example, speaking rate was faster in trusted speech from all groups except native English speakers. For most features, however, there was considerable variation across speaker groups. For example, when considering only native Chinese speakers, increased pitch mean and increased NHR were strong indicators of perceived deception; this was not the case for any other group. Another interesting finding was that jitter and shimmer were increased in perceived truthful speech, when considering female speakers and native English speakers but not male speakers or native speakers of Chines.

Next, we studied the role of interviewer traits in perception of deception: **Are there differences in characteristics of trusted speech across gender or native language of the *listener?*** Table 5 shows the results comparing acoustic-prosodic features in believed and disbelieved speech, this time for specific groups of *interviewers*.

Table 5: *T-tests for individual differences in trust: perceived truthful vs. perceived deceptive interviewee responses, segmented by interviewer traits.*

Feature	Male	Female	English	Chinese	All
Pitch Max			F		(F)
Pitch Mean	(F)				
Intensity Max	(F)		(F)	(F)	F
Intensity Mean					
Speaking Rate	T		T	(T)	T
Jitter		F			
Shimmer		(F)			
NHR					

We see in this table that some characteristics of perceived deceptive speech are consistent across several interviewer groups. For example, speaking rate was increased in speech that interviewers trusted, and intensity maximum was increased in speech that was *not* trusted, for all interviewer groups except females. There was also considerable variation. For example, jitter and shimmer were increased in disbelieved speech only by female interviewers. Pitch max was increased in disbelieved speech only by interviewers who were native speakers of English. And pitch mean was increased in disbelieved speech only by male interviewers. Overall, we observed greater differences in gender than native language when considering listener traits, and greater differences in native language than gender when considering speaker traits.

6. Deception Classification with Acoustic-prosodic Features

Motivated by our analysis showing significant differences between truthful and deceptive responses to interviewer questions, we trained machine learning classifiers with the task of distinguishing between truthful and deceptive interviewee speech, using acoustic-prosodic features. Because our analysis also demonstrates that acoustic-prosodic indicators for truthful and deceptive speech differs by gender and native language, we 1) incorporated gender and native language in our classification experiments and 2) also trained and tested classifiers on different demographic segments of interviewees (e.g. training and testing on just Male interviewees) to explore the role of interviewee demographic data in classification performance.

Features	Segment	A	P	R	F1
OS	single	74.71	78.58	66.97	72.31
	multiple	74.3	78.04	65.49	71.22
OS-Male	single	74.86	74.84	69.25	71.94
	multiple	73.61	74.43	66.96	70.5
OS-Female	single	73.33	77.54	63.02	69.53
	multiple	74.79	74.87	70	72.35
OS-English	single	72.85	76.36	60.97	67.8
	multiple	73.41	76.59	65.27	70.48
OS-Chinese	single	75.1	72.84	69.94	71.36
	multiple	76.43	77.56	70.14	73.67
OS+G+L	single	75.21	79.3	67.23	72.77
	multiple	73.85	77.44	64.86	70.6

Table 6: *Deception classification of single turn and multiple turn segmentations, using a Random Forest classifier trained on openSMILE (OS) features and demographic features of gender and native language. Additionally, classification results for training on openSMILE features for specific speaker demographics (e.g. OS-Male refers to models trained on just the openSMILE features of male interviewees). OS+G+L refers to classification using openSMILE, gender, and native language features.*

We compared classification performance for the two segmentation methods described in section 2: first turn and multiple turns. This allowed us to explore the role of context in automatic deception and trust detection. When classifying interviewee response-segments, should the immediate response only be used for classification, or is inclusion of surrounding turns helpful? This has implications not only for deception and trust classification, but for practitioners as well. Should human interviewers make use of responses to follow up questions when determining response veracity, or should the initial response receive the most consideration?

We use the Interspeech 2009 (IS09) ComParE Challenge baseline feature set, which contains 384 acoustic-prosodic features from the computation of various functionals over low-level descriptor (LLD) contours extracted from openSMILE [11]. The LLD features include pitch (fundamental frequency), intensity (energy), spectral, cepstral (MFCC), duration, and voice quality. This standard feature set was designed for emotion recognition, which is a relevant computational paralinguistic task. We compared the performance of 3 classification algorithms: Random Forest, Logistic Regression, and SVM (sklearn implementation). Random Forest was the best performing classifier, so we report only those results due to space constraints.

There were 7,878 question segments for both single turn and multiple turns after balancing the data such that there were an equal number of true and false responses. For each of the classifiers trained on different demographic segments of interviewee speakers, we had 3,596 question segments by male interviewees, 4,244 female, 4,174 question segments by native English speaking interviewees, and 3,666 from native Mandarin speakers. We divided the question segments into 80% train and 20% test, and used the same fixed test set in experiments for single-turn and multiple-turns segmentation in order to directly compare results. The random baseline performance is 50%, since the dataset is balanced. A stricter baseline is human performance, which is 56.75% in this corpus.

Table 6 presents the deception classification performance for openSMILE features and combinations of openSMILE features with demographic features for both single turn and multiple turn segmentations. Our results demonstrate that acoustic-prosodic features are highly effective at this deception classification task. The single turn classifier achieved an F1-score of 72.31, and the classifier trained on multiple turns achieved an F1-score of 71.22, both well above the random as well as human performance baselines. It seems that the prosody of the first turn is more useful than the prosody of the full set of turns in an interviewee response.

We found that, for single turn classification, training gender-specific and language-specific classifiers was not helpful, resulting in either slightly worse or equivalent performance to training on all speakers. This was not the case for multiple turn classification: female-specific and Chinese-specific classifiers performed better than the generic classifier. This suggests that when more context is available, classifiers that are trained on homogenized speakers can improve over a classifier trained on a diverse set of speakers. Finally, we found that adding gender and native language as features did not significantly improve over using acoustic-prosodic features alone. It is likely that gender and native language may be captured to some extent by the acoustic-prosodic feature set; therefore, adding this information as features is perhaps redundant.

7. Conclusions

In this paper we have presented a study of deception and trust in interview dialogues. Our analysis of acoustic-prosodic characteristics of deceptive and truthful speech provides insight into the nature of deceptive speech. We also analyzed the acoustic-prosodic characteristics of speech that is *perceived* as deceptive or truthful, which is important for understanding the nature of trustworthy speech. We explored individual variation across gender and native language in acoustic-prosodic cues to deception, highlighting cues that are specific to a particular subset of speakers. We also explored individual variation in perception of deception by analyzing features of trustworthy speech by speaker traits as well as listener traits. Finally, we trained classifiers to automatically distinguish between truthful and deceptive speech using acoustic-prosodic features, and explored leveraging individual traits for classification. Our best classifier achieved an F1-score of 72.77, well above both the random baseline and above human performance at this task. This work contributes to the critical problem of automatic deception detection, and increases our scientific understanding of deception, deception perception, and individual differences in deceptive behavior.

In future work, we plan to focus on the problem of trust classification, training machine learning models to automatically identify trustworthy vs. not trustworthy speech. We also plan to conduct similar analyses using deception corpora in other domains, in order to identify consistent domain-independent deception indicators. In addition, we plan to conduct cross-corpus machine learning experiments, to evaluate the robustness of these and other feature sets in deception detection. We also would like to explore additional feature combinations, such as adding linguistic features.

8. Acknowledgements

This work was partially funded by AFOSR FA9550-11-1-0120 and by NSF DGE-11-44155.

9. References

[1] A. Glass, D. L. McGuinness, and M. Wolverton, "Toward establishing trust in adaptive agents," in *Proceedings of the 13th international conference on Intelligent user interfaces*. ACM, 2008, pp. 227–236.

[2] R. H. Gálvez, Š. Benuš, A. Gravano, and M. Trnka, "Prosodic facilitation and interference while judging on the veracity of synthesized statements," *Proc. Interspeech 2017*, pp. 2331–2335, 2017.

[3] 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*, 2018.

[4] 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, 2016, pp. 1166–1170.

[5] 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*. ACM, 2015, pp. 1–8.

[6] P. Costa and R. McCrae, "Neo five-factor inventory (neo-ffi)," *Odessa, FL: Psychological Assessment Resources*, 1989.

[7] G. Mendels, S. I. Levitan, K.-Z. Lee, and J. Hirschberg, "Hybrid acoustic-lexical deep learning approach for deception detection," *Proc. Interspeech 2017*, pp. 1472–1476, 2017.

[8] A. S. Maredia, K. Schechtman, S. I. Levitan, and J. Hirschberg, "Comparing approaches for automatic question identification." SEM, 2017.

[9] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer.[computer program]. version 6.0. 19," 2016.

[10] A. Vrij, G. R. Semin, and R. Bull, "Insight into behavior displayed during deception," *Human Communication Research*, vol. 22, no. 4, pp. 544–562, 1996.

[11] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 1459–1462.