Acoustic-Prosodic Indicators of Deception and Trust in Interview Dialogues

Sarah Ita Levitan\(^1\), Angel Maredia\(^1\), Julia Hirschberg\(^1\)

\(^1\)Department of Computer Science, 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 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. We previously studied the linguistic characteristics of deception and perceived deception in interview dialogues [2]; here we focus on acoustic-prosodic characteristics. In a related work, [3] 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 [4, 5]. 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 gender and native language 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 differences across 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 [6], 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 [7].

The experiment was recorded in a sound booth, where the two subjects were separated by a curtain to ensure no visual contact. 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 [6, 8]. 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 [9] 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 [10], an open-source audio processing toolkit, and z-score normalized by gender. 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 $\frac{k}{m}$. In all the tables in this paper, we use $F$ to indicate that a feature was significantly increased in deceptive speech, and $T$ to indicate a significant indicator of truth or perceived truth. We consider a test to result in significance if its uncorrected $p$ value is less than 0.05 and indicate this with ($\cdot$) in the tables.

4. Acoustic-prosodic Deception Indicators

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 Chinese. Table 2 shows the results of these experiments, along with the results across all speakers for comparison.

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 similar groups of speakers. 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 [11]. 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.

---

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

<table>
<thead>
<tr>
<th>Feature</th>
<th>$t$</th>
<th>$df$</th>
<th>$p$</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pitch Max</td>
<td>4.37</td>
<td>7111</td>
<td>1.2E-05</td>
<td>F</td>
</tr>
<tr>
<td>Pitch Mean</td>
<td>0.56</td>
<td>7110</td>
<td>0.58</td>
<td></td>
</tr>
<tr>
<td>Intensity Max</td>
<td>3.45</td>
<td>7118</td>
<td>0.0006</td>
<td>F</td>
</tr>
<tr>
<td>Intensity Mean</td>
<td>1.33</td>
<td>7131</td>
<td>0.18</td>
<td></td>
</tr>
<tr>
<td>Speaking Rate</td>
<td>-1.69</td>
<td>7135</td>
<td>0.09</td>
<td></td>
</tr>
<tr>
<td>Jitter</td>
<td>-1.31</td>
<td>6757</td>
<td>0.19</td>
<td></td>
</tr>
<tr>
<td>Shimmer</td>
<td>-1.39</td>
<td>6731</td>
<td>0.17</td>
<td></td>
</tr>
<tr>
<td>NHR</td>
<td>0.35</td>
<td>7074</td>
<td>0.73</td>
<td></td>
</tr>
</tbody>
</table>

**Table 2:** T-tests for individual differences in deception: truthful vs. deceptive interviewee responses. M=Male, F=Female, E=English, C=Chinese.

<table>
<thead>
<tr>
<th>Feature</th>
<th>M</th>
<th>F</th>
<th>E</th>
<th>C</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pitch Max</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pitch Mean</td>
<td></td>
<td></td>
<td>(F)</td>
<td></td>
<td>F</td>
</tr>
<tr>
<td>Intensity Max</td>
<td>F</td>
<td>(F)</td>
<td>F</td>
<td></td>
<td>F</td>
</tr>
<tr>
<td>Intensity Mean</td>
<td>(F)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Speaking Rate</td>
<td></td>
<td></td>
<td>T</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jitter</td>
<td></td>
<td></td>
<td></td>
<td>(T)</td>
<td></td>
</tr>
<tr>
<td>Shimmer</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NHR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

We observed these trends across all speakers. In our next
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.

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

Interviewee responses that were judged as deceptive 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 [3], 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.

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.

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

### Table 4: T-tests for gender and native language differences in trust: perceived truthful vs. perceived deceptive interviewee responses, segmented by interviewee traits. M=male, F=female, E=English, and C=Chinese.

<table>
<thead>
<tr>
<th>Feature</th>
<th>M</th>
<th>F</th>
<th>E</th>
<th>C</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pitch Max</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(F)</td>
</tr>
<tr>
<td>Pitch Mean</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intensity Max</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>F</td>
</tr>
<tr>
<td>Intensity Mean</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>F</td>
</tr>
<tr>
<td>Speaking Rate</td>
<td>T</td>
<td>T</td>
<td>(T)</td>
<td>(T)</td>
<td>T</td>
</tr>
<tr>
<td>Jitter</td>
<td>(T)</td>
<td>(T)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shimmer</td>
<td>T</td>
<td>T</td>
<td>(T)</td>
<td>(T)</td>
<td>T</td>
</tr>
<tr>
<td>NHR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>F</td>
</tr>
</tbody>
</table>

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
and 20% test, and used the same fixed test set in experiments. English speaking interviewees, and 3,666 from native Mandarin viewees, 4,244 female, 4,174 question segments by native En- classiers trained on different demographic segments of inter-

sifier, so we report only those results due to space constraints. Random Forest was the best performing clas-

acoustic-prosodic features, and explored leveraging individual characteristics of deceptive and truthful speech provides insight into the nature of deceptive speech. We also analyzed the acoustic-prosodic characteristics of the speed that is perceived as deceptive or truthful, which is important for understanding the nature of trustworthy speech. We explored 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 variation in perception of deception by analyzing features of trustworthy speech by speaker traits as well as listener traits. Finally, we trained classifiers to automati-cally 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 hu-

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 demo-graphics (e.g. OS-M refers to models trained on just the openS- MILE features of male interviewees). OS+G+L refers to clas-sification using openSMILE, gender, and native language fea-

features. Table 6 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 differ-

demographic segments of interviewees (e.g. training and testing on just Male interviewees) to explore the role of interviewee demographic data in classification performance.

We compared classification performance for the two seg-

mentation methods described in section 2: first turn and mul-
tiple turns. This allowed us to explore the role of context in automatic deception and trust detection. When classifying in-

terviewee 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 inter-


eviewers make use of responses to follow up questions when determining response veracity, or should the initial response receive the most consideration?

We used the Interspeech 2009 (IS09) ComParE Challenge baseline feature set, which contains 384 acoustic-prosodic features from the computation of various functionalities over low-level descriptor (LLD) contours extracted from openSMILE [12]. The LLD features include pitch (fundamental frequency), in-
tensity (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 algo-rithms: Random Forest, Logistic Regression, and SVM (sklearn implemen-tation). Random Forest was the best performing clas-sifier, 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 inter-

viewees, 4,244 female, 4,174 question segments by native En-
glish 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 di-

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 multi-

ple turn segmentations. Our results demonstrate that acoustic-prosodic features are highly effective at this deception classifi-

cation 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 interview 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 sig-
nificantly 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 char-

acteristics of deceptive and truthful speech provides insight into the nature of deceptive speech. We also analyzed the acoustic-prosodic characteristics of the speed that is perceived as deceptive or truthful, which is important for understanding the nature of trustworthy speech. We explored 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 variation in perception of deception by analyzing features of trustworthy speech by speaker traits as well as listener traits. Finally, we trained classifiers to automati-
cally 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 hu-

man performance at this task. This work contributes to the criti-
cal 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 automati-
cally 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. We also wanted to explore additional feature combinations, such as adding linguistic fea-

8. Acknowledgements

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


