Identifying Individual Differences in Gender, Ethnicity, and Personality from Dialogue for Deception Detection

Sarah Ita Levitan & Yocheved Levitan
Columbia University

Guozhen An
CUNY Graduate Center

Michelle Levine
Columbia University

Rivka Levitan
Brooklyn College CUNY

Andrew Rosenberg
Queens College CUNY

Julia Hirschberg
Columbia University

Abstract

When automatically detecting deception, it is important to model individual differences across speakers. We explore the automatic identification of individual traits such as gender, native language, and personality, using acoustic-prosodic and lexical features from an initial non-deceptive dialogue. We also explore predicting success at deception and at deception detection, using the same features.

1 Introduction

Automatic deception detection is a critical area of research for a variety of disciplines, from computational linguistics and psychology to law enforcement, military, and intelligence agencies. One of the challenges in this work is the variation across individuals and cultures in deceptive behavior. In previous work on deception in American English speech, (Hirschberg et al., 2005) developed automatic deception detection procedures trained on spoken cues. These procedures have accuracies 20% better than human judges. While identifying common characteristics of deceivers, they also noticed many individual differences in deceptive behavior, e.g., some subjects raised their pitch when lying, while some lowered it; some tended to laugh when deceiving, while others laughed more while telling the truth. They also discovered that human judges' accuracy in judging deception could be predicted from their scores on the NEO-FFI Personality Inventory (Costa and MacCrae, 1992), suggesting that such simple personality tests might also provide useful information in predicting individual differences in deceptive behavior itself (Enos et al., 2006).

Differences in verbal deceptive behavior in across cultures have been identified by several researchers (Feldman, 1979; Cody et al., 1989), who have found that culture-specific deception cues do exist. (Fornaciari et al., 2013) found that personality information can be successfully used as features for deception detection. In our previous work (Levitan et al., 2015a), we conducted a study of cross-cultural deceptive speech, using a large scale corpus of Mandarin Chinese (MC) and Standard American English (SAE) native speakers. We found that including gender, native language, and personality scores in addition to acoustic-prosodic features improved classification performance.

Such work is promising, but requires ground-truth knowledge of these individual traits. For example, it requires NEO-FFI personality scores, which may be impractical to collect in a real-world deception situation. In this work, we address this problem. Specifically, we aim to answer the following question: How much information can be automatically learned from a short dialogue with a subject? We use a portion of the corpus described in (Levitan et al., 2015b) for this study. This part is an initial dialogue between an experimenter and each subject, a 3-4 minute truthful conversation in which the subject answers simple, open-ended questions. Using this subset, we extract acoustic-prosodic and lexical features, and train classifiers to identify gender, native language (American English or Chinese), and personality. We are also interested in whether we can automatically predict from a short dialogue if a subject will be better or worse than average at detec-

Proceedings of NAACL-HLT 2016, pages 40–44,
San Diego, California, June 12-17, 2016. ©2016 Association for Computational Linguistics

tion deception and at lying. All of this information can be useful for downstream deception detection.

In Section 2, we describe the corpus collection and the methods used for transcription and segmentation of the data. Section 3 describes the features used in classification experiments, and Section 4 presents the results of our classification experiments for gender, native language, personality, success at deception, and success at deception detection. We conclude in Section 5 and discuss future research.

2 Corpus

The Columbia deception corpus (Levitan et al., 2015b) consists of within-subject deceptive and non-deceptive speech from native speakers of SAE and MC, both speaking English. (Native language is defined as spoken at home until age 5.) It includes data from 172 subject pairs — 122.5 hours of speech. To our knowledge, this is by far the largest corpus of cleanly recorded deceptive and non-deceptive speech collected and transcribed, with self-identified truth/lie labels. We employed a fake resume paradigm in which pairs of subjects were recorded playing a lying game, alternating between interviewing their partner and being interviewed about answers to a set of 24 biographical questions. As interviewees, they are instructed to convince their interviewer that everything they say is true. As interviewers, they are told to try to identify when the interviewee is lying and when they are telling the truth. To motivate them, their compensation depends on their ability to deceive while being interviewed, and to judge truth and lie correctly while interviewing. As interviewer, they receive $1 each time they correctly identify an interviewees answer as either lie or truth and lose $1 for each incorrect judgment. As interviewee, they earn $1 each time their lie is judged to be true, and lose $1 each time their lie is correctly judged to be a lie by the interviewer.

We collected demographic data from each subject and administered a NEO-FFI (5 factor) personality test (Costa and MacCrae, 1992), assessing: *Neuroticism, Extraversion, Openness to Experience, Agreeableness Conscientiousness*.

We also collected a 3-4 minute baseline sample of speech from each subject for use in speaker normalization, in which the experimenter asks the subject open-ended questions (e.g., What do you like best/worst about living in NYC?). Subjects are instructed to be truthful in answering. Once both subjects have completed all the questionnaires and we have collected both baselines, they begin the lying game.

Transcripts for the recordings were obtained using Amazon Mechanical Turk (https://www.mturk) (AMT). Three transcripts for each audio segment from different 'Turkers' were obtained, and combined using rover techniques (Fiscus, 1997), producing a rover output score measuring the agreement between the initial three transcripts. For clips with a score lower than 70%, transcripts were manually corrected; we needed to hand correct 9.7% of clips.

Speech was segmented into InterPausalUnits (IPUs) – speech from a single speaker separated by 50ms or more (Hirschberg et al., 2005) – using Praat (Boersma and Weenink, 2002). The silence detection was done using intensity thresholding with Praat. We observed IPU segmentation errors resulting from this method, in which areas of speech were identified as silence and had humans hand-correct the IPU segmentation. The experiments in this paper use the 238 baseline files that have been corrected to date. This consists of 25,424 IPUs, in total about 16 hours of speech.

3 Features

We extracted two feature sets for our machine learning experiments: acoustic-prosodic and lexical. Acoustic-prosodic features are extracted from IPUs. We used Praat to extract the following acoustic-prosodic features from each IPU: f0 min, max, mean, median, stdv, mean absolute slope; intensity min, max, mean, stdv; jitter, shimmer. The first six features are different measures of the fundamental frequency, the physical correlate of pitch. The next four are measures of a correlate of perceived loudness. The last two features are measures of voice quality, variation in vocal fold behavior which leads to listeners' perception of the harshness or creakiness or breathiness of the voice. We also estimated speaking rate by calculating the ratio of voiced to total frames and included this as a feature. All these features have been proposed in the literature on deception as possible indicators of deception (DePaulo et al., 2003). Once we extracted these features at the IPU level, we aggregated them into a feature vector for each speaker by averaging the features across the speaker's IPUs. We did this for all features except for minf0 and maxf0, where we used the min of all minf0 and max of maxf0 as aggregated features.

We used Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2001) to extract the lexical features. LIWC is a text analysis program that analyzes text by computing word counts for 72 linguistic dimensions. LIWC dimensions have been used in many studies to predict outcomes including personality (Pennebaker and King, 1999), deception (Newman et al., 2003), and health (Pennebaker et al., 1997). We performed minor pre-processing of the transcripts before running LIWC, as recommended by the manual (e.g. removing spaces from phrases, such as ”Idon'tknow”).

4 Classification Experiments

Using these features, we experimented with different machine learning models and feature sets to predict the following information about the speakers: gender, native language, personality, success at deception, and success at deception detection. Gender and native language ground truth were collected via demographic forms, and personality scores were obtained from the NEO-FFI. Success at deception is measured by the percentage of lies that the subject tells which are judged as true by the interviewer. Success at deception detection is measured by the percentage of correct judgments that the subject makes as interviewer. To conduct our experiments, we needed to convert the numeric values for personality scores, lies, and judgments into nominal values. For NEO-FFI scores, we used the population norms for personality detailed in (Locke, 2015) to divide subjects into high, average, and low for each of the five personality factors, for male and for female subjects separately. For success at deception and detection, we binned subjects into high and low, using the mean as a cutoff point. For each classification task, we compared the performance of four machine learning algorithms, all using the scikit-learn implementation (Pedregosa et al., 2011): SVM, logistic regression, AdaBoost, Random Forest. All evaluation was done using 10-fold cross validation.

4.1 Gender and Language Identification

We found that the Random Forest classifier performed best for gender and language identification for all three feature sets; therefore only those results are reported here. Table1 displays the gender classification results, measured by accuracy, precision, recall, and f-score. The majority class baseline is 61% for gender (female), and 57% for ethnic origin (SAE). Intuitively, the acoustic-prosodic features are highly predictive of gender. It is interesting that LIWC categories are also somewhat predictive of gender, with an f-score of .76 – despite the fact that all subjects answered almost the same questions. Some of the most useful categories were ‘home’, ‘anxious’, and ‘negative emotion’, which all were predictive of ‘female’. On the other hand, ‘anger’, ‘certainty’, and ‘money’ categories were predictive of male. Note that combining LIWC features with acoustic-prosodic features very slightly improves over using acoustic-prosodic features alone. On the other hand, as shown in Table 2, LIWC features are more predictive than acoustic-prosodic features for native language/ethnicity identification. Combining both feature sets results in the best performance, with an f-score of .78. When analyzing the LIWC categories, we find that the punctuation counts are especially useful for this task. For example, use of ‘apostrophe’ is predictive of English, while dash is predictive of Mandarin. In our transcription, dashes are used to represent false starts, so it makes sense that this is a marker of less fluent speech. Apostrophes were used exclusively for transcribing contractions, which are less commonly used by non-native speakers of English, so it is plausible that apostrophes are predictive on English.

Table 1: Gender classification results

Features	acc	prec	rec	f-score
Prosodic	.94	.94	.96	.95
LIWC	.70	.69	.86	.76
Combined	.94	.94	.98	.96

It is important to note that our subjects are all answering the same questions in their baseline sam-

Table 2: Language classification results

Features	acc	prec	rec	f-score
Prosodic	.65	.64	.40	.49
LIWC	.84	.82	.7	.76
Combined	.81	.88	.68	.78

ples of speech. Therefore, it is quite impressive that LIWC features can predict both gender and language with .76 f-score by capturing the variation in how the questions were answered. We find that gender can be predicted with 94% accuracy and language with 84% accuracy using simple features from 3-4 minutes of conversational speech.

4.2 Personality Identification

We repeated the same experiments for personality detection, this time with a 3-class classification problem (low, med, high) for each of the 5 factors. Table 3 shows the results for AdaBoost and Random Forest, which were the two best performing classifiers for personality. As expected, the three classes are highly unbalanced, with the majority of subjects falling into the average class, and a small percentage in the high and low classes. Because of this, we focus our analysis on comparing the f-scores of the different classifiers and feature sets, in order to obtain a more meaningful comparison of the performance of the classifiers over the max-frequency baseline. As seen in Table 3, the best performing model for each factor is significantly better than the baseline. The baseline classifier always predicts the majority class for each of the five factors, and table shows the f-score of this baseline classifier. The relative improvements in f-score range from 38% (Extroversion) to 64% (Openness to experience), and the absolute improvements in f-score range from 10% to 22%. When we rank the LIWC features that contribute the most to the models, we find interesting results. For N-score, ‘power’ and ‘money’ dimensions are the most useful; for E-score, ‘drives’ and ‘focusfuture’. For O-score, ‘interrogation’ and ‘focus past’ are useful; for A-score, ‘social’ and ‘assent’; and for C-score, ‘work’ and ‘time’ are highest. Most of these are intuitive and show the power of using LIWC features for personality detection. Although these results show significant improvements over the baseline, there is much more to investigate in predicting personality traits. A possible reason for the low performance is that self-reported personality scores might not correlate with personality perceptions of others (including machine learning models). Additionally, we use classes based on population distributions that might not be the best fit for our data. For example, the scores are based on participants of all ages, while our subjects are mostly college students. Our approach might benefit from using personality scores from a student population if these were available.

Table 3: Personality classification results, f-scores

Features	Model	N	E	O	A	C
Prosodic	AB	**.43**	.30	.48	.46	.35
	RF	.39	.33	.45	.40	.33
LIWC	AB	.38	.34	.45	.41	**.42**
	RF	.38	**.36**	**.56**	**.47**	.32
Combined	AB	**.43**	.33	.48	.40	.34
	RF	.41	**.36**	.53	**.47**	.39
Baseline	–	.28	.26	.34	.32	.27

4.3 Deception and Deception Detection

We repeated the experiments to predict success at lying and at detecting lies simply from the norming data, in which people were presumably telling the truth. The baseline is 53% for success at deceiving, and 52% for success at detecting deception. For deception detection, our best classifier (AdaBoost, prosodic+LIWC) achieved an accuracy of 61%. When we included binned NEO-scores, gender and language as features in addition to the prosodic and LIWC feature sets, we achieved an accuracy of 65%, a 25% relative increase over the majority class baseline and a 13% absolute increase. We find that none of models using any feature sets were able to beat the baseline accuracy for predicting success at lying. Despite previous findings that success at deception is correlated with success at deception detection (Levitan et al., 2015b), our classifiers were only able to predict success at judgments. When we examine the ranked features for deception detection, the most useful features include speaking rate and LIWC categories‘focus present’, and ‘bio’. Perhaps people who are present-focused are more conscientious and therefore better judges of others' behavior. This is supported by our previous findings that conscientiousness is somewhat correlated with success at deception detect deception.

5 Conclusions and Future Work

In this work we have identified acoustic-prosodic and lexical features which can predict gender, native language/ethnicity, and personality using only a short non-deceptive dialogue. We experimented with using all these features to predict success at deception detection and at deception itself, with significant results only for deception detection. In future work we will experiment with different methods of binning personality scores and different ways of modeling deception detection and deception success. We will also evaluate if our prediction of personality, gender, and ethnic background can help in predicting truth and lie in our larger corpus as effectively as ground truth.

Acknowledgments

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

References

Paul Boersma and David Weenink. 2002. Praat, a system for doing phonetics by computer. *Glot international*, 5(9/10):341–345.

Michael J Cody, Wen-Shu Lee, and Edward Yi Chao. 1989. Telling lies: Correlates of deception among chinese. *Recent advances in social psychology: An international perspective*, pages 359–368.

Paul T Costa and Robert R MacCrae. 1992. *Revised NEO personality inventory (NEO PI-R) and NEO five-factor inventory (NEO FFI): Professional manual*. Psychological Assessment Resources.

Bella M DePaulo, James J Lindsay, Brian E Malone, Laura Muhlenbruck, Kelly Charlton, and Harris Cooper. 2003. Cues to deception. *Psychological bulletin*, 129(1):74.

Frank Enos, Stefan Benus, Robin L Cautin, Martin Graciarena, Julia Hirschberg, and Elizabeth Shriberg. 2006. Personality factors in human deception detection: comparing human to machine performance. In *INTERSPEECH*.

Robert S Feldman. 1979. Nonverbal disclosure of deception in urban koreans. *Journal of Cross-Cultural Psychology*, 10(1):73–83.

Jonathan G Fiscus. 1997. A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover). In *Automatic Speech Recognition and Understanding, 1997. Proceedings., 1997 IEEE Workshop on*, pages 347–354. IEEE.

Tommaso Fornaciari, Fabio Celli, and Massimo Poesio. 2013. The effect of personality type on deceptive communication style. In *Intelligence and Security Informatics Conference (EISIC), 2013 European*, pages 1–6. IEEE.

Julia Bell Hirschberg, Stefan Benus, Jason M Brenier, Frank Enos, Sarah Friedman, Sarah Gilman, Cynthia Girand, Martin Graciarena, Andreas Kathol, Laura Michaelis, et al. 2005. Distinguishing deceptive from non-deceptive speech. In *EUROSPEECH*.

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

Sarah Ita Levitan, Michelle Levine, Julia Hirschberg, Nishmar Cestero, Guozhen An, and Andrew Rosenberg. 2015b. Individual differences in deception and deception detection. In *Proceedings of Cognitive 2015*. Cognitive.

Kenneth Locke. 2015. NEO scoring. `http://www.webpages.uidaho.edu/klocke/neo_scoring.htm`. 2016-03-07.

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.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, 12:2825–2830.

James W Pennebaker and Laura A King. 1999. Linguistic styles: language use as an individual difference. *Journal of personality and social psychology*, 77(6):1296.

James W Pennebaker, Tracy J Mayne, and Martha E Francis. 1997. Linguistic predictors of adaptive bereavement. *Journal of personality and social psychology*, 72(4):863.

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71:2001.