Distinguishing Deceptive from Non-Deceptive Speech

Julia Hirschberg, Stefan Benus*, Jason M. Brenier**, Frank Enos*, Sarah Friedman*,*
Sarah Gilman, Cynthia Girand**, Martin Graciarena†, Andreas Kathol†,*
*Laura Michaelis**, Bryan Pellom**, Elizabeth Shriberg†§, Andreas Stolcke†§*

*Columbia University, **University of Colorado, †SRI, §ICSI

julia@cs.columbia.edu

Abstract

To date, studies of deceptive speech have largely been confined to descriptive studies and observations from subjects, researchers, or practitioners, with few empirical studies of the specific lexical or acoustic/prosodic features which may characterize deceptive speech. We present results from a study seeking to distinguish deceptive from non-deceptive speech using machine learning techniques on features extracted from a large corpus of deceptive and non-deceptive speech. This corpus employs an interview paradigm that includes subject reports of truth vs. lie at multiple temporal scales. We present current results comparing the performance of acoustic/prosodic, lexical, and speaker-dependent features and discuss future research directions.

1. Introduction

In recent years, there has been considerable interest in the speech community in the automatic identification of affective speech [5]. Promising research is currently underway using corpus-based machine learning approaches to identify frustration and anger in spoken dialogue systems and call centers; to determine whether students using automatic tutoring systems are confident or uncertain; or, more generally, to decide whether speakers are expressing 'positive' or 'negative' emotions[10, 2, 3, 11]. Other categories of speaker state have also been the subject of considerable interest, particularly in the psychological literature. One such state is that of deception, which has been associated with manifestations of both fear and elation [7]. However, studies of deceptive behavior have so far primarily involved human perception studies or descriptive analyses of facial, gestural, and biometric data rather than acoustic information. Corpus-based, machine learning approaches to detecting any aspect of deception have yet to be undertaken. In particular, such approaches have not been possible in the spoken domain, since corpora to support such studies have not been available.

In this paper we describe a corpus-based, machine learning approach to the detection of deceptive speech, using acoustic/prosodic, lexical, and subject-dependent cues. In Section 2 we briefly discuss the current state of research on the characteristics of deceptive speech. In Section 3, we describe the collection of the Columbia/SRI/Colorado (CSC) Corpus of deceptive speech for training and testing. In Section 4, we discuss the features we have extracted for analysis and in Section 5, we describe our machine learning approach and current results. We conclude with a discussion of further work (Section 6).

2. Previous Research

Distinguishing deceptive from non-deceptive speech automatically is of considerable practical interest, especially to law enforcement and other government agencies — to identify potential deception at border crossings and in military scenarios in the field and elsewhere, and to evaluate reports from informants at embassies and consulates throughout the world. To date, the main research efforts in studying deceptive behavior have been pursued in the psychological literature, where the focus has largely been on identifying visual cues (body and facial gestures) through laboratory experiments (Cf. [6] for a review of much of this literature.)

A few studies have included audio analysis: Ekman et al. [8] found a significant increase in pitch for deceptive speech over truthful speech. Streeter et al. [18] reported similar results, with stronger findings for more highly motivated subjects. DePaulo et al., in their meta-study of previous research findings in deception [6], reported significant effects for increased pitch and vocal tension in their overall examination of evidence of subject 'tenseness' during deception. There is also some literature and much lore among members of law enforcement agencies and the military identifying auditory and lexical cues to deception, the most widely followed including: response latency, filled pauses, coherence of discourse, passive voice, and use of contractions [1, 17]. Voice stress analysis procedures attempt to rely upon low level indicators of stress as indirect indicators of deception [9], and commercial systems promise to distinguish truth from lie — or love from indifference — with little independent evidence of success. However, despite some evidence from the research community and belief among practitioners, there has been little work on the automatic identification of deceptive speech from such acoustic, prosodic, and lexical cues.

3. The CSC Corpus

One of the primary obstacles to research on the automatic detection of deceptive speech is the lack of a cleanly-recorded corpus of deceptive and non-deceptive speech to use for training and testing. Existing corpora are difficult to analyze due to poor recording conditions. While early studies were better able to utilize scenarios with 'high stakes' deception (in which subjects could be motivated by fear or shame) in the laboratory [12], more recent studies have been limited to less stressful scenarios by human subjects protocols and privacy considerations. In these studies subjects are motivated to deceive primarily by financial reward.

Our collection paradigm was designed to elicit within-subject deceptive and non-deceptive speech from subjects who

had both financial incentive and motivation in terms of what De-Paulo [6] terms the 'self-presentational' perspective to do well at deception. Thirty-two native speakers of Standard American English were recruited for the study. They were asked to perform a series of tasks (activities and question-answering) in six areas. They were told that their performance would be compared to a target profile based on a survey of the twenty-five 'top entrepreneurs of America' performing similar tasks, results of which they would be shown later. Task difficulty was manipulated so that subjects scored more poorly than the target in two task areas, better than the target in two others, and the same in another two of the six; this manipulation was balanced across task categories.

In the next phase of the experiment, subjects were shown their own score and the target, which were invariably quite different in four areas. They were told that the study's actual goal was to compare people who have certain skills and knowledge with people who are good at convincing others that they do. They were told that they could continue to the second stage of the study and also be eligible for a $100 prize if they could convince an interviewer that, instead of scoring as they had, they had in fact performed just as the target entrepreneurial profile. Thus, each subject was motivated to tell the truth in two task areas and to deceive the interviewer in four others. They were told that the interviewer had no knowledge either of the target profile or of their performance (the latter true). The interviewer's task was to determine how he thought the subjects had actually performed, and he was allowed to ask them any questions other than those that were actually part of the tasks they had performed. Finally, for each question, subjects were asked to indicate whether the reply was factually true or contained any false information by pressing one of two pedals hidden from the interviewer under the table.

The interviews, which lasted between 25 and 50 minutes, comprised 15.2 hours of interviewer/subject dialogue and yielded approximately 7 hours of subject speech. They were recorded to digital audio tape on two channels using a Crown CM311A Differoid headworn close-talking microphone and downsampled to 16kHz. They were subsequently orthographically transcribed and sentence-like units (EARS SLASH UNITS [14]) were labeled. The transcription was then automatically aligned with the audio data. Currently we have divided the data into word, slash unit, 'breath group' (phrasal units determined automatically from pause and intensity and subsequently hand-corrected), and turn units, by combining automatic procedures and hand transcriptions. There are 79,488 words, 9687 SUs[1], 19,170 breath groups, and 3882 SPEAKER TURNS. Breath group and speaker turn units were derived semi-automatically. The former were inferred from pausal information in the alignment files plus intensity features; some spurious breaks were subsequently removed by hand. The latter were derived by comparing the two channels of the recordings.

To our knowledge, this corpus is the first audio corpus recorded under conditions that will permit sophisticated speech analyses to be performed (e.g. to extract pitch, intensity, and voice quality characteristics). It is also unusual in including ground-truth information indicated by subjects on a per-turn basis, as well as knowledge of when subjects were trying to deceive the interviewer about how they had performed on the task area as a whole.[2] We term this distinction 'Little Lie' versus

'Big Lie' below. The unit divisions into Little Truth and Lie are: words (47,188/31,200), SUs (5709/3782), breath groups (11,612/7180), and turns (2230/1573). The Big Truth/Lie divisions are words (22,435/52,894), SUs (2748/6184), breath groups (5362/11,966), and turns (1092/2358). So, for example, there are 2358 turns in which speakers are trying to persuade the interviewer they performed a task differently from their actual performance, and 1573 turns in which they made factually inaccurate statements according to their self-report.

4. Feature Extraction and Analysis

Previous research and practitioner experience suggest that acoustic/prosodic and lexico/syntactic cues may signal that speakers when speakers are deceptive. While some of these cues are proposed as general, at least within a culture, there is also some evidence from diverse findings for phenomena such as pitch variation and disfluency production [6] that there is considerable individual variation as well. Deceptive speakers, for example, may raise or lower their pitch significantly from their normal speaking style. This variation is not surprising, since it is well known to characterize the conveyance of particular emotions, such as anger and frustration. Below we describe lexical and acoustic/prosodic cues we have found in our corpus as a whole. In Section 5 we will discuss speaker-dependent cues that we have explored.

4.1. Lexical Cues to Deception

It has often been noted in the literature and by practitioners that speakers who intend to deceive have different patterns of word usage than speakers who are telling the truth ([13], [16], [22]); thus, analysis of lexical content can be used to predict whether or not a speaker is being deceptive or truthful. We have performed lexical analysis of the CSC corpus using a lexical categorization program, *Linguistic Inquiry and Word Count* (LIWC) ([15]). This program classifies words in a text according to textual, semantic, and syntactic categories shown to be relevant for predicting subjects' deceptive intent (e.g., positive-emotion words, words denoting cognitive activity, prepositions, pronouns). Of the 68 LIWC categories examined, preliminary results suggest that the presence of positive emotion words is the best indicator of deception; deceptive speech has a greater proportion of positive emotion words than does truthful speech (p = .0074). Other categories which also appeared promising but will require further analysis are those of word count and lexical items relating to causation.

Another claim in the literature is that FILLED PAUSES (e.g. *um*, *uh*) signal discomfort with a topic or signal the beginning of a deceptive utterance [19, 20]. Our corpus contains 3614 filled pauses. Surprisingly, they correlate more strongly with truthful than with deceptive speech in the Little Lie/Truth condition, with $\chi^2(1, N = 78, 396) = 26.03, p = .0001$. This may be consistent with suggestions by practitioners [17] that deceptive speech is more careful or planned. In the Big Lie/Truth condition the difference is not significant.

Another tool for lexical analysis that we have used focuses on the emotive content of speech, a feature that our early experiments with LIWC have suggested is a promising one for deception detection. This second tool is Whissell's *Dictionary of Affect in Language* (DAL) [21]. Its focus is more narrow than LIWC in that it only addresses the emotional connotation

---

[1]These figures include 196 segments from the introductory sections of the interviews, which are labeled neither truth nor lie.

[2]For example, claiming to have scored well on the interactive task

section (false), the subject might report her years of having played varsity basketball (true) as supporting evidence.

of words along the three dimensions of pleasantness, activation, and imagery. The dictionary assigns ratings along these dimensions, using a continuous scale that ranges from 1 to 3 (determined by human judgment), rather than classifying entries categorically, as is done in LIWC. With 8742 entries, which were determined by general corpus frequency, it is designed to cover about 90% of an average English text.

Based on the premise that different emotional states experienced during deceptive vs. truthful speech are reflected in word choice, we investigated the distribution of DAL scores calculated for the three dimensions. Preliminary findings appear to point to pleasantness as the most promising factor in predicting deception. Considering DAL scores on a per-SU basis, we have found that the minimum pleasantness score per SU appears to be predictive of deception. Specifically, for each one unit increase in minimum pleasantness score an utterance is 1.20 times more likely to be deceptive (p = .001). When controlling for SU length, for each one unit increase in the average pleasantness score, an utterance is 1.29 times more likely to be deceptive (p = .001) and for each one unit increase in the pleasantness standard deviation, an utterance is 54% less likely to be deceptive (p = 0). A somewhat weaker finding is that for each one unit increase in maximum pleasantness score, an utterance is 23% less likely to be deceptive (p = .085). No significant effect was found for the imagery or activation dimensions.

4.2. Acoustic and Prosodic Features

Observations in the literature suggest that pitch, energy, speaking rate, and other stylistic factors (e.g. "muffled" voice) vary when speakers deceive. Our features attempt to capture this variation as well as to explore other potential cues. We consider wide range of potential acoustic and prosodic features, taking advantage of tools available from automatic speech recognition, to extract and model features including durational, pausing, intonational, and loudness, associated with multiple time scales, from a few milliseconds to an entire speaker turn. Features are automatically normalized, taking into account long-term speaker-specific habits as well as segmental context.

To extract these features, the speech was first segmented into SUs by chopping at punctuation marks (ellipses, periods and question marks) in the hand-transcribed corpus. For each SU, we computed roughly 150 prosodic features involving pitch, energy, and duration patterns. Pitch and energy were obtained from the ESPS/Waves pitch tracker *get_f0*; duration features were obtained via forced alignment of hand transcripts using the SRI automatic speech recognition system. Pitch features were computed from the voiced regions in the SU, and were then used in one of three forms: raw, median-filtered, or stylized using an approach that fits linear splines to the median-filtered pitch. From these pitch sequences we computed a large set of features, including maximum pitch, mean pitch, minimum pitch, range of pitch number of frames that are rising/falling/doubled/halved/voiced, length of the first/last slope, number of changes from fall to rise, and value of first/last/average slope. Features were normalized by five different approaches: no normalization, divide by the mean, subtract the mean, and *z-scores* (subtract the mean and divide by the standard deviation). Two basic energy features were computed. The first was the raw energy in the SU and the second was the raw energy only of the voiced regions. The second feature type was used in one of three forms: raw, median-filtered, or stylized using the linear spline approach. From these values we computed several derived features, including the maximum,

energy, minimum energy, mean energy, and other features similar to those just mentioned for pitch. Finally several duration features were computed. The maximum and the average phone duration in the SU were first computed. They were then used either as raw values, normalized using speaker specific durations or normalized using durations computed from the whole corpus. The normalization was applied by dividing by the mean or subtracting by the mean and dividing by the variance.

5. Machine Learning Experiments

We have performed machine learning experiments on lexical, acoustic/prosodic, and speaker-dependent features, using the *Ripper* rule-induction classifier [4]. Experiments described below were performed on the 9491 SUs in our corpus for which we have Little Lie/Truth labels. The baseline error for this task, when we predict the majority class of true, is 39.8%. While the literature suggests that speakers vary in their styles of deception, results here pool all our data, although we will reconsider speaker dependencies below. We divided the data 90%/10% into training and test sets five times (with replacement), trained on the former and tested on the latter, and then averaged the results for the numbers presented below.

We first examined the usefulness of the acoustic/prosodic features described in Section 4.2 in distinguishing deceptive from non-deceptive speech. Results for our Acoustic/Prosodic feature-set averaged over our test sets were 38.5% error — only slightly below the baseline. The rules which proved important in this model included energy and f0 features.

We next considered the predictive power of lexical features, including some of those described in Section 4.1, and others derived from prior descriptive analyses and practitioner observations, as indicators of deceptive speech. Our Lexical feature-set includes simple part-of-speech and word features, such as the presence of different types of pronouns, cue phrases, contractions, verb tense, and particular phrases (e.g. "I did not"), all observed to vary importantly in Statement Analysis [1], and a simple bag-of-words feature. We also looked at positive and negative emotion words, as described in Section 4.1, for whether or not the utterance was a syntactic question or a question following an interviewer question, for the number of words repeated from the interviewer's previous query, indicating hedging behavior [17], and for the presence of disfluencies (self-repairs or filled pauses) or laughter in the phrase [20, 7, 17]. Averaged error over our five test sets was also around the baseline at 39.0%. Features used in the rule-sets produced from these experiments included the number of words repeated from the interviewer's queries, verb tense, the presence of filled pauses, and the presence of the specific word '*poor*' in the SU.

We next considered whether a combination of lexical and acoustic features might perform better than either feature-set alone. Indeed, the error on the test set is reduced to 37.2% when we use all of the lexical and acoustic features described above. Compared to a baseline of 39.8% however, this improvement is still rather modest. In all of the rulesets produced in this experiment, the acoustic features dominate over the lexical markedly.

Because of the intuitions among researchers and practitioners that deceptive speech exhibits a high degree of individual variation, we next considered a speaker-dependent feature-set, to test this hypothesis. Our Speaker-Dependent feature-set includes subject id, subject gender, and a number of subject-dependent ratios, designed to capture subjects' lexical habits, including ratios of filled pauses in Little Lie and Little Truth conditions for a given speaker, and similar ratios of laughter

and cue phrase use, as well as the ratio of phrases containing filled pauses to all phrases spoken by that speaker, and similar ratios of laughter and cue phrase use. When we include this feature-set with our Acoustic/Prosodic and Lexical feature-sets, we do indeed see a considerable reduction in error, from 39.8% to 33.6% averaged over the five test sets. Sample rule-sets from these experiments show that speaker-dependent filled pause and cue phrase ratios, alone or combined with acoustic energy and pitch features, produce this reduction in error. These results support the hypothesis that deceptive behavior in speech is an individualized phenomenon.

6. Conclusion

In this paper we have described experiments in distinguishing deceptive from non-deceptive speech in the CSC Corpus, a data collection designed to elicit within-speaker deceptive and non-deceptive speech. Preliminary analyses of this data indicate several statistically significant differences between the two classes of speech: differences in the use of positive emotion words, of filled pauses, and in a pleasantness score calculated using Whissel's Dictionary of Affect. In pursuit of our primary goal of automatic modeling of deceptive speech, we have also examined three feature-sets for use in machine learning experiments, including acoustic/prosodic, lexical, and speaker-dependent sets of features. While neither the acoustic/prosodic nor the lexical feature sets currently perform much above the baseline, their combination does exhibit improved performance. The most dramatic improvement, however, comes with the addition of speaker-dependent features, which reduces the baseline error by over 6%. Our future work will focus not only on exploring new features, such as voice quality features and additional lexical features, but on developing clustering and feature-selection techniques which can address the speaker-dependent nature of deceptive speech.

7. Acknowledgements

8. References

[1] S. H. Adams. Statement analysis: What do suspects' words really reveal? *FBI Law Enforcement Bulletin*, October 1996.

[2] J. Ang, R. Dhillon, A. Krupski, E. Shriberg, and A. Stolcke. Prosody-based automatic detection of annoyance and frustration in human-computer dialog. In *Proc. ICSLP 2002*, pages 2037–39, Denver, 2002.

[3] A. Batliner, R. Fischer, R. Huber, J. Spilker, and E. Nöth. How to find trouble in communication. *Speech Communication*, 40:117–143, 2003.

[4] W. A. Cohen. Fast effective rule induction. In *Proc. Twelfth International Conference on Machine Learning*, 1995.

[5] R. Cowie and E. Douglas-Cowie and N. Campbell. *Speech Communication: Special Issue on Speech and Emotion*. 40, 2003.

[6] 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.

[7] P. Ekman. *Telling lies: clues to deceit in the marketplace, politics, and marriage*. Norton, New York, 1992.

[8] P. Ekman, M. Sullivan, W. Friesen, and K. Scherer. Face, voice, and body in detecting deception. *Journal of Nonverbal Behaviour*, 15(2):125–135, 1991.

[9] D. Haddad and R. Ratley. Investigation and evaluation of voice stress analysis technology, March 2002. National Criminal Justice Reference Service, http://www.ncjrs.org/pdffiles1/nij/193832.pdf.

[10] C. M. Lee and S. S. Narayanan. Combining acoustic and language information for emotion recognition. In *Proc. ICSLP 2002*, pages 873–76, Denver, 2002.

[11] D. Litman and K. Forbes-Riley. Predicting student emotions in computer-human dialogues. In *Proc. ACL-2004*, Barcelona, 2004.

[12] A. Mehrabian. Nonverbal betrayal of feeling. *J. Experimental Research in Personality*, 5:64–73, 1971.

[13] M. L. Newman, J. W. Pennebaker, D. S. Berry, and J. M. Richards. Lying words: Predicting deception from linguistic style. *Personality and Social Psychology Bulletin*, 29:665–675, 2003.

[14] 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.

[15] J. W. Pennebaker, M. E. Francis, and R. J. Booth. *Linguistic Inquiry and Word Count*. Erlbaum Publishers, Mahwah, NJ, 2001.

[16] T. Qin, J. K. Burgoon, and J. F. Nunamaker. An Exploratory study on promising cues in deception detection and application of decision tree. In *Proc. 37th Annual Hawaii International Conference on System Sciences*, pages 23–32, 2004.

[17] J. E. Reid and Associates. *The Reid Technique of Interviewing and Interrogation*. Reid, John E. and Associates, Inc., Chicago, 2000.

[18] L. A. Streeter, R. M. Krauss, V. Geller, C. Olson, and W. Apple. Pitch changes during attempted deception. *Journal of Personality and Social Psychology*, 35(5):345–350, 1977.

[19] J. F. F. Tree. Interpreting pauses and ums at turn exchanges. *Discourse Processes*, 34:37–55, 2002.

[20] A. Vrij and F. W. Winkel. Cultural patterns in Dutch and Surinam non-verbal behavior: Analysis of simulated police/citizen encounters. *Journal of Nonverbal Behavior*, 15:169–184, 1991.

[21] C. Whissell. The Dictionary of Affect in Language. In R. Plutchnik and H. Kellerman, editors, *Emotion: Theory, Research and Experience*, pages 113–131. Academic Press, New York, 1989.

[22] L. Zhou, J. K. Burgoon, D. P. Twitchell, T. Qin, and J. F. Nunamaker. A Comparison of classification methods for predicting deception in computer-mediated communication. *Journal of Management Information Systems*, 20(4):139–165, 2004.