Deceptive Speech: Acoustic, Prosodic Lexical, Gender, Ethnicity, and Personality Factors

Julia Hirschberg

Computer Science

Columbia University

14 June 2016
Collaborators

• Colleagues: Sarah Ita Levitan, Guozhen An, Michelle Levine, Andrew Rosenberg, Yocheved Levitan, Bingyan Hu

• And thanks to: Zoe Baker-Peng, Lingshi Huang, Leighanne Hsu, Melissa Kaufman-Gomez, Rivka Levitan, Gideon Mendels, Yvonne Missry, Elizabeth Petitti, Sarah Roth, Molly Scott, Jennifer Senior, Grace Ulinski, Christine Wang, Mandi Wang
Goals

• Identify acoustic, prosodic and lexical characteristics of deceptive vs. non-deceptive speech
  – How do these differ by gender, ethnicity, and personality profile?
  – Taking these features into account, can we distinguish lies from truth with accuracy better than current technology? Better than humans?
Defining Deception

• Deliberate choice to mislead
  – *Without prior notification*
  – To gain some *advantage* or to avoid some *penalty*

• *Not*:
  – Self-deception, delusion, pathological behavior
  – Theater
  – Falsehoods due to ignorance/error
Why is deception difficult?

• Hypotheses:
  – Our cognitive load is increased when lying because…
    • Must keep story straight
    • Must remember what we have said and what we have not said
  – Our fear of detection is increased if…
    • We believe our target is hard to fool or suspicious
    • Stakes are high: serious rewards and/or punishments
  – These make it hard for us to control indicators of deception (e.g. in speech)
Cues to Deception in Many Modalities

- **Body posture and gestures** (Burgoon et al ‘94)
  - Complete shifts in posture, touching one’s face,…
- **Microexpressions** (Ekman ‘76, Frank ‘03)
  - Fleeting traces of fear, elation,…
- **Biometric factors** (Horvath ‘73)
  - Increased blood pressure, perspiration, respiration, odor…other correlates of stress → polygraph
- **Changes in brain activity:** true vs. false stories
- **Variation in what is said and how** (Adams ‘96, Pennebaker et al ‘01, Streeter et al’77, Hirschberg et al ‘05, Levitan et al ‘16)
Current Approaches to Deception Detection

- Training humans
  - John Reid & Associates
    - Behavioral Analysis: Interview and Interrogation
- Laboratory studies: Production and Perception
- ‘Automatic’ methods
  - Polygraph
  - Nemesysco and the Love Detector
  - No evidence that any of these work....but publishing this statement can be dangerous! (Anders Eriksson and Francisco La Cerda)
Our Research

• Conduct *objective, experimentally verified* studies of cues to deception which predict better than *humans* or *polygraphs*

• *Our method:*
  – Collect speech data and extract *acoustic, prosodic, and lexical cues* automatically
  – Take *gender, ethnicity, and personality factors* into account
  – Use *Machine Learning* techniques to train models to classify deceptive vs. non-deceptive speech
Our Previous Work on Deception

• Created **Columbia/SRI/Colorado Deception Corpus (LDC2013S09)**
  – Interviews eliciting deceptive and non-deceptive speech (15.2h; 7h of speech from 32 adult native speakers of Standard American English)

• **Machine learning classifiers** identified deceptive vs. non-deceptive speech with **70% accuracy** using acoustic, prosodic, and lexical features
  – Out-performed humans (58.2% acc) on same data
  – **Best human judges** scored high in NEO-FFI personality inventory in agreeableness and openness to experience
  – Could deceptive behavior also be predicted from personality?
A Meta-Study of Human Deception Detection  
(*Aamodt & Mitchell 2004*)

<table>
<thead>
<tr>
<th>Group</th>
<th>#Studies</th>
<th>#Subjects</th>
<th>Accuracy %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Criminals</td>
<td>1</td>
<td>52</td>
<td>65.40</td>
</tr>
<tr>
<td>Secret service</td>
<td>1</td>
<td>34</td>
<td>64.12</td>
</tr>
<tr>
<td>Psychologists</td>
<td>4</td>
<td>508</td>
<td>61.56</td>
</tr>
<tr>
<td>Judges</td>
<td>2</td>
<td>194</td>
<td>59.01</td>
</tr>
<tr>
<td>Cops</td>
<td>8</td>
<td>511</td>
<td>55.16</td>
</tr>
<tr>
<td>Federal officers</td>
<td>4</td>
<td>341</td>
<td>54.54</td>
</tr>
<tr>
<td>Students</td>
<td>122</td>
<td>8,876</td>
<td>54.20</td>
</tr>
<tr>
<td>Detectives</td>
<td>5</td>
<td>341</td>
<td>51.16</td>
</tr>
<tr>
<td>Parole officers</td>
<td>1</td>
<td>32</td>
<td>40.42</td>
</tr>
</tbody>
</table>
Current Study: Hypotheses

• Ethnicity, gender and personality factors play a role in deceptive behavior as well as deception detection

• New corpus: Pairs of Standard American English and Mandarin Chinese native speakers (E/E, E/M, MM), male and female, speaking English, interviewing each other
Experiment
Experiment

Background survey
Biographical Questionnaire
Baseline sample
Lying game
Survey
**Biographical Questionnaire**

**Participant No.______**  
**Date______**

**Instructions**

Please carefully look through the questions. Write down the true answer to each question in the "True Answer" column. When you have finished that, for all the questions that have don't have "X's in the "False Answer" column, make up an answer. Consult the additional sheet you have been given. You want to choose a lie that you are not as familiar with as the true answer.

This experiment is completely anonymous—your name will never be linked to the data.

<table>
<thead>
<tr>
<th>No.</th>
<th>Questions</th>
<th>True Answer</th>
<th>False Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Where were you born?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>How many years did you live in your first home?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>What is your mother's job?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>What is your father's job?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Have your parents divorced?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Have you ever broken a bone?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Do you have allergies to any foods?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Have you ever stayed overnight in a hospital as a patient?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Have you ever tweeted? (posted a message on twitter)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Have you ever bought anything on eBay?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Do you own an e-reader of any kind?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Who was the last person you were in a physical fight with?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Have you ever gotten into trouble with the police?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Who ended your last romantic relationship?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Whom do you love more, your mother or father?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>What is the most you have ever spent on a pair of shoes?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>What is the last name do you say that you really hated?</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
NEO-FFI

OCEAN

openness to experience

conscientiousness

extraversion

agreeableness

neuroticism
Experiment
Experiment
Example: “Where were you born?”

True or False?
Example: “Where were you born?”

False!
Annotation

• Transcribe using crowd-sourcing (Amazon Mechanical Turk)
  – 5 Turkers per utterance, combined using Rover techniques
• Automatically segmented using Praat into Inter-Pausal Units (IPUs) at 50ms silence
• Automatically aligned with speech and truth/lie labels using ASR aligner
# Rover Example

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>its</th>
<th>really</th>
<th>fun</th>
<th>um</th>
<th>I</th>
<th>go</th>
<th>like</th>
<th>to</th>
<th>a</th>
<th>place</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>its</td>
<td>really</td>
<td>fun</td>
<td></td>
<td></td>
<td></td>
<td>go</td>
<td>like</td>
<td></td>
<td>a</td>
<td>place</td>
</tr>
<tr>
<td>3</td>
<td>it’s</td>
<td>really</td>
<td>fun</td>
<td>um</td>
<td>I</td>
<td></td>
<td>go</td>
<td>like</td>
<td>to</td>
<td>a</td>
<td>place</td>
</tr>
<tr>
<td>Rover output</td>
<td>its</td>
<td>really</td>
<td>fun</td>
<td>um</td>
<td>I</td>
<td>go</td>
<td>like</td>
<td>to</td>
<td>a</td>
<td>place</td>
<td></td>
</tr>
<tr>
<td>score</td>
<td>l</td>
<td>l</td>
<td>l</td>
<td>2/3</td>
<td>2/3</td>
<td>l</td>
<td>l</td>
<td>2/3</td>
<td>l</td>
<td>l</td>
<td>l</td>
</tr>
</tbody>
</table>

\[
\text{ROVER Score} = \frac{(1+1+1+2/3+2/3+1+1+2/3+1+1)}{10} = 0.9
\]
Corpus

• Two current versions:
  – Balanced corpus: 140 pairs; 112h speech
  – Larger corpus: 172 pairs; 122h speech
  – *Largest existing cleanly recorded corpus of deceptive/non-deceptive speech*
Correlations: Deceptive Ability with Deception Detection

- People’s ability to detect deception correlates with their ability to deceive \( r(280) = 0.12, p = 0.05 \)
  - Ability to lie \( \rightarrow \) as interviewee, \# lies believed true by interviewer
  - Ability to detect lies vs. truth \( \rightarrow \) as interviewer, \# correct guesses for truth or lie

- Holds across all subjects but strongest for **females**
  \( r(140) = 0.24, p = 0.005 \)
Correlations: Gender, Ethnicity, Personality with Ability to Deceive

• *Extraversion* is significantly *negatively correlated* with ability to deceive for English males $r(70) = -0.24$, $p = 0.04$

• Tendencies for
  – Chinese/female *extraversion* *positively* correlated with ability to deceive
  – American/female *conscientiousness* *negatively* correlated with ability to deceive
Gender, Ethnicity, Personality & Deception Detection

- *No effect of personality factors in deception detection* found so far
  - Contra earlier findings for English speakers (Enos et al ’06)
  - But this is real-time detection vs. later judgments
  - Working on better ways representations
    - Clustering using all factors
    - Quantizing raw scores: high, medium, low
Confidence in Judgments

• Ability to detect deception negatively correlates with confidence in judgments for all subjects $r(278) = -0.15, p = 0.01$
  – Strongest for females $r(140) = -0.25, p = 0.003$

• Ability to deceive negatively correlated with confidence for males $r(138) = -0.17, p = .05$
  – Strongest for Chinese males $r(70) = -0.27, p = 0.02$

• Less confident interviewers may ask more follow-up questions and obtain more evidence for decisions?
Personality Factors and Confidence

- Neuroticism *negatively* correlates with confidence for Chinese female subjects $r(70) = -0.29$, $p = 0.02$
- Openness to experience *negatively* correlates with confidence for all subjects $r(277) = -0.14$, $p = 0.02$
  - Strongest for females $r(140) = -0.20$, $p = 0.02$
  - Strongest for Chinese females $r(70) = -0.29$, $p = 0.02$
Acoustic/Prosodic Features

- **Duration** features
  - Phone / Vowel / Syllable Durations
  - Normalized by Phone/Vowel Means, Speaker
- **Speaking rate** features (vowels/time)
- **Pause** features (cf Benus et al ‘06)
  - Speech to pause ratio, number of long pauses
  - Maximum pause length
- **Energy** features (RMS energy)
- **Pitch** features
  - Pitch stylization (Sonmez et al. ‘98)
  - Model of F0 to estimate speaker range
  - Pitch ranges, slopes, locations of interest
- **Spectral tilt** features
Lexical Features

- Presence and # of filled pauses
- Is this a question? A question following a question
- Presence of pronouns (by person, case and number)
- A specific denial?
- Presence and # of cue phrases
- Presence of self repairs
- Presence of contractions
- Presence of positive/negative emotion words
- Verb tense
- Presence of ‘yes’, ‘no’, ‘not’, negative contractions
- Presence of ‘absolutely’, ‘really’
- Presence of hedges
- Complexity: syls/words
- Number of repeated words
- Punctuation type
- Length of unit (in sec and words)
- # words/unit length
- # of laughs
- # of audible breaths
- # of other speaker noise
- # of mispronounced words
- # of unintelligible words
Subject-Dependent Features

- % units with cue phrases
- % units with filled pauses
- % units with laughter
- Lies/truths with filled pauses ratio
- Lies/truths with cue phrases ratio
- Lies/truths with laughter ratio
- Gender
Classification Results (Unbalanced Corpus)

- **Features (z-score normalized):**
  - **Acoustic features:** f0, intensity, voice quality, speaking rate – raw and normalized 2 ways
  - **Gender:** subject and partner
  - **Ethnicity:** subject and partner
  - **Personality scores**
  - **Lexical features** not yet available for entire corpus

- **Weka experiments**
  - J48 decision trees
  - Random Forests
  - Bagging
## Classification Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Raw</th>
<th>SessionNorm</th>
<th>BaselineNorm</th>
</tr>
</thead>
<tbody>
<tr>
<td>J48</td>
<td>59.89</td>
<td>62.09</td>
<td>62.19</td>
</tr>
<tr>
<td>Bagging</td>
<td>58.65</td>
<td>61.19</td>
<td>61.01</td>
</tr>
<tr>
<td>RF</td>
<td>61.23</td>
<td><strong>63.03</strong></td>
<td>62.79</td>
</tr>
</tbody>
</table>

*Baseline accuracy: 59.9%*
Added features

• Speaker gender
• Speaker native language
• NEO-FFI personality scores
## Adding Gender, Ethnicity and Personality Features (SessionNorm)

<table>
<thead>
<tr>
<th>Model</th>
<th>Acoustic/prosodic</th>
<th>Acoustic/prosodic + gender/ethn/NEO</th>
</tr>
</thead>
<tbody>
<tr>
<td>J48</td>
<td>62.09</td>
<td>64.86</td>
</tr>
<tr>
<td>Bagging</td>
<td>61.19</td>
<td>63.9</td>
</tr>
<tr>
<td>RF</td>
<td>63.03</td>
<td>65.86</td>
</tr>
</tbody>
</table>

- **Majority class baseline accuracy:** 59.9%
# Accuracy Predictions (Baseline 59.9%)

3 ML Models, Raw vs. Norm’d Acoustic Features

<table>
<thead>
<tr>
<th>Model</th>
<th>Raw</th>
<th>Session Norm</th>
<th>Baseline Norm</th>
</tr>
</thead>
<tbody>
<tr>
<td>J48</td>
<td>59.89</td>
<td>62.09</td>
<td>62.19</td>
</tr>
<tr>
<td>Bagging</td>
<td>58.65</td>
<td>61.19</td>
<td>61.01</td>
</tr>
<tr>
<td>RandomForest</td>
<td>61.23</td>
<td><strong>63.03</strong></td>
<td>62.79</td>
</tr>
</tbody>
</table>

All Features, (Session Norm’d Acoustic)

<table>
<thead>
<tr>
<th>Model</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>J48</td>
<td>64.86</td>
</tr>
<tr>
<td>Bagging</td>
<td>63.9</td>
</tr>
<tr>
<td>RandomForest</td>
<td><strong>65.86</strong></td>
</tr>
</tbody>
</table>
Interviewer Correctness and Confidence by Question

Confidence Score vs Question Number

- Correct Guessing Rate
- False Only Correct Guessing Rate
- Total avg Confidence Score

Question Number

Correct Guessing Rate

Average Confidence Score

0.7

0.6

0.5

0.4

0.3

0.2

0.1

0

1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

3.8

3.6

3.4

3.2

3

2.8

2.6

2.4

2.2

2

1.8

1.6

1.4

1.2

1

0.8

0.6

0.4

0.2

0
Which Question Types are Most Difficult?

• Yes/no vs. open-ended questions
  – Do you own an e-reader of any kind?
  – Who ended your last romantic relationship?

• Sensitive (intrusive, threat of disclosure, social desirability) vs. non-sensitive
  – Who ended your last romantic relationship?
  – Do you own an e-reader of any kind?
Current Findings

• All interviewers better at guessing T/F for yes/no questions than for open-ended questions
  – Only English interviewers detected lies to ynq’s better than to open-ended questions
  – Many differences in pair types wrt detecting lies to ynq’s (e.g. English females did better when interviewing Mandarin males than English males)

• All interviewers detected T/F better in sensitive questions than non-sensitive, with some differences for pair types
Intriguing Results on Baseline Interviews

• 3-4m pre-experiment interviews between subject and experimenter to use for normalizing interview data
• Can we detect gender, ethnicity, personality types and even ability to lie or detect lies from this data alone?
Baseline Interview

- Tell me how you decided to come to Columbia.
- What are your favorite things to do in New York City?
- What do you like the most about living in New York City?
- What do you like the least about living in New York City?
- Describe a typical weekend for you, from Friday night through Sunday night.
- What was the best food you ever ate, where did you have it, and what made it so good?
- Where was the last place you traveled, and what are some things you did while you were there?
- What was the last movie you saw, and what was the premise of it?
- Besides work or school, what do you do with your time?
- What did you do this past summer?
Findings

• 3-4 minutes of ’normal’ speech turn out to yield good classification of 
ethnicity, gender, and success at deception detection, with some success at predicting 
NEO-FFI personality scores as well

• Promising results for future work on classifying speakers from small amounts of ‘normal’ speech
Gender Classification

<table>
<thead>
<tr>
<th>Features</th>
<th>acc</th>
<th>prec</th>
<th>rec</th>
<th>f-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prosodic</td>
<td>.94</td>
<td>.94</td>
<td>.96</td>
<td>.97</td>
</tr>
<tr>
<td>LIWC</td>
<td>.70</td>
<td>.69</td>
<td>.86</td>
<td>.76</td>
</tr>
<tr>
<td>Combined</td>
<td>.94</td>
<td>.94</td>
<td>.98</td>
<td>.96</td>
</tr>
</tbody>
</table>

- Baseline: 61% female
- \( f_0 \) is the best feature
- Useful LIWC categories:
  - ‘anger’, ‘money’, ‘certainty’ -> male

Native Language

<table>
<thead>
<tr>
<th>Features</th>
<th>acc</th>
<th>prec</th>
<th>rec</th>
<th>f-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prosodic</td>
<td>.65</td>
<td>.64</td>
<td>.40</td>
<td>.52</td>
</tr>
<tr>
<td>LIWC</td>
<td>.84</td>
<td>.82</td>
<td>.7</td>
<td>.76</td>
</tr>
<tr>
<td>Combined</td>
<td>.81</td>
<td>.88</td>
<td>.68</td>
<td>.78</td>
</tr>
</tbody>
</table>

- Baseline: 57% English
- Useful LIWC categories:
  - ‘apostrophe’ -> English (contractions)
  - ‘dash’ -> Mandarin (false starts)
Personality Classification (F measure)

<table>
<thead>
<tr>
<th>Features</th>
<th>Model</th>
<th>N</th>
<th>E</th>
<th>O</th>
<th>A</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prosodic</td>
<td>AB</td>
<td>.43</td>
<td>.30</td>
<td>.48</td>
<td>.46</td>
<td>.35</td>
</tr>
<tr>
<td></td>
<td>RF</td>
<td>.39</td>
<td>.33</td>
<td>.45</td>
<td>.40</td>
<td>.33</td>
</tr>
<tr>
<td>LIWC</td>
<td>AB</td>
<td>.38</td>
<td>.34</td>
<td>.45</td>
<td>.41</td>
<td>.42</td>
</tr>
<tr>
<td></td>
<td>RF</td>
<td>.38</td>
<td>.36</td>
<td>.56</td>
<td>.47</td>
<td>.32</td>
</tr>
<tr>
<td>Combined</td>
<td>AB</td>
<td>.43</td>
<td>.33</td>
<td>.48</td>
<td>.40</td>
<td>.34</td>
</tr>
<tr>
<td></td>
<td>RF</td>
<td>.41</td>
<td>.36</td>
<td>.53</td>
<td>.47</td>
<td>.39</td>
</tr>
<tr>
<td>Baseline</td>
<td>–</td>
<td>.28</td>
<td>.26</td>
<td>.34</td>
<td>.32</td>
<td>.27</td>
</tr>
</tbody>
</table>

- Binning: high, medium, low for each factor, based on population means
- Compare AdaBoost and RandomForest classifiers
- Useful LIWC categories:
  - Neuroticism: ‘power’, ‘money’
  - Extraversion: ‘drives’, ‘focusfuture’
  - Openness to experience: ‘interrogation’, ‘focuspast’
  - Agreeableness: ‘social’, ‘assent’
  - Conscientiousness: ‘work’, ‘time’
Success at Deception Detection Ability

• Baseline for ability to detect deception: 52%
  • AdaBoost, prosodic + LIWC features: 61%
  • Adding knowledge of gender, native language, personality: 65%
• Useful features: ‘focuspresent’, speaking rate
• Currently unable to predict success at deception production above chance – but not yet using full set of lexical features
Related and Future Work

- **Laughter and deception studies** (with Greg Bryant & Dan Fessler, UCLA): can subjects distinguish *laughter in truth* vs. *laughter in deception*

- More classification experiments to **predict deception** (and DNNs) on entire corpus
  - Additional features:
    - *Lexical* and *subject-dependent* features
    - *Entrainment* as a factor in deception: Do subjects who *entrain* make better deceivers or deception detectors?
  - **Clustering** subjects by gender, ethnicity, and personality features to build different models for each cluster
Deception Detection and Trust

- Spoken indicators of trust across cultures: which factors (personality, gender, ethnicity) might be used to identify trust vs. distrust vs. lack of trust
  - Previous work on trust has focused on nonverbal cues
- **Belief** in deception corpus may provide hypotheses for trust research: what acoustic, prosodic and lexical features correlate with/predict interviewer **belief** in interviewee statements
- Collect new corpus of spoken cues in interactions preceding classic economic trust games
Trust Analyses with Current Corpus

• Assumption: Believing someone is truthful is equivalent to trusting someone (regardless of whether it is true or false).

• Subjects will be considered *having a high level of trust* when:
  – They believe their partner truthful more than X% of the time (where X varies between 50-100)
  – They judge their partner’s statements truthful with high levels of confidence

• Subjects will be classified as *being highly trusted* when:
  – Their partner believes they are truthful more than X% of the time
  – Their partner judged their truthful answers with high levels of confidence
Data Collection: New Trust Corpus

• NEO-FFI personality inventory
• Initial conversation between subjects: to provide speech features similar to those in the first corpus (pitch, intensity, speaking rate, voice quality, lexical features)
• Some economic-based trust game: 2 rounds of a typical Give-Some Game (e.g., DeSteno et al., 2012) or Investment Game (e.g., Buchan, Croson, & Solnick, 2008
  – Participants hear but do not see each other
  – Maximal communal benefit vs. maximal individual gain: how much do I give to my partner?
Publications


• Many invited talks...
Thank you!