Personality Factors in Human Deception Detection: Comparing Human to Machine Performance

Frank Enos, Stefan Benus,* Robin L. Cautin,** Martin Graciarena,*[†] *Julia Hirschberg,* Elizabeth Shriberg*[†§]

*Columbia University, **Manhattanville College, [†]SRI, [§]ICSI

frank@cs.columbia.edu

Abstract

Previous studies of human performance in deception detection have found that humans generally are quite poor at this task, comparing unfavorably even to the performance of automated procedures. However, different scenarios and speakers may be harder or easier to judge. In this paper we compare human to machine performance detecting deception on a single corpus, the Columbia-SRI-Colorado Corpus of deceptive speech. On average, our human judges scored worse than chance — and worse than current best machine learning performance on this corpus. However, not all judges scored poorly. Based on personality tests given before the task, we find that several personality factors appear to correlate with the ability of a judge to detect deception in speech.

Index Terms: deception, deceptive, perception, personality.

1. Introduction

Interest continues to grow in the research community in the detection of deceptive speech. Such work also has important implications for law enforcement and national security. However, despite a fair number of studies (c.f. [7]), relatively little is known about how deception is revealed in the speech signal. How well humans or machines may ultimately perform at the task of detecting deceptive speech remains an open question.

DePaulo [7] catalogs a large number of psychological studies of deception, from a long tradition focused primarily on visual cues. More recently, work has been under way to apply speech technologies and machine learning techniques to a new, cleanly recorded corpus of deceptive speech, the Columbia-SRI-Colorado (CSC) Corpus [9, 3, 8]. Previous research on this corpus has produced two machine learning systems that achieve classification accuracies of 66.4% [9] and 64.0% [8] (see Section 5).

In this paper, we describe a perception study in which judges attempted to classify as deceptive or truthful the interviews that compose the CSC Corpus. The present work examines human performance at classifying the CSC Corpus with respect to two levels of truth/lie judgments. These results contextualize both previous machine learning experiments and future work on this corpus. In addition we present several strong results suggesting that particular personality factors may contribute significantly to a judge's success at classification.

2. Previous Research

A recent meta-analysis [1] examines the results of 108 studies that attempted to determine if individual differences exist in the ability to detect deception. Ability (where chance is 50%) ranged from that of parole officers (40.41%, one study) to that of secret service agents, teachers, and criminals (one study each) who scored in the 64–70% range. The bulk of studies (156) used students as judges; they scored on average 54.22%.

3. The CSC Corpus

The CSC corpus was designed to elicit within-speaker deceptive and nondeceptive speech [9]. Speakers received a financial incentive to deceive successfully, and the instructions were designed to link successful deception to the 'self-presentational' perspective [7]. That is, speakers were told that the ability to succeed at deception indicated other desirable personal qualities.

The corpus comprises interviews of thirty-two native speakers of Standard American English who were recruited from the community and the Columbia University student population in exchange for payment. Interviewees were told that the study sought individuals who fit a profile based on the twenty-five 'top entrepreneurs of America'. Interviewees answered questions and performed tasks in six areas. The difficulty of tasks was manipulated so that interviewees scored too high to fit the profile in two areas, too low in two, and correctly in two. Four target profiles existed so that interviewees' lies were balanced among the six areas.

In the second phase of the study, interviewees were told that their scores did not fit the target profile, but that the study also sought interviewees who did not fit the profile but who could convince an interviewer that they did. They were told that those who succeeded at deceiving the interviewer would qualify for a drawing to receive an additional $100. Interviewees then attempted to convince the interviewer that their scores in each of the six categories matched the target profile. Two kinds of lies are implicit in this context. The 'global lie' is the interviewee's overall intention to deceive with respect to each score. The 'local lie' represents statements in support of the reported score; these statements will be either true or false.[1] The distinction between these types of lie is subtle but important, since interviewees do not always lie at the local level to convey a global lie. For example, an interviewee may truthfully claim that she has lived in New York City her whole life to support her false claim that she scored well on her knowledge of NYC geography. Interviewees indicated whether each statement they made was entirely true or contained some element of deception by pressing one of two pedals hidden beneath the table (one for **TRUTH**, the other for **LIE**); these labels correspond to the lo-

[1] Hirschberg et al. [9] termed these 'big lie' and 'little lie', respectively.

cal lie category. This data was timestamped and synchronized with the speech signal in post-processing. Ground truth was known a priori for the global lie category, since the interviewees' scores on each section were known. The interviews lasted between 25 and 50 minutes, and comprised approximately 15.2 hours of dialogue; they yielded approximately 7 hours of subject speech.

Interviews were digitally recorded using headworn microphones and transferred to disk. They were orthographically transcribed by hand; labels for local lies were obtained automatically from the pedal-press data, those for global lies were annotated during transcription. Several segmentations were created: the implicit segmentation of the pedal presses, which was hand-corrected to align with corresponding sets of statements; word segments, from the automatic alignment of the transcription using a telephone speech recognizer adapted for full-bandwidth recordings; sentence-like units (EARS SLASH-UNITS or SUs [11]), which were hand labeled; 'breath groups' which were identified from ASR word alignments plus intensity and pauses, and subsequently hand-corrected. The corpus thus consists of lexical transcription, global and local lie labels, segmentations, and the speech itself.

4. Methods and Materials

For the current perception study, we recruited thirty-two native speakers of American English from the community to participate in a 'communication experiment' in exchange for payment. Each judge listened to two complete interviews from the CSC corpus that were selected in order to balance the length of interviews as much as possible (i.e., one long, one short) so that judges could complete the task within two hours. Judges were asked to indicate their judgments on both local and global lies for these interviews. They marked local truth/lie via a labeling interface constructed in Praat[2] [4]. Judges indicated their judgments with respect to global truth/lie (that is, the interviewees' claimed score in each section) on a paper form. For one of the two interviews, each judge received a section of training, or immediate feedback, with respect to the correctness of his or her judgments, so that we could test the effect of training on their judgments. Each judge rated two interviewees and each interviewee was rated by two judges.

In order to examine individual differences among judges, prior to the perception task judges were administered the NEO-FFI form, measuring the Costa & McCrae five-factor personality model, a widely used personality inventory for nonclinical populations [5, 6]. Judges next filled out a brief questionnaire that asked if they had work experience in which detecting deception was relevant and, if so, what that experience was. They were also asked questions intended to determine their preconceptions with respect to lying (*How often can you spot a lie in daily life?* and *How often do you think people lie in daily life in order to achieve some undeserved gain, either material or social?*) and asked to respond on a five-point Likert scale.

Next, judges received written and oral instructions on the perception task: the CSC Corpus (Section 3) was described to them in layman's terms; then, the task and method of labeling each section (global lies) and each segment (local lies) was explained.

Each judge received 'training' for one section of one of the interviews judged. The training consisted of immediate feedback via the interface on the correctness of their ratings. Training was balanced: odd-numbered judges received training on the first interviewee and even-numbered judges on the second, in both cases on a section in which the interviewee lied about performance.

Table 1: *Judges' aggregate performance classifying* **TRUTH / LIE**.

Lie Category	**Chance Baseline**	**Mean**[a]	**Median**	**Std. Dev.**	**Min.**	**Max.**
Local	63.87 [b]	58.23	57.42	7.51	40.64	71.48
Global	63.64 [c]	47.76	50.00	14.82	16.67	75.00

[a] Each judge's score is his or her average over two interviews; as percentages.

[b] Guessing **TRUTH** each time.

[c] Guessing **LIE** each time.

After judging two interviews, judges were asked *Did you find it easy to use the interface?* (all judges responded 'yes'). Judges were also asked to rate their confidence on their performance: *In your opinion, how many of the judgments you made today are correct?* Again, judges responded on a five-point Likert scale.

5. Results on Deception Detection

We now consider accuracy by examining each judge's average performance over two interviews, the average performance of two judges on each interviewee, and judges' performance in the context of machine learning results on the corpus. As noted in Section 2, previous studies have shown that most of the population performs quite poorly at the deception detection task. Our study on the CSC Corpus supports this conclusion. Table 1 shows the aggregate performance of judges on both levels of truth/lie[3]. Most notable is that judges perform *worse* than chance on both local and global lies (where chance is understood to mean guessing the majority class for the aggregate data). The data reflect considerable variability among judges, particularly on the level of the global lie, where standard deviation is quite large, and the difference is great between the best and worst performers. Likewise, the low maximum scores on both levels indicate the difficulty of the task.

Previous studies [9, 8] have presented machine learning results in the detection of local lies in SUs in the CSC corpus. Hirschberg et al. [9] report a classification accuracy of 66.4% versus a chance (majority class) baseline of 60.2% when classifying SUs using lexical, acoustic, and subject-dependent features. A study by Graciarena et al. [8] reports an accuracy of 64.0% versus a chance baseline of 60.4%[4] combining acoustic, lexical, and cepstral learners.

Although the present study focuses on pedal-press-defined units, comparison of results with respect to the difference between classification accuracy and baseline serve to relate human performance to current best machine performance. Even given the limitations of the comparison, we interpret the current finding — that humans perform worse than chance on both levels of lie — to suggest that machine learning results are promising. Work is currently under way to perform machine classification of global and local lies with respect to the pedal-press units.

We now consider the question of whether some interviewees are more or less difficult to classify. Although we hesitate to make strong statistical inferences in this respect (since each interview

[2] Here judges labeled segments delimited by interviewee pedal presses, as described in Section 3. They were able to replay sections at will.

[3] No effect was found for the length of the interview.

[4] The discrepancy of 0.2% in the baselines can be attributed to adjustments in the definition of SUs between studies.

Table 2: *Aggregate performance by interviewee.*

Lie Type	Mean[a]	Median	Std. Dev.	Min.	Max.
Local	58.23	58.58	9.44	35.86	87.79
Global	44.83	45.58	17.40	10.00	81.67

[a] Each interviewee's score is the average over two judges; as percentages.

was labeled by only two judges), a comparison of Table 2 with Table 1 suggests directions for future work. Inspection shows that the range of scores on interviewees is greater than that of the range of scores among judges. In addition, these results suggest a greater variance (shown as standard deviation) among interviewees than among judges. And indeed, O'Sullivan and Ekman [12], have found evidence that extraordinarily good human deception detectors pay close attention to individual differences in determining what cues are relevant. We have work currently under way that seeks to identify such differences.

6. Personality Factors and Performance

An important finding of the present study relevant to human performance in detecting deception is a set of strong correlations between three personality factors and performance or other behaviors in the detection of global lies. The five-factor model [5, 6] is an empirically-derived and comprehensive taxonomy of personality traits. It was developed by applying factor analysis to thousands of terms taken from subject self-descriptions, using words found in a standard English dictionary. Five personality dimensions emerged: Openness to Experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. This model and associated measures appear extensively in the psychology literature.

Table 3 displays the correlations of the factors Openness, Agreeableness, and Neuroticism with performance measures and with a measure of the proportion of sections labeled **LIE** by the judge. Table 4 shows regression models constructed on the factors and measures shown in Table 3.[5] We draw the reader's attention to the particularly strong predictive power of the models using the factor Openness, i.e. those for accuracy and F-measure for **LIE**.

The factor Openness measures the degree to which an individual is available to new experience and able to adjust viewpoints; in addition it correlates with intelligence [5]. We hypothesize that this factor enhances the ability of the judge to base labeling decisions on the available data rather than on preconceptions, hence its presence in the models for accuracy and F-measure for **LIE**.

Table 3: *Correlations between personality factors and judge performance at labeling global lies.*

Factor	Measure	Pearson's corr. coef.	p-value
Neuroticism	**Proportion of segments judged LIE**	-0.44	0.012
Openness	**Accuracy**	0.51	0.003
Agreeableness		0.41	0.021
Neuroticism	**F-measure for TRUTH**	0.37	0.035
Agreeableness		0.41	0.019
Openness	**F-measure for LIE**	0.52	0.003

Individuals who score high in Agreeableness tend to be 'compassionate, good natured, and eager to cooperate and avoid conflict' [5]. Initially, then, it seems unintuitive that Agreeableness would be a predictor of success at deception detection. However, an extremely high score in Agreeableness is associated with a pathology known as dependent personality disorder [5]. This pathology manifests itself in extreme attention to the opinions and affective state of others [2]; likewise, the qualities of compassion and eagerness to cooperate entail sensitivity to affect. We hypothesize that it is this sensitivity that enhances the judge's ability to perceive cues to deception. This is consistent with prior evidence [1] that suggests that people who are highly self-monitoring (individuals who are particularly attuned to the impressions and attitudes of others) do well at the deception detection task.

There is an interesting negative correlation between Neuroticism and the proportion of sections labeled **LIE** by judges. We wondered whether this was a function of behavior at the time of labeling, or of the judges' prior expectations that an interviewee would lie. We found, in fact, a negative correlation (Pearson's cor: -0.39, p=0.0277) between Neuroticism and judges' pre-test report of their expectation of the frequency with which people lie in general.[6] This correlation clearly merits further investigation. We speculate that Neuroticism may entail an inflated need to believe that people are generally truthful, since the neurotic individual suffers more than others when faced with upsetting thoughts or negative perceptions [6]. In addition there is a positive correlation between Neuroticism and F-measure for **TRUTH**; this is fairly intuitive, since a bias toward guessing **TRUTH** may well impact a measure that can favor prediction of **TRUTH**.

7. Conclusions and Future Work

We have examined the performance of humans in distinguishing truth from lie in the CSC corpus of deceptive speech. Our findings have important implications for research in machine detection of deceptive speech and for the understanding of human performance on the deception task. One of the best-documented claims in the literature is that the deception detection task is extremely difficult for humans (c.f. [7, 1]), particularly when speech is the only channel of communication available. In the present study, judges per-

[5] Standard assumptions with respect to normality, variance, and absence of covariance among the independent variables were met in the current data. Regression models were subjected to standard diagnostic measures [10] (DFFITS, DFBETAS, Studentized residuals, Cook's D). In each model one or two potentially influential cases were identified, so we applied robust regression techniques [10]: least median of squares, least trimmed squares, and simply removing the suspect points. In all cases, results were comparable, and in some cases better, than the ordinary least squares models reported here. Although our sample represents 32 judges, we feel the size is mitigated by the extremely small p-value for the F-statistic of the R^2 values, except in the case of the model of proportion of lies guessed, where we warn against making strong inferences.

[6] No other correlations between personality factors and judges' priors were found.

Table 4: *Regression models of performance on global lies.*

Proportion of Segments Judged LIE

	Value	*Std. Err.*	*t-value*	*p-value*
(Int.)	0.7092	0.1065	6.6606	0.0000
Neurot.	-0.0056	0.0021	-2.6749	0.0120

Multiple R^2: 0.19 p-value: 0.0120
F-statistic: 7.16, 1 and 30 deg. of freedom

Classification Accuracy

	Value	*Std. Err.*	*t-value*	*p-value*
(Int.)	-0.2508	0.1427	-1.7572	0.0894
Agree.	0.0056	0.0016	3.4713	0.0016
Open.	0.0079	0.0019	4.1929	0.0002

Multiple R^2: 0.48 p-value: < 0.0001
F-statistic: 13.39, 2 and 29 deg. of freedom

F-measure for TRUTH

	Value	*Std. Err.*	*t-value*	*p-value*
(Int.)	-0.0029	0.1224	-0.0237	0.9813
Neurot.	0.0044	0.0018	2.4251	0.0218
Agree.	0.0047	0.0018	2.6686	0.0123

Multiple R^2: 0.31 p-value: < 0.0046
F-statistic: 6.50, 2 and 29 deg. of freedom

F-measure for Lie

	Value	*Std. Err.*	*t-value*	*p-value*
(Int.)	-0.1469	0.1896	-0.7747	0.4446
Open.	0.0101	0.0031	3.2906	0.0026

Multiple R^2: 0.27 p-value: < 0.0026
F-statistic: 10.83, 1 and 30 deg. of freedom

form on average worse than chance. We thus note the success of machine learning methods in predicting deception in the CSC corpus, since results exceed both chance and human performance.

There is also considerable evidence that individual differences must be taken into account in deception detection, whether by humans or machines [12]. This appears to be supported by the variability of our judges' success in detecting individual interviewees in the present study, and supports our own future efforts to model such individual differences in automatic deception detection.

From the point of view of improving human efforts at detection, we are intrigued by evidence that personality variables have an impact on a judge's success. This finding may help to identify good human detectors of deception and point toward ways individuals can be trained to become better detectors. Further, knowledge of what kinds of people are good detectors may lead to better identification of reliable objective cues to deception in speech.

In addition, we are interested in examining the efficacy of cues individuals believe to be predictors of deception. For example Benus et al. [3] found that pauses correlated with *truthful* speech in the CSC corpus. A number of judges in the present study reported using pauses as an indicator of deception. It thus seems possible that the relatively poor performance achieved by human judges can be attributed in part to the discrepancy between strategies used in perceiving and producing deceptive speech. We hope to use the data described here to shed further light on this question.

In future research on deception detection, we will examine the effects of several factors — the impact of training, prior experience, gender, and the type of cues judges reported using in making decisions — on deception detection. We also will continue to study the role of personality variables in deception, as both attributes of deception detectors and of deception producers.

8. Acknowledgments

The authors thank Maureen O'Sullivan for helpful discussions of the paradigm used here. This research was funded in part by NSF IIS-0325399 and the Department of Homeland Security.

9. References

[1] M. Aamodt, H. Custer, "Who Can Best Catch a Liar?", *Forensic Examiner*, 15(1), 6–11, 2006.

[2] American Psychiatric Association, *(DSM-IV-TR) Diagnostic and Statistical Manual of Mental Disorders*, 4th edition, text revision. American Psychiatric Press, Inc. ,Washington, DC, 2000.

[3] S. Benus, F. Enos, J. Hirschberg, E. Shriberg. "Pauses in Deceptive Speech", To appear in *Speech Prosody 2006*, Dresden.

[4] P. Boersma, D. Weenink. Praat: doing phonetics by computer [Computer program], http://www.praat.org/, 2005.

[5] P.T. Costa, R.R McCrae. *Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO-FFI) Professional Manual*, Psychological Assessment Resources, Inc. Odessa, FL, 1992.

[6] P.T. Costa, R.R McCrae. *Personality in Adulthood: A Five-Factor Theory Perspective*, 2nd Edition. Guilford Publications, New York, 2002.

[7] B. M. DePaulo, J. J. Lindsay, B. E. Malone, L. Muhlenbruck, K. Charlton, H. Cooper. "Cues to Deception", *Psychological Bulletin*, 129(1):74–118, 2003.

[8] M. Graciarena, E. Shriberg, A. Stolcke, F. Enos, J. Hirschberg, S. Kajarekar. "Combining Prosodic, Lexical and Cepstral Systems for Deceptive Speech Detection", To appear in *Proc. IEEE ICASSP*, Toulouse, 2006.

[9] J. Hirschberg, S. Benus, J. M. Brenier, F. Enos, S. Friedman, S. Gilman, C. Girand, M. Graciarena, A. Kathol, L. Michaelis, B. Pellom, E. Shriberg, A. Stolcke, "Distinguishing Deceptive from Non-Deceptive Speech", *Proc. Eurospeech*, Lisbon, 2005.

[10] J. Neter, M. Kutner, C. Nachtsheim, W. Wasserman, *Applied Linear Statistical Models*, 4th Ed. Irwin, Chicago, 1996.

[11] NIST. Fall 2004 Rich Transcription (RT-04f) evaluation plan, August 2004. http://www.nist.gov/speech/tests/rt/rt2004/fall/docs/rt04f-eval-plan-v14.pdf.

[12] M. O'sullivan, P. Ekman. The Wizards of Deception Detection. In *The Detection of Deception in Forensic Contexts*, Cambridge University Press, Cambridge, 2004.

[13] A. Stolcke, X. Anguera, K. Boakye, O. Cetin, F. Grezl, A. Janin, A. Mandal, B. Peskin, C. Wooters, J. Zheng. "Further Progress in Meeting Recognition: The ICSI-SRI Spring 2005 Speech-to-Text Evaluation System", *Proc. NIST MLMI Meeting Recognition Workshop*, Edinburgh, 2005.