Acoustic-Prosodic and Lexical Cues to Deception and Trust: Deciphering How People Detect Lies

Xi (Leslie) Chen, Sarah Ita Levitan, Michelle Levine, Marko Mandic & Julia Hirschberg
Department of Computer Science
Columbia University
New York, NY, USA
{xi_chen, sarahita, michelle}@cs.columbia.edu,
mm5305@columbia.edu, julia@cs.columbia.edu

Abstract

Humans rarely perform better than chance at lie detection. To better understand human perception of deception, we created a game framework, LieCatcher, to collect ratings of perceived deception using a large corpus of deceptive and truthful interviews. We analyzed the acoustic-prosodic and linguistic characteristics of language trusted and mistrusted by raters and compared these to characteristics of actual truthful and deceptive language to understand how perception aligns with reality. With this data we built classifiers to automatically distinguish trusted from mistrusted speech, achieving an F1 of 66.1%. We next evaluated whether the strategies raters said they used to discriminate between truthful and deceptive responses were in fact useful. Our results show that, although several prosodic and lexical features were consistently perceived as trustworthy, they were not reliable cues. Also, the strategies that judges reported using in deception detection were not helpful for the task. Our work sheds light on the nature of trusted language and provides insight into the challenging problem of human deception detection.

1 Introduction

Humans are notoriously poor lie detectors, most performing at chance level or worse (Bond Jr and DePaulo, 2006). This result has been found across a wide variety of deception detection tasks, in multiple modalities, and in different cultures. Although poor performance has been well-attested, very little work has been done to understand *why* humans perform so poorly at detecting lies.

Because humans are so poor at deception detection, there have been many efforts to develop automated methods to detect deception in multiple modalities. Biometric indicators, typically measured by the *polygraph* (a device used to detect lies by measuring blood pressure, pulse, respiration, and skin conductivity), have been shown to perform poorly at deception detection (Eriksson and Lacerda, 2007). Facial expressions (Ekman, 2009a), gestures and body posture (Lu et al., 2005; Tsechpenakis et al., 2005), and even brain imaging (Meijer and Verschuere, 2017) have been explored as potential indicators of deception. Some of these features are difficult or expensive to capture automatically, or are too invasive to be practical for general use. In recent years, automatic deception detection has gained popularity in the speech and NLP communities. Language cues have the advantage of being inexpensive, non-invasive, and easy to collect automatically. More importantly, prior research examining linguistic cues to deception has been promising. Researchers have used machine learning to identify deceptive language in various domains, including court testimonies (Fornaciari and Poesio, 2013), hotel reviews (Ott et al., 2011), and interview dialogues (Levitan et al., 2018b). These automated methods have demonstrated that machine learning classifiers can indeed identify deceptive language with accuracy between 70% and 90%, depending on the task—much better than human performance on the same task. These studies have also identified specific characteristics of deceptive language.

Despite these important advances in understanding and automatically identifying deception, there has been little work investigating *human perception* of deception. What linguistic and prosodic characteristics of an utterance lead listeners to

Transactions of the Association for Computational Linguistics, vol. 8, pp. 199–214, 2020. https://doi.org/10.1162/tacl_a_00311
Action Editor: Eric Fosler-Lussier. Submission batch: 8/2019; Revision batch: 11/2019; Published 4/2020.

believe that it is true—to trust it—regardless of whether it is true or not? Why do people frequently believe lies? How do the strategies humans use in lie detection align with actual indicators of deception and how do they relate to people's performance in lie detection? Can we in fact train machine learning classifiers to automatically identify speech that will be perceived as truth (trusted) or lie (mistrusted) by humans?

To investigate these questions, we created a lie detection game, **LieCatcher**, to conduct a large-scale study of human perception of deception. The stimuli for this game were drawn from a large corpus of previously collected truthful and deceptive dialogues; players were asked to judge whether single utterance spoken responses to written questions were truthful or deceptive. We distributed the game on Amazon's Mechanical Turk crowd-sourcing platform to collect large scale judgments of deceptive or true responses to a set of biographical questions. We systematically analyzed a number of linguistic and prosodic features in the rated responses to understand the characteristics of trusted vs. mistrusted speech. We compared these features to the actual characteristics of truthful and deceptive responses presented in the game to identify the similarities and differences between human perception of deception and the actual production of deception. We also examined player-reported strategies to discover which the raters believed to be useful and which were in fact useful or not useful for detecting deception. Finally, we trained machine learning classifiers using a large set of lexical and speech features to automatically identify human-trusted speech.

The contributions of this paper include: 1) A large-scale analysis of linguistic and prosodic cues to trust compared with cues to deception; this adds considerably to our scientific understanding of human perception of deception. Our results show that there are several prosodic and lexical features that were consistently perceived as trustworthy, but that these were not reliable cues to deceptive speech. 2) A game framework for studying deception perception, which can be extended to other speech and language perception studies. 3) A classifier that uses lexical and acoustic-prosodic features to identify speech that was trusted by humans, achieving an F1 of 66.1%. 4) An analysis of successful and unsuccessful human strategies for detecting deception, showing that strategies that judges reported using in deception detection were not helpful for the task. We further believe that this latter analysis may be useful for training humans to detect lies more successfully.

2 Related Work

Previous studies have examined deceptive language in various domains, including fake reviews (Ott et al., 2011), public trials (Pérez-Rosas et al., 2015), TV shows (Pérez-Rosas et al., 2015), Twitter (Addawood et al., 2019), opinions on controversial topics (Mihalcea and Strapparava, 2009), online games (Zhou et al., 2004), and interviews (Levitan et al., 2018a, b). Machine learning classifiers have been shown to outperform human judges by a large margin. For example, Ott et al. (2011) trained a deception classifier that achieved nearly 90% accuracy on a corpus of fake hotel reviews, whereas human accuracy was about 60%.

Researchers have also examined various features that are characteristic of truthful vs. deceptive language. A meta-study by Bond Jr and DePaulo (2006) highlighted several patterns of deceptive language found in multiple studies, such as shorter responses, fewer details, and more negative emotions. Other cues to deception that have been identified include language that is less sensory or concrete (Ott et al., 2011; Vrij et al., 2006). Truthful language has been found to contain more linguistic markers of certainty (Levitan et al., 2018b; Rubin et al., 2006). Syntactic features such as lexicalized production rules and part of speech tags have also been shown to be useful in predicting deception (Feng et al., 2012; Pérez-Rosas and Mihalcea, 2015). Linguistic Inquiry and Word Count (LIWC) (Pennebaker and King, 1999), which groups words into psychologically meaningful dimensions, has also been used extensively in deception studies (Ott et al., 2011; Pérez-Rosas and Mihalcea, 2015; Pérez-Rosas et al., 2015). Prosodic cues to deception have also been identified; for example, Levitan et al. (2018a) found increased pitch maximum and intensity maximum are indicators of deception. Though these studies are critical for advancing the state of machine deception detection and for understanding the nature of deceptive language, they do not address the question of human perception of deception, which is the focus of this work. We aim to gain insight into why humans are poor judges

T	my dad works i don't i never really know how to say it he works with computers um in information technology
T	he's a technical engineer at draper laboratory
F	he works for um it's like um a subsidiary of walgreens kind of it's very it's very corporate it's like a big big very impersonal company which is i think he doesn't like about it
F	uh my dad is an official in uh in the government system

Table 1: CXD corpus example responses to the question, ''What is your father's job?''

of deception by comparing actual cues to deception with characteristics of language trusted and mistrusted by humans.

Psychology research of human deception detection has traditionally focused on facial expression cues (Ekman et al., 1991; Frank et al., 2008) and personal beliefs about what characterizes deceptive behavior (The Global Deception Research Team, 2006; Granhag and Strömwall, 2004; Wright et al., 2014). Based on worldwide survey studies, The Global Deception Research Team (2006) found pan-cultural deception stereotypes that liars tend to be nervous with flawed speech. However, Hartwig and Bond (2011) pointed out the methodological limitation of such studies: We cannot be certain that what people report reflects their actual decision process (Nisbett and Wilson, 1977). Our work attempts to decipher the cues people actually use to detect lies by examining features of utterances that are labeled as true by participants, compared with features of utterances rated as lies.

3 CXD Corpus

We used deceptive and truthful utterances from the Columbia X-Cultural (CXD) Corpus for our deception perception study (Levitan et al., 2015). The CXD Corpus is a collection of interviews between native speakers of Standard American English and Mandarin Chinese, all speaking in English. It contains 122 hours of conversational speech between 340 individuals. Previously unacquainted pairs of participants were brought into the lab to interview one another. They were first surveyed for gender and native language and asked to complete the NEO-FFI personality inventory (Costa and McCrae, 1989). They were then asked to provide true answers to a set of 24 biographical questions and then to provide false answers for a random half we chose. Interviews took place in a soundproof booth and each pair of participants took turns playing the role of interviewer and interviewee. During the game, the interviewer asked the 24 questions in any order and was encouraged to ask follow-up questions to help determine whether the interviewee was lying or telling the truth about each question. Participants were financially compensated for both successful deception and successful deception detection. Table 1 provides sample responses to one of the questions.

The recorded interviews were orthographically transcribed using Amazon Mechanical Turk (AMT) crowd-sourcing and the transcripts were force-aligned with the audio recordings using the Kaldi Speech Recognition Toolkit (Povey et al., 2011). The interviews were segmented using a question identification classifier (Maredia et al., 2017). All interviewee turns were automatically identified using the question identification system and subsequently hand-corrected. The corpus was segmented into: 1) question responses: The single interviewee turn directly following the question; 2) question chunks: All interviewee turns in (1) plus answers to subsequent follow-up questions. We used the single turn question response segmentation for our deception perception study, so as not to influence raters' responses with interviewers' follow-up questions.

4 LieCatcher

Using the data described in Section 3, we created a lie detection game called LieCatcher.[1] LieCatcher is a Game With A Purpose that allows players to assess their overall ability to detect lies, while simultaneously providing deceptive speech judgments that we then use to study deception perception. We developed LieCatcher in Unity[2] and hosted the game on the Web. Figure 1 shows a screenshot from the game. In the game, players are shown a text version of a question asked

[1]The LieCatcher game framework is publicly available at `https://github.com/sarahita/LieCatcherGame`.

[2]`https://unity.com`.

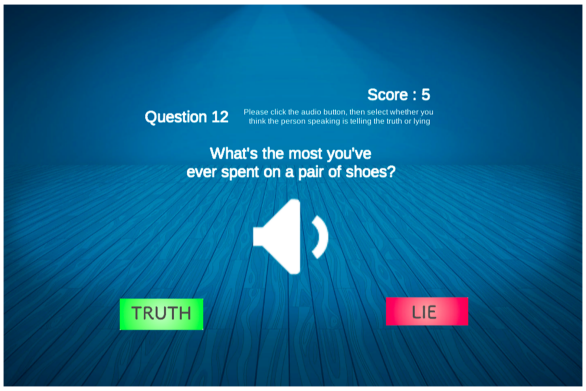

Figure 1: Screenshot from LieCatcher gameplay.

by one interviewer and then listen to the single-turn spoken interviewee response. After listening, the player selects a ‘‘Truth’’ or ‘‘Lie’’ button, indicating their perception of the speech sample as truthful or deceptive. The game was designed so that players could submit their decision only after the audio had finished playing, so they could not make a judgment without hearing the full response. This feature of the game also provided information about raters’ behavior when making judgments, as we recorded the time interval between the end of the audio clip and the time that the player entered their response for each decision. After the gameplay, a score report is displayed summarizing all their judgments for that task, giving players feedback about their performance at the end of each multi-question task.

4.1 Crowdsourcing Experiment

We used the game to collect deception judgments via crowd-sourcing on AMT. On AMT we first vetted potential raters by giving them a language background questionnaire and restricting raters to those who had spoken English fluently since the age of 5 years. In the game, each player was shown a series of 13 questions, one at a time, with the audio recording of the interviewee response. Audio samples were balanced by gender, native language of speaker, and question number (there were no duplicate questions within a game), with half of the responses true and half false. For quality control, we included a randomly placed check question instructing the annotator to select a certain answer for that question (e.g., ‘‘wait 5 seconds and then press False’’) to help ensure that raters were actually paying attention to the game with their audio on. Annotators were also given a post-game survey including questions on previous experience in jobs related to deception detection, their gender, their own confidence level in spotting lies, and the strategies they used in making judgments. We manually filtered out annotators who answered the check question incorrectly or who failed to finish the game or survey. We obtained institutional review board approval for our deception perception study and followed all human subject protection guidelines.

Each response was rated by three annotators and each annotator was limited to a maximum of 10 total tasks of 13 questions each. In total, 5340 utterances were annotated by 431 total annotators; 4.8% of the raters said they had had previous experience in law enforcement. In our sample, 38.9% of the annotators self-identified as male, 59.1% female, and 2.1% other. On average, annotators judged 49.93% of the utterances correctly, roughly at chance. In cases where all three annotators agreed on a judgment, the accuracy was 50.75%, slightly higher than the overall accuracy but still at chance. This is consistent with decades of research in deception detection (Bond Jr and DePaulo, 2006).

4.2 Inter-annotator Agreement

We used Fleiss’ kappa to measure inter-annotator agreement on whether an utterance was truthful or deceptive. The annotators had a Fleiss’ kappa of 0.135, indicating slight agreement as Landis and Koch (1977) suggests, showing that this task is highly subjective. We also computed inter-annotator agreement across utterances from female vs. male speakers and from native English versus native Mandarin speakers. We found only slight agreement (Fleiss’ kappa in the range (0.10, 0.15)) across all speaker traits, indicating that people did not agree more on speakers with certain traits. Lastly, we considered whether inter-annotator agreement might be affected by utterance length, since annotators might find it more difficult to judge short utterances (e.g., one word utterances like ‘‘yes’’ or ‘‘no’’) for lack of sufficient information. However, we found that agreement was uniformly low across all utterance lengths and that longer utterance length tended to result in even lower agreement (Fleiss’ kappa: 0.138 for length ≤ 5 words vs. 0.067 for length ≥ 30 words).

We plotted the distribution of trust levels over all responses (Figure 2) and found that the distribution is skewed toward trust, indicating that

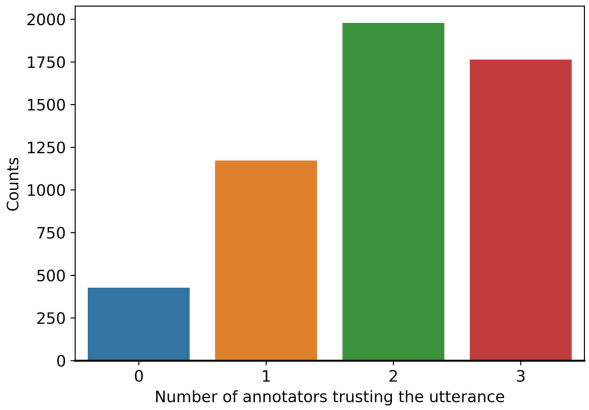

Figure 2: Distribution of utterance trust levels.

annotators tend to be more trusting than mistrusting, with 33% of utterances trusted by all annotators and 70% trusted by at least two, consistent with the Truth Default Theory (Levine, 2014), which posits that humans operate on a default presumption that others are basically honest.

5 Textual and Prosodic Indicators of Trust and Deception

In this section, we consider the following questions: What are the characteristics of trusted and mistrusted speech? How do these compare with the characteristics of truthful and deceptive speech? Of all features raters believed to indicate lies, which are valid cues and which not? Also, what are the deceptive cues that raters failed to perceive? We compared features of trusted and mistrusted utterances and features of truthful and deceptive utterances using paired t-tests. Labels for trust were computed using majority vote (i.e., an utterance is considered trusted if at least 2 annotators believed it is true, otherwise mistrusted). We also compared utterances trusted by all and mistrusted by all, observing differences in complexity and prosodic features. For complexity features, #verbs, #nouns, #num, concreteness ceased to be significant, and type-token became significant with $p < 0.05$. For prosodic features, pitch mean and pitch std ceased to be significant, and intensity max became significant with $p < 0.05$. Notice that all features that differ had relative small significance levels $p > 0.001$. To prevent the inflation of false positive errors caused by conducting multiple comparisons, we present only features that are statistically significant after Benjamini–Hochberg correction (Benjamini and Hochberg, 1995).

Disfluency

Social psychologists hypothesize that telling a lie can be more cognitively demanding than truth-telling (Hauch et al., 2015; M DePaulo et al., 2003). False responses are hypothesized to be less fluent than true responses because fabricating a story takes more mental effort than recalling an actual event. We considered a wide range of features indicative of disfluencies and report those that were statistically significant in either the responses raters labeled as lies or those that actually appear in lying responses in the corpus:

Filled Pauses: We curated a list of filler words based on previous studies of deception (Enos, 2009; Bachenko et al., 2008); we included the binary indicator and the total count.

Response Latency: For deception, this is defined as the time span between the interviewer question and the first non-filler word of the interviewee response; for trust, it is defined as the time span between the start of the audio and the first non-filler word of the interviewee response. Note that latency is defined differently for trust since annotators only heard the segmented version of the response which did not include the silence between the question and the response.

False Starts: A type of speech disfluency where a speaker begins an utterance or a phrase and then self-corrects it; annotation of disfluencies was included in the corpus transcription.

Repetitions: The number of identical, consecutive words or bigrams (e.g., "he he has a...").

Table 2 shows which features appeared in responses raters believed to be deceptive (Column Trust) and which appeared in responses that actually **were** deceptive (Column Deception).

Overall, our findings are consistent with Zuckerman et al. (1981): Speech hesitations and errors are perceived as signs for incompetence and cues for deception. Of all the disfluency features, filled pauses proved to be the strongest reliable indicator of deception and this feature was also perceived correctly by raters. For raters, response latency was the strongest cue for mistrust among

Feature name	Trust	Deception
has filled pause	↓↓↓↓	↑↑↑↑
#filled pauses	↓↓↓↓	↑↑↑↑
has false start	↓↓↓	↑↑
response latency	↓↓↓↓	
repetitions	↓↓↓↓	↑

Table 2: Statistically significant indicators of trust and deception for disfluency features. For this and subsequent tables, the direction of the arrow indicates whether the relationship is positive or negative. The number of arrows indicates the significance level, ↓: < 0.05, ↓↓: < 0.01, ↓↓↓: < 0.001, ↓↓↓↓: < 0.0001.

all disfluency features; however, this was not a reliable indicator of lies. This is also consistent with the findings of Zuckerman et al. (1981) and of Hartwig and Bond (2011). Word and bigram repetition, as well as false starts, were traits of speech raters mistrusted, even though they were only weak cues to deception.

Complexity

Previous research suggests that deceptive statements tend to be simpler and less complex than true ones. This is because of the theory that cognitive load is increased during deception, which can limit creative and complex utterance production (Hauch et al., 2015; M DePaulo et al., 2003; Hartwig and Bond, 2011). Based on these findings, one might expect lies to be less lexically diverse, shorter, and less elaborate than true responses. Do raters appear to use these cues in judging deception? We used utterance length as the most direct indicator of response complexity and also counted the number of words with more than six characters and the number of content words in the utterance as a more fine-grained indication of complexity. We used type-token ratio to capture lexical diversity. We also considered word entropy but found it to be strongly correlated with number of words so decided not to include it as a separate feature. In addition, we used Flesch reading ease (Kincaid et al., 1975) to identify readability, specificity score (Li and Nenkova, 2015) as an indication of the level of detail on the sentence level, and concreteness score (Brysbaert et al., 2014) as an indicator of the level of details of the speakers' visual and haptic experiences. We also used discourse markers (causation and conjunction) extracted from LIWC (Pennebaker and King, 1999), inspired by the hypothesis that liars might use fewer discourse markers in their utterances (Newman et al., 2003).

Features	Trust	Deception
#sent	↓↓↓↓	↑↑↑↑
#word	↓↓↓	↑↑↑↑
#word per sent	↓↓↓	↑↑↑↑
#word>6	↓↓	↑↑↑↑
type-token		↑↑↑
#verb	↓↓	↑↑↑↑
#noun	↓↓	↑↑↑↑
#adj		↑↑↑↑
#num	↓	↑↑↑
#proper nouns	↓	
concreteness	↓↓	↑↑↑↑
specificity	↓↓	↑↑↑↑
#conj		↑↑

Table 3: Statistically significant indicators of trust and deception for complexity features.

As shown in Table 3, overall, raters were more likely to judge longer and more complex responses as deceptive. Contrary to previous research, we found that lies tended to be more complex than true utterances: they tended to be longer, included more specific Language, and were more lexically diverse. They were also more concrete and contained higher numbers of verbs, nouns, adjectives, numbers, and conjunctions. Although raters were apparently using these cues to predict lies, they were relatively weak ones.

Sentiment

When lying, people may experience feelings of guilt and fear of being caught, which may result in their use of more negative words (M DePaulo et al., 2003; Hauch et al., 2015; Ekman, 1988, 2009b). Abe et al. (2007) also found that the act of deceiving is uniquely associated with neural structures associated with heightened emotion. We extracted positive emotion and negative emotion using LIWC (Pennebaker and King, 1999). We also extracted Pleasantness, Activation, and Imagery scores for each utterance from the Dictionary of Affect (DAL) (Whissell, 1989) by summing up the scores of all words. We normalized all features to reduce length effect.

As shown in Table 4, truthful utterances in the corpus contained more visually descriptive

Features	Trust	Deception
DAL-imagery	↑↑↑↑	↓↓↓↓
DAL-activation	↑↑↑↑	↓↓↓↓
DAL-pleasant	↑↑↑↑	

Table 4: Statistically significant indicators of trust and deception on sentiment features.

Features	Trust	Deception
has hedge phrase	↓↓↓↓	↑↑↑
#hedge phrases	↓	↑↑↑↑
certain		↓↓↓

Table 5: Statistically significant indicators of trust and deception on uncertainty features.

words than deceptive utterances, and listeners were more apt to rate utterances with descriptive words as truthful. This is consistent with findings in Masip et al. (2005) that providing sensory details is more difficult when fabricating a story. Truthful utterances also contained words with higher activation scores than deceptive utterances, and trusted utterances also had higher activation scores than mistrusted ones. Consistent with Hartwig and Bond (2011), raters judged more pleasant utterances as truthful, although this was not a valid cue in the CXD corpus.

Uncertainty

Our previous research (Levitan et al., 2018b) has found that linguistic markers of certainty and uncertainty are significant indicators of deception. So we measured certainty and uncertainty in two ways: words from LIWC's ''certainty'' category as linguistic markers of certainty (e.g., always, never) and hedge words and phrases (e.g., possible, sort of) (Ulinski et al., 2018) as indicators of uncertainty. As shown in Table 5, there was a match between rater trust and true responses for hedge words and phrases. In the CXD corpus, lies included hedge phrases more often than true responses did, and we found that listeners did mistrust responses containing hedge phrases. However, although linguistic markers of certainty in the corpus (e.g., ''always,'' ''never''—which are the opposite of hedge words) were indicators of truth, raters failed to perceive this.

Creativity

Do liars tend to rely upon certain ''templates'' or generic responses when answering questions for lack of a more detailed story to present? Do truth-tellers provide more creative responses based on reality? To measure creativity of responses, we examined how similar a response was to other responses to the same question. For each question, we converted all responses to TF-IDF vectors on unigrams and bigrams. We built a lexical graph for each question with responses as nodes and cosine similarities between TF-IDF vectors as edge weights. Then we computed the eigenvalue centrality for each node and used its negative value as the measure of creativity. The intuition here is that the more central a response is, the more similar it is to its neighbors and thus less "creative."

We found liars to be more creative than truth-tellers. We verified this result by counting the number of neighbors within a certain cosine distance in the TD-IDF space. This result is robust against various threshold of cosine distance (0.1–0.9 with 0.1 as the step size). The difference was not due to response length, as we found no correlation between creativity and response length (spearman, $\rho = 0.007$, $p > 0.05$). However, we did find that judgments were not influenced by whether the response was creative or not. Perhaps when people lie they try to tell a compelling story, which results in a more creative response regardless of length.

Prosody

Previous studies have shown that pitch maximum and intensity maximum are significant indicators of deception (Levitan et al., 2018a). We examined whether prosodic features impacted listeners' trust. We extracted a set of 14 features from Praat (Boersma and Weenink, 2009), an open-source audio processing toolkit, and z-score normalized the features by gender. We used the total number of words divided by duration of utterance as a measure of speaking rate. As shown in Table 6, raters judged speech that was loud (high intensity min and mean) and had less variation in intensity (low std) as trustworthy, perhaps because louder speech can sound more confident. Trusted speech also had higher degrees of jitter, shimmer, and noise to harmonics ratio (NHR), which are measures of voice quality. Though not a valid indication of truth, faster speaking rate was also trusted,

Features	Trust	Deception
speaking rate	↑↑↑↑	
pitch max	↑↑↑↑	↑↑↑↑
pitch min	↑↑↑↑	↓↓
pitch mean	↑↑	
pitch std	↑↑	↑↑
intensity max		↑↑↑
intensity min	↑↑↑↑	↓
intensity mean	↑↑↑↑	
intensity std	↓↓↓↓	↑
NHR	↑↑↑↑	
jitter	↑↑↑↑	
shimmer	↑↑↑↑	

Table 6: Statistically significant indicators of trust and deception on prosodic features.

perhaps because raters expected speakers to speak more slowly when lying. This is consistent with previous findings that, while faster speaking rate is trusted, it is not an actual cue to trustworthiness (Zuckerman et al., 1981; Hartwig and Bond, 2011). Listeners also trusted speech with higher pitch (max, min, and mean) and greater pitch variance (std). However, a higher pitch max and greater pitch std were in fact signs of deception.

Of the 11 prosodic cues of mistrust, only 3 (27%) are actually valid indicators of deception; of the 6 prosodic cues of deception, 3 (50%) are also valid indicators of mistrust. This is in contrast to the overlap we see across all features reported. Of the 31 cues that are significant indicators of mistrust, 20 (65%) are also valid indicators of deception; of the 28 cues that are indicative of deception in the data, 20 (71%) are actually indicative of mistrust. This suggests that, although there were several characteristics of trusted speech that were in fact associated with truth, and also many characteristics of mistrusted speech that were associated with deception, prosodic cues to deception are far more difficult for humans to correctly perceive than other cues.

5.1 Can We Predict Trust and Deception?

Based upon the analysis of the linguistic and prosodic characteristics of trusted and deceptive utterances above, we developed predictive models of trust and deception to identify the relative strengths of each type of feature. In addition, we included additional speaker traits (gender, native language, personality), which have also been shown to identify significant differences in speaker trust (Levitan et al., 2018a). We observed several differences in trust behavior across these speaker traits.

- **Gender:** We observed a gender difference in trust ($\chi^2(1) = 5.16$, N = 5340, $p < 0.05$) with female speakers trusted (71.50% of all utterances) more than males (68.61%).

- **Native language:** We observed a native language difference in speaker trust by raters ($\chi^2(1) = 30.22$, N = 5340, $p < 0.0001$) with native American English speakers trusted (73.52% of all utterances) more than native Chinese speakers (66.59%).

- **Personality:** We partitioned CXD speakers into the NEO-FFI Five Factor personality groups by binning personality scores into "high,‘‘ ’’average,‘‘ and ’’low" in each dimension as described in Levitan (2019). We observed significant differences in speakers' responses trusted by raters in the following dimensions:

  - *Conscientiousness*: Speakers with low scores (71.91%) were more trusted than people with neutral (69.12%) or high scores (67.66%). ($\chi^2(2) = 7.22$, N = 5340, $p < 0.05$)
  - *Openness*: Speakers with high scores (71.55%) were more trusted than people with neutral (68.64%) or low scores (67.40%). ($\chi^2(2) = 6.40$, N = 5340, $p < 0.05$)
  - *Neuroticism*: Speakers with high scores (71.75%) were more trusted than people with neutral (68.65%) or low scores (66.48%). ($\chi^2(2) = 8.93$, N = 5340, $p < 0.05$)

In addition, we included a large number of **data-driven** features extracted from the spoken utterances and from their text transcripts. In contrast to the features that were analyzed in Section 5, which were specifically motivated by the deception detection literature, these data-driven features were chosen because of their usefulness for a wide range of NLP and speech processing tasks. Data-driven features included: dependency triples backed off

Feature Sets	Precision	Recall	macro-F1
random	50.19	50.29	44.97
majority	40.27	50.00	44.61
data-driven (9538)	73.45	59.72	61.51
disfluency (5)	76.83	56.73	57.28
+ prosody (16)	78.90 (**80.14**)	60.52 (58.66)	62.74 (60.20)
+ sentiment (5)	78.75 (40.27)	61.19 (50.00)	63.61 (44.61)
+ uncertainty (3)	78.11 (40.27)	61.80 (50.00)	64.36 (44.61)
+ creativity (1)	77.94 (40.27)	61.90 (50.00)	64.48 (44.61)
+ complexity (16)	77.71 (55.86)	62.24 (50.31)	64.87 (45.77)
+ speaker traits (7)	77.55 (40.27)	**63.34** (50.00)	**66.10** (44.61)
all (9591)	74.26	60.34	62.37

Table 7: Prediction results for trust averaged over 5 cross-validation splits. The number of features in each set is included in parentheses in the feature set column. We incrementally added each feature set and also included the individual performance of each feature set in parentheses in the precision/recall/macro-F1 columns.

to parts of speech; one-hot encoded unigrams and bigrams; average of word vectors using GloVe embedding pretrained on Twitter; Interspeech 2013 (IS13) ComParE Challenge baseline feature set, which contains 6373 features resulting from the computation of functionals over low-level descriptor contours extracted from openSMILE (Eyben et al., 2010). The prosodic feature analysis in Section 5 was conducted on a small set of prosodic features extracted using Praat. These were a subset of the openSMILE feature set and were used for clarity and interpretability, while the openSMILE features used here are a larger set for classification experiment.

Because of the low agreement of the annotations, we took an approach similar to Danescu-Niculescu-Mizil et al. (2013) and considered only utterances that annotators reached consensus on for experiments on trust. In total, 1762 utterances were trusted by all annotators and 427 utterances were mistrusted by all annotators. Due to the small size of this dataset, we randomly divided the speakers into five bins of similar size and performed cross-validation with utterances from speakers in each bin as the test data and the rest as the training data; thus, learned models were always evaluated on unseen speakers. On each speaker split, we trained logistic regression models and tuned penalty parameter C and the type of regularization using five cross-validation folds on the training data. We also experimented with linear SVM, gradient boosting classifier, and random forest classifier and found no improvement. We normalized all features to have zero mean and unit-variance. For data-driven features, feature selection was performed to prune the feature space. Because this is an unbalanced task, we evaluated our models using precision, recall, and macro-F1 score.

As shown in Table 7, we can predict whether an utterance is trusted with an average macro-F1 score of 66.10%. We found that 5 disfluency features combined with 16 prosody features outperformed 9538 data-driven features, demonstrating the efficacy of the feature sets specifically designed for this task. Prosody (macro-F1, 60.20%) is the strongest feature set that is predicative of trust, with intensity mean as the strongest feature (45.50%) within that set. Disfluency (57.28%) is the next strongest feature set, with response latency (58.40%) as the strongest feature. Of the individual feature sets, we found that speaker traits, sentiment, uncertainty, and creativity did not perform better than the random baseline, which suggested they are not useful cues on their own for predicting rater trust. However, we found that including speaker traits helped improving the classifier's performance when combined with other features.

In addition to classifying trust, we trained a logistic regression model using these same features to classify deception. Our best model achieved

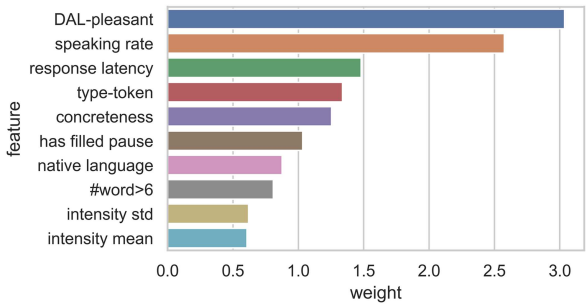

Figure 3: The most important 10 features for predicting trust. The *x*-axis denotes the absolute values of the feature weights averaged over 5 cross-validation splits.

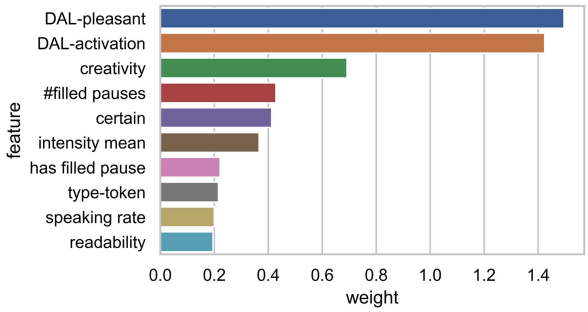

Figure 4: The most important 10 features for predicting deception. The *x*-axis denotes the absolute values of the feature weights averaged over 5 cross-validation splits.

an F1 score of 55.5% using a combination of disfluency, prosody, and uncertainty features. We note that this performance is substantially lower than the trust classification results, suggesting that distinguishing between truthful and deceptive utterances is much harder than distinguishing between utterances trusted and mistrusted by all annotators. However, we also note that this model was not optimized for deception classification (e.g., no feature selection or parameter tuning) and was trained on a very small amount of data. Our previous work obtained better performance at deception detection (69.8 F1) using more data and using models that were optimized for the task (Levitan, 2019). The purpose of this current experiment was to directly compare trust and deception classification using the same data and features.

Figures 3 and 4 show the most important 10 features for predicting trust and deception. To compute feature importance, we averaged the absolute values of the feature weights across the 5 cross-validation splits. We found that DAL-pleasant, has filled pauses, intensity mean, and type-token

	Features	Successful?
1.	#sent	↓↓↓↓
2.	response latency	↓↓↓↓
3.	has filled pause	↓↓↓↓
4.	#filled pauses	↓↓↓↓
5.	repetitions	↓↓↓↓
6.	intensity min	↑↑↑↑
7.	intensity mean	↑↑↑↑
8.	intensity std	↓↓↓↓
9.	speaking rate	↑↑↑↑
10.	shimmer	↑↑↑↑

Table 8: Top 10 statistically significant features for lies that successfully deceived human judges.

are important for both prediction tasks. Of the top three features in either task, speaking rate and response latency are only important for predicting trust, and DAL-activation and creativity are only important for predicting deception.

5.2 The Mechanism Behind a Successful Lie

To better understand why people are ‘‘vulnerable’’ to deceptive utterances, and to understand the characteristics of successful lies, we ran paired t-tests with Benjamini–Hochberg correction to compare the linguistic and prosodic features of successful vs. unsuccessful lies. As shown in Table 8, we found that successful lies differed most from unsuccessful lies in that they contained fewer sentences (1) and were shorter in duration (2). Successful lies were also louder (6, 7), faster (9), had fewer filled pauses (3, 4), varied less in intensity (8), and were harsher (10) in voice quality. When people were more successful at lying, they tended to respond quicker (2) and did not repeat themselves (5).

We also conducted this analysis for classifier judgments, to understand the characteristics of lies that successfully deceived a lie detection classifier. Table 9 shows the top features that discriminate between successful and unsuccessful lies. There are several features that were similar for both human and machine judgments. For example, successful lies had fewer filled pauses (2, 3) and were shorter in duration (5) and number of sentences (9). Some features, however, were unique to deceiving a classifier. For example, lies that successfully deceived the deception classifier were

	Features	Successful?
1.	creativity	↓↓↓↓
2.	has filled pause	↓↓↓↓
3.	#filled pauses	↓↓↓↓
4.	specificity	↓↓↓↓
5.	duration	↓↓↓↓
6.	pitch max	↓↓↓↓
7.	#word	↓↓↓↓
8.	#word per sent	↓↓↓↓
9.	#sent	↓↓↓↓
10.	concreteness	↓↓↓↓

Table 9: Top 10 statistically significant features for lies that successfully deceived a deception classifer.

less creative (1) and less specific (4). It seems that different kinds of lies were successful at deceiving humans and an automated deception classifier.

5.3 Individual Differences in Lie Detection Ability

We found that gender was not significantly related to accuracy (Mann–Whitney U, $p > 0.05$), but female participants took longer to judge (Mann–Whitney U, $p <$ 1e-20). In addition, we did not observe female participants to be more or less trusting than male participants (Mann–Whitney U, $p > 0.05$) or significantly different from male participants in their level of confidence (Mann–Whitney U, $p > 0.05$). Raters with previous job experience related to lie detection did not perform better than those without such experience (1-sample t-test, $p > 0.05$), but they did take longer to make judgments (1-sample t-test, $p < 0.05$). They were at the same level of trust (1-sample t-test, $p > 0.05$) and same level of confidence (1-sample t-test, $p > 0.05$) as those without prior experience. This is consistent with previous findings that persons in the legal professions are no better at detecting deception than others (Aamodt and Custer, 2006).

6 Characterizing Strategies for Detecting Lies

We summarized annotator-provided strategies for detecting lies based on previous work in deception detection (Hauch et al., 2015; M DePaulo et al., 2003; Albrechtsen et al., 2009; Blair et al., 2010; Vrij et al., 2006) and annotators' responses. The strategies were manually labeled by domain experts with previous research experience in deception detection and all ambiguities were discussed by three people. For each strategy, we computed the average percentage of utterances judged correctly by annotators who reported using this strategy and compared it with the average percentage of utterances judged correctly across all annotators. We performed the same analysis for the percent of utterances the raters believed to be true (trusted). In addition, we reported the percentage of annotators who claimed to have used these strategies in Table 10. Prosody, response latency, pauses, disfluency, and intuition were the top five strategies mentioned by annotators, and in Section 5 we verified that the set of features related to prosody, response latency, pauses, and disfluency were indeed significant indications of trust. As shown in Table 10, none of the reported strategies was associated with an improved deception detection performance. However, we did find that using speaker ''confidence'' as a cue to deception was negatively associated with the annotators' performance.

Which strategies are reported by raters who are more or less trusting over all?

We found that people who reported using response latency, pauses, and disfluency when judging deception trusted a smaller percentage of utterances. This could be because of the high prevalence of disfluencies in spontaneous speech, regardless of whether the utterance was deceptive or not. We also found that raters who used the level of detail in a response as a cue were more mistrusting in their judgments. Conversely, those who used clarity of a response and prior domain knowledge were more trusting.

Does complex reasoning correlate with accuracy in lie detection or trust level?

We examined two measures of player behavior that approximate complex reasoning: How long do people take to make judgments? How many strategies do they report in total?

We measured how long people took to judge responses using the time interval between the end of the audio clip and the time that annotator entered his/her response. We found no correlation between response time and the percentage of

Strategy	%Correct	%Trust	%Used	Example
Prosody	-0.25	-0.17	45.74%	**voice tone and pattern**
Response latency	$+0.11$	-2.13^{**}	30.71%	listened for **delays in the speakers response**
Pauses	-0.52	-2.95^{**}	24.66%	I listened for **pauses** to see...
Disfluency	-0.59	-1.88^{*}	22.87%	If they said **''um''** I thought they were lying
Intuition	$+1.09$	$+0.52$	22.87%	My **gut instinct**...
Details	$+0.81$	-2.95^{*}	17.26%	...how much or how little **detail** they used...
Prior	$+1.95$	$+2.85^{*}$	13.90%	How **realistic** the answers were
Style	-0.65	$+0.86$	11.88%	**Anxiety** in voice
Confidence	-2.83^{*}	-1.60	11.21%	paying attention to the person's **confidence**..
Duration	-0.94	-2.80	9.41%	**length** of answer
Speaking rate	$+0.39$	-0.64	6.72%	**Speed of answer**
Speaker traits	$+0.07$	-0.00	6.05%	how **relaxed** they were
Lexical	$+1.53$	$+1.00$	5.16%	Look for **context around the words**
Laughter	$+1.04$	$+0.40$	1.79%	if they **laugh** its false
Clarity	$+2.52$	$+9.71^{*}$	1.35%	People usually give more and **clearer** details...
Breathing	$+5.33$	-2.73	1.12%	I tried to notice when they **breathe so deeply**..
Repeat question	$+0.36$	$+6.10$	0.67%	I id notice one person **repeat the question**..
Contradictions	$+0.04$	$+1.24$	0.67%	...the person blatantly **contradicted** themselves...
Repetition	$+1.24$	$+6.52$	0.44%	**repetition** when lying

Table 10: For each strategy, we show the increase or decrease in the average percentage of utterances trusted/judged correctly for annotators reporting that strategy compared with all annotators. We also show percentage of annotators reporting the strategy and a sample response from one. For % correct and % trust, the statistical significance is computed by comparing the annotators who said they used the strategy and annotators who did not with a Mann–Whitney–Wilcoxon U test. * $p < 0.05$; **; $p < 0.01$; ***; $p < 0.001$.

answers correct. However, we did find a negative correlation between response time and the percentage of answers trusted (spearman, $\rho = -0.101$, $p < 0.0001$). We found a similar result using the number of strategies as a proxy for complex reasoning. There was no correlation between the number of strategies reported and the accuracy score, but we did discover a negative correlation between the percentage of answers trusted and the number of strategies raters reported using (spearman, $\rho = -0.133$, $p < 0.01$). These findings indicate that complexity of reasoning process does not correlate with lie detection performance but negatively correlates with trust level.

7 Conclusion

In this paper we presented a framework for understanding human deception perception. We created a lie detection game, **LieCatcher**, and used it to collect large-scale judgments of deceptive speech. We analyzed a large set of linguistic and prosodic cues to deception and identified some mismatches between the responses people perceived as deceptive and those that were actual deceptive responses. Particularly notable in these mismatches were prosodic features, suggesting that humans have difficulty interpreting prosodic cues to deception.

We built a predictive model of trust with a macro-F1 score of 66.1%, and showed that disfluencies and prosody were most useful for predicting trust. We summarized and manually annotated annotator-provided strategies and found that none of them were associated with an improvement in lie detection ability; however, some were associated with raters' tendency to trust. The identified mismatches between features of trusted vs. deceptive speech, as well as the lack of useful strategies reported by raters, shed new light on the poor performance of humans at deception detection. In addition, we showed that complex reasoning did not correlate with accuracy in deception detection but negatively correlated with trust level.

This work has implications for several applications in multiple disciplines. In business, politics, and interpersonal relationships, it is critical to cultivate the trust of others. Our empirically identified characteristics of trusted language provide useful information for training individuals

who want to speak in a more trustworthy manner. Furthermore, we are interested in using these findings to synthesize voices that are likely to be trusted by others and we have already begun that process. Potential applications that can benefit from trustworthy voices include dialogue systems and robots, especially for assistive technologies (e.g., for individuals with disabilities, elderly individuals) where trust is crucial for successful interactions. Our LieCatcher game was a useful framework for studying perceived deception in an engaging format. The experiments presented in this paper were conducted using stimuli from the CXD corpus. In future experiments, we plan to conduct a cross domain analysis to see if these findings generalize to other domains and corpora. Because of the lack of corpora with annotations of human perceptions of deception, we plan to conduct similar perception studies using the LieCatcher framework to enable this cross-domain analysis. In addition, the LieCatcher game can be extended to explore the perception of other aspects of spoken language. We are currently exploring its use for training purposes. In the future, we will provide more immediate feedback to players about their judgments for each response, with the goal of training practitioners to improve their performance at deception detection.

Acknowledgments

This work was funded by AFOSR grant FA9550-18-1-0039, ''Spoken Indicators of Trust Across Cultures.'' We thank the anonymous reviewers for their helpful feedback. Thanks also to James Shin, Ivy Chen, and William Wang (ww2742) for their help implementing LieCatcher.

References

Michael Aamodt and Heather Custer. 2006. Who can best catch a liar? a meta-analysis of individual differences in detecting deception. *The Forensic Examiner*, 15:6–11.

Nobuhito Abe, Maki Suzuki, Etsuro Mori, Masatoshi Itoh, and Toshikatsu Fujii. 2007. Deceiving others: Distinct neural responses of the prefrontal cortex and amygdala in simple fabrication and deception with social interactions. *Journal of Cognitive Neuroscience*, 19(2):287–295.

Aseel Addawood, Adam Badawy, Kristina Lerman, and Emilio Ferrara. 2019. Linguistic cues to deception: Identifying political trolls on social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 15–25.

Justin S. Albrechtsen, Christian Meissner, and Kyle Susa. 2009. Can intuition improve deception detection performance? *Journal of Experimental Social Psychology*, 45:1052–1055.

Joan Bachenko, Eileen Fitzpatrick, and Michael Schonwetter. 2008. Verification and implementation of language-based deception indicators in civil and criminal narratives. In *Proceedings of the 22Nd International Conference on Computational Linguistics - Volume 1*, COLING '08, pages 41–48. Stroudsburg, PA, USA. Association for Computational Linguistics.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300.

J. Pete Blair, Timothy Levine, and Allison Shaw. 2010. Content in context improves deception detection accuracy. *Human Communication Research - HUM COMMUN RES*, 36:423–442.

Paul Boersma and David Weenink. 2009. Praat: doing phonetics by computer (version 5.1.13).

Charles F. Bond Jr and Bella M. DePaulo. 2006. Accuracy of deception judgments. *Personality and Social Psychology Review*, 10(3):214–234.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46(3):904–911.

P. T. Costa and R. R. McCrae. 1989. NEO five-factor inventory (Neo-FFI). Odessa, FL: Psychological Assessment Resources, 3.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 250–259.

Sofia, Bulgaria. Association for Computational Linguistics.

Bella M. DePaulo, James J. Lindsay, Brian E. Malone, Laura Muhlenbruck, Kelly Charlton, and Harris Cooper. 2003. Cues to deception. *Psychological Bulletin*, 129:74–118.

Paul Ekman. 1988. Lying and nonverbal behavior: Theoretical issues and new findings. *Journal of Nonverbal Behavior*, 12(3):163–175.

Paul Ekman. 2009a. Lie catching and micro-expressions. *The Philosophy of Deception*, pages 118–133.

Paul Ekman. 2009b. *Telling lies: Clues to Deceit in the Marketplace, Politics, and Marriage (revised edition)*. WW Norton & Company.

Paul Ekman, Maureen O'Sullivan, Wallace V. Friesen, and Klaus R. Scherer. 1991. Face, voice, and body in detecting deception. *Journal of Nonverbal Behavior*, 15(2):125–135.

Frank Enos. 2009. *Detecting Deception in Speech*. Ph.D. thesis, Columbia University, New York, NY, USA. AAI3348430.

Anders Eriksson and Francisco Lacerda. 2007. Charlatanry in forensic speech science: A problem to be taken seriously. *International Journal of Speech, Language and the Law*, 14(2): 169–193.

Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: The Munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM'10, pages 1459–1462. ACM, New York, NY, USA.

Song Feng, Ritwik Banerjee, and Yejin Choi. 2012. Syntactic stylometry for deception detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2*, ACL '12, pages 171–175. Stroudsburg, PA, USA. Association for Computational Linguistics.

Tommaso Fornaciari and Massimo Poesio. 2013. Automatic deception detection in italian court cases. *Artificial intelligence and law*, 21(3):303–340.

Mark G. Frank, Melissa A. Menasco, and Maureen O'Sullivan, New York. 2008. Human behavior and deception detection. *Wiley Handbook of Science and Technology for Homeland Security*.

Pär Anders Granhag and Leif A. Strömwall. 2004. *The detection of deception in forensic contexts*. Cambridge University Press.

Maria Hartwig and Charles Bond. 2011. Why do lie-catchers fail? a lens model meta-analysis of human lie judgments. *Psychological Bulletin*, 137:643–59.

Valerie Hauch, Iris Blandon-Gitlin, Jaume Masip, and Siegfried Sporer. 2015. Are computers effective lie detectors? a meta-analysis of linguistic cues to deception. *Personality and Social Psychology Review*, 19:307–342.

J. Peter Kincaid, Robert P. Fishburne Jr, Richard L. Rogers, and Brad S. Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. *Institute for Simulation and Training, University of Central Florida*, 56.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 159–174.

Timothy R. Levine. 2014. Truth-default theory (TDT) a theory of human deception and deception detection. *Journal of Language and Social Psychology*, 33(4):378–392.

Sarah I. Levitan, Guzhen An, Mandi Wang, Gideon Mendels, Julia Hirschberg, Michelle Levine, and Andrew Rosenberg. 2015. Cross-cultural production and detection of deception from speech. In *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection*, WMDD '15, pages 1–8. New York, NY, USA. ACM.

Sarah Ita Levitan. 2019. *Deception in Spoken Dialogue: Classification and Individual Differences*. Ph.D. thesis, Columbia University.

Sarah Ita Levitan, Angel Maredia, and Julia Hirschberg. 2018a. Acoustic-prosodic indicators of deception and trust in interview dialogues. In *Interspeech*, pages 416–420.

Sarah Ita Levitan, Angel Maredia, and Julia Hirschberg. 2018b. Linguistic cues to deception and perceived deception in interview dialogues. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1941–1950. New Orleans, Louisiana. Association for Computational Linguistics.

Junyi Jessy Li and Ani Nenkova. 2015. Fast and accurate prediction of sentence specificity. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, pages 2281–2287. AAAI Press,

Shan Lu, Gabriel Tsechpenakis, Dimitris N. Metaxas, Matthew L. Jensen, and John Kruse. 2005. Blob analysis of the head and hands: A method for deception detection. In *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*, pages 20c–20c. IEEE.

Angel Maredia, Kara Schechtman, Sarah Ita Levitan, and Julia Hirschberg. 2017. Comparing approaches for automatic question identification. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (* SEM 2017)*, pages 110–114.

Jaume Masip, Siegfried Sporer, Eugenio Garrido, and Carmen Herrero. 2005. The detection of deception with the reality monitoring approach: A review of the empirical evidence. *Psychology Crime and Law*, 11:99–122.

Ewout H. Meijer and Bruno Verschuere. 2017. Deception detection based on neuroimaging: Better than the polygraph? *Journal of Forensic Radiology and Imaging*, 8:17–21.

Rada Mihalcea and Carlo Strapparava. 2009. The lie detector: Explorations in the automatic recognition of deceptive language. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 309–312. Suntec, Singapore. Association for Computational Linguistics.

Matthew L. Newman, James W. Pennebaker, Diane S. Berry, and Jane M. Richards. 2003. Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin*, 29(5):665–675. PMID: 15272998.

Richard E. Nisbett and Timothy D. Wilson. 1977. Telling more than we can know: Verbal reports on mental processes. *Psychological review*, 84(3):231.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 309–319. Stroudsburg, PA, USA. Association for Computational Linguistics.

James Pennebaker and Laura A. King. 1999. Linguistic styles: language use as an individual difference. *Journal of Personality and Social Psychology*, 77(6):1296–1312.

Verónica Pérez-Rosas, Mohamed Abouelenien, Rada Mihalcea, Yao Xiao, C. J. Linton, and Mihai Burzo. 2015. Verbal and nonverbal clues for real-life deception detection. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2336–2346.

Verónica Pérez-Rosas and Rada Mihalcea. 2015. Experiments in open domain deception detection. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1120–1125. Lisbon, Portugal. Association for Computational Linguistics.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. 2011. The kaldi speech recognition toolkit. Technical report. IEEE Signal Processing Society.

Victoria L. Rubin, Elizabeth D. Liddy, and Noriko Kando. 2006. Certainty identification in texts: Categorization model and manual tagging results, In *Computing Attitude and Affect in Text: Theory and Applications*, pages 61–76, Springer.

The Global Deception Research Team. 2006. A world of lies. *Journal of Cross-Cultural Psychology*, 37(1):60–74. PMID: 20976033.

Gabriel Tsechpenakis, Dimitris Metaxas, Mark Adkins, John Kruse, Judee K. Burgoon, Matthew L. Jensen, Thomas Meservy, Douglas P. Twitchell,

Amit Deokar, and Jay F. Nunamaker. 2005. Hmm-based deception recognition from visual cues. In *2005 IEEE International Conference on Multimedia and Expo*, pages 824–827. IEEE.

Morgan Ulinski, Seth Benjamin, and Julia Hirschberg. 2018. Using hedge detection to improve committed belief tagging. In *Workshop on Computational Semantics beyond Events and Roles. NAACL HLT 2018*.

Aldert Vrij, Ronald Fisher, Samantha Mann, and Sharon Leal. 2006. Detecting deception by manipulating cognitive load. *Trends in Cognitive Sciences*, 10(4):141–142.

Cynthia M. Whissell. 1989. The dictionary of affect in language, *The Measurement of Emotions*, Elsevier, pages 113–131.

Clea Wright, Graham Wagstaff, and Jacqueline Wheatcroft. 2014. Subjective cues to deception/honesty in a high stakes situation: An exploratory approach. *The Journal of Psychology*, 149:1–18.

Lina Zhou, Judee K. Burgoon, Jay F. Nunamaker, and Doug Twitchell. 2004. Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communications. *Group Decision and Negotiation*, 13(1):81–106.

Miron Zuckerman, Bella M. DePaulo, and Robert Rosenthal. 1981. Verbal and nonverbal communication of deception. In *Advances in Experimental Social Psychology*, volume 14, pages 1–59. Elsevier.