Human Detection of Deceptive Speech

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

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

frank@cs.columbia.edu

Abstract

We report a perception study in which judges attempted to label as deceptive or truthful the recorded interviews of the Columbia-SRI-Colorado Corpus of deceptive speech. In general, judges performed very poorly at the task, scoring on average worse than chance. We use these results to contextualize 'current best' machine learning performance on this corpus. In addition, we report strong findings that suggest that certain personality factors influence the ability of a judge to detect deception in speech.

1. Introduction

Interest continues to grow in the topic of the detection of deceptive speech. Research in this area has important implications for law enforcement, national security, and basic science. Despite a fair number of studies (c.f. [7]), relatively little is known with much certainty about how deception is (or is not) revealed in the speech signal, and it is an open question as to how well humans or machines may ultimately perform at the task of detecting deceptive speech.

DePaulo[7] catalogs a large number of psychological studies of deception, and indeed, this has been a topic of interest to psychologists for years. More recently, work has been underway to apply machine learning and other statistical techniques to a new, cleanly recorded corpus of deceptive speech, the Columbia/SRI/Colorado (CSC) Corpus of deceptive speech. [9, 3, 8] . 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 judgements. These results help to 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 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 54.22 % on average; police officers did little better.

3. The CSC Corpus

To our knowledge, prior to the collection of the CSC Corpus, no cleanly-recorded, labeled corpus of deceptive and non-deceptive speech existed. The corpus was designed to elicit within-speaker deceptive and non-deceptive speech [9]. Speakers received financial incentive to deceive successfully, and the instructions were designed to link successful deception to what DePaulo [7] calls the 'self-presentational' perspective. That is, speakers were told that the ability to succeed at deception indicated other desirable personal qualities.

The study 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. Speakers were told that the study sought individuals who fit a profile based on the twenty-five 'top entrepeneurs of America'. Speakers answered questions and performed tasks in six areas. The difficulty of tasks was manipulated so that speakers scored too high to fit the profile in two areas, too low in two, and correctly in two. Four target profiles existed so that speakers' lies were balanced among the six areas.

In the second phase of the study, speakers were told that their scores did not fit the target profile, but that the study also sought speakers 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. The interviewer had no knowledge of the speaker's performance at the time of the interview. During the interview, speakers 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 speaker'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]

[1] The distinction between global and local lie is subtle but important, since it is possible that a speaker would tell the truth on the local level in support of a global lie, for example the speaker might claim that she has lived in New York City her whole life (true) in support of the claim of a high score in the NYC geography section (false). Speakers were instructed to indicate whether each statement they made was entirely true or contained some element of deception by pressing pedals hidden beneath the table; one for **TRUTH**, the other for **LIE**; these labels correspond to the local lie category. This data was timestamped and synchronized with the speech signal in post-processing. Ground truth was established a priori for the global lie category, since speakers were instructed to lie or tell the truth with respect to their scores on each section. Here we use the terminology of [3]; Hirschberg et al. [9] employ the terms 'big lie' and 'little lie' to denote the global and local lies, respectively.

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 computer disk. They were orthographically transcribed by hand, and labled for global and local lies using a GUI. Various segmentations were created: 'breath groups' which were detected automatically based on intensity and pauses, and subsequently hand-corrected; sentence-like units (EARS SLASH-UNITS [11]); and the implicit segmentation of the pedal presses, which was hand corrected to align with corresponding sets of statements.

Raw data thus consists of lexical transcription, global and local lie labels, segmentations, and the speech signal itself.

4. Methods and Materials

The present paper reports a perception study conducted using the CSC Corpus. Thirty-two native speakers of American English were recruited from the community to participate in a 'communication experiment' in exchange for payment. Each judge listened to two complete interviews that were selected in order to balance the length of interviews as much as possible (i.e. one long, one short). Judges indicated their judgments with respect to the local lie via a labeling interface constructed in Praat[2] [4]. Judges indicated their judgments with respect to the global lies (that is, the speakers' claimed score in each section) on a paper form. For one of the two speakers, each judge received a section of training, or immediate feedback, with respect to the correctness of his or her judgments. Speakers were assigned to judges so that each judge rated two speakers and each speaker was rated by two judges.

4.1. Pre-judgment questionnaires

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 non-clinical populations [5, 6]. Judges next filled out a brief questionnaire that asked if they had work experience in which detecting deception was relevant, and sought to determine their preconceptions with respect to lying.

4.2. Judges' Instructions

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

4.3. Post-judgment questionnaire

After judging two speakers, subjects were asked *Did you find it easy to use the interface?* (all subjects responded 'yes'). Subjects were also asked to rate their confidence on their performance.

5. Accuracy at Detecting Deception

We consider accuracy from three standpoints: the performance of judges; judges' performance in the context of machine learning results on the corpus; and accuracy at detecting the deception of individual speakers.

Table 1: *Judges' aggregate performance.*

Lie Type	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 speakers; as percentages.
[b] Guessing 'truth' each time.
[c] Guessing 'lie' each time.

5.1. Judges' performance

As discussed in Section 2, previous studies (c.f. [7] and [1]) show that most of the population performs quite poorly at the deception detection task. Our study using the CSC Corpus draws a similar conclusion. Table 1 shows the aggregate performance of judges on both levels of truth. Most notable is that judges perform *worse* than chance in both cases (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 maximum scores on both levels suggest the difficulty of the task.

5.2. Machine learning results in context

Two previously published studies [9, 8] report results using machine learning techniques to detect local lies in sentence-like units 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 sentence-like units 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%[3] using combined 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 approaches are progessing in a promising fashion. Work is now underway to perform machine classification of global lies and of local lies with respect to the pedal-press units.

5.3. Accuracy at classifying individual speakers

We now consider the performance of multiple judges in evaluating the speech of individual speakers. Although we hesitate to make strong statistical inferences in this respect (since each speaker was only labeled by two judges), a comparison of Table 2 with Table 1 suggests directions for future work. Inspection shows that

[2] Here judges labeled segments delimited by speaker pedal presses; see Sec. 3.

[3] 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 speaker.*

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 speaker's score is the average over two judges; as percentages.

the range of scores among speakers is greater than that of the range of scores among judges. In addition, these results suggest a greater variance (shown as standard deviation) among subjects 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 underway that seeks to identify such cues in the speech signal.

6. Personality Factors and Judges' Performance

Possibly the most compelling results of the present study are strong correlations between three personality factors and performance or other behaviors in the detection of global lies.

6.1. The Five Factor Model of personality

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 descriptive terms taken from subject self-descriptions. All terms were words found in a standard English dictionary. Five personality dimensions emerged: Openness to Experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. This model and the associated measures appear extensively in the psychology literature.

6.2. Correlations

Table 3 displays the correlations between the factors Openness, Agreeableness, and Neuroticism with various performance measures and a measure of the proportion of sections labeled **LIE** by the judge (essentially the posterior probability of the judge's having chosen **LIE**).

6.3. Regression Models

Table 4 shows regression models constructed on the factors and measures shown in Table 3[4] 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**.

6.4. Discussion

The factor Openness measures the degree to which an individual is available to new experience and willing to adjust viewpoint or values; 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**.

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 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, that is, 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 labled **LIE** by judges. We wondered whether this was a function of behavior at the time of labeling, or of the judges' prior expectations that a speaker 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[5]. This correlation clearly merits further investigation, but 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 draw several conclusions from the work presented here. The most obvious, and best documented in the deception literature, is that the deception detection task is extremely difficult (c.f. [7, 1]. This is particularly true in the case where speech is the only channel of communication available; in the present study, judges perform on average worse than chance. We are encouraged, then, by the progress shown in machine learning methods on the CSC corpus, since they exceed chance and human performance.

Next, we continue to believe that the best approach to deception detection is one that will take into account individual differences in deceptive behavior. This seems to be supported by the variability of success in detecting individual speakers in the present study, and is supported in the literature [12] and bolstered by conversations with practitioners.

Finally, we are intrigued by the evidence that personality vari-

[4] 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, and we thus 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 of the sample is mitigated by the extremely small p-value for the F-statistic associated with the R^2 values, except possibly in the case of the model of proportion of lies guessed, where we warn against making very strong inferences.

[5] No other correelations between personality factors and judges' priors were found

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

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

Proportion of Segments Guessed '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				

ables have an impact on a judge's success at the task. This finding may help to identify good human detectors of deception and may point towards ways individuals can be trained to become better detectors. In addition, it may provide a key to the computational modeling of relevant features.

A future paper will examine the effects of other factors, including the impact of training, prior experience, gender, and the type of cues judges reported using in making decisions. In addition, we believe that further study is warranted on the impact of personality variables, from the standpoint of both deceiver and detector.

8. Acknowledgements

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?",

[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]. Retrieved from 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", Pyschological Assesment 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, and 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 and P. Ekman The Wizards of Deception Detection. In *The Detection of Deception in Forensic Contexts*, Cambridge University Press, Cambridge, 2004.