Individual Differences in Deception and Deception Detection in Spoken Dialogue

Sarah Ita Levitan
Department of Computer Science
Columbia University
March 9, 2018
Spoken Language Processing
Research Goals

I leverage large amounts of language data, both spoken and written, in order to:

• Learn about human behavior
• Solve real-world problems
Projects

- Entrainment in Supreme Court Oral Arguments
- Automatic Identification of Age and Gender in Call Center Dialogues
- Acoustic Event Detection in YouTube Videos
- Automatic Deception Detection in Interview Dialogues
Motivation
Modalities

- **Body posture and gestures** (Burgoon et al., ‘94)
- **Facial expressions** (Ekman, ‘76; Frank, ‘03)
- **Biometric factors** (Horvath, ‘73)
- **Brain imaging technologies** (Bles & Haynes, ‘08)
- **Language-based**
  - **Text** (Adams, ‘96, Pennebaker et al., ‘01)
  - **Speech** (Enos et al., ‘06)
“Have you ever tweeted?”

TRUE or FALSE?
“Have you ever tweeted?”
Research Goals

• Develop robust techniques to detect deceptive behaviors in spoken dialogue

• Advance understanding of deceptive behavior
  - Individual differences: gender, personality, native language
Challenges

• Hard problem! Human performance ~50%
• Data
• Ground truth annotation
• Individual differences
Contributions

• Columbia X-Cultural Deception Corpus
• Automatic deception classification >70% accuracy
• Empirical study of linguistic indicators of deception
• Individual differences in deceptive behavior
• Entrainment in deceptive speech
Outline

• Introduction
• Deception detection from text and speech
• Individual differences in deceptive behavior
• Acoustic-prosodic indicators of trust
Deception Detection from Text and Speech

Corpus collection → Annotation, feature extraction → Classification → Analysis
Columbia X-Cultural Deception (CXD) Corpus

Levitan et al. 2015, “Cross-Cultural Production and Detection of Deception from Speech”
Levitan et al. 2015, “Individual Differences in Deception and Deception Detection”
Corpus annotation

- Crowdsourced transcription, speech alignment
- TF keypress alignment
- Segmentation
  - Inter-pausal unit (IPU)
  - Turn
  - Question-level (single, chunk)

<table>
<thead>
<tr>
<th>Unit</th>
<th>Interviewee</th>
<th>Interviewer</th>
<th>Total</th>
<th>Avg. duration (s)</th>
<th>Avg. # words</th>
</tr>
</thead>
<tbody>
<tr>
<td>IPU</td>
<td>111,479</td>
<td>81,536</td>
<td>193,015</td>
<td>1.4</td>
<td>4.5</td>
</tr>
<tr>
<td>Turn</td>
<td>43,706</td>
<td>41,753</td>
<td>85,459</td>
<td>4.7</td>
<td>12.4</td>
</tr>
<tr>
<td>Question</td>
<td>7,418</td>
<td>7,418</td>
<td>14,836</td>
<td>single: 3.2, chunk: 20.9</td>
<td>single: 7.9, chunk: 56.5</td>
</tr>
</tbody>
</table>
Machine Learning Experiments

• What are the best classification models?
• What are the optimal segmentation units?
• Which feature sets are most useful?
Machine Learning Experiments

• What are the best classification models?
  • Statistical machine learning – Logistic Regression, SVM, Random Forest
  • Neural networks – DNN, LSTM, hybrid

• What are the optimal segmentation units?

• Which feature sets are most useful?
Machine Learning Experiments

• What are the best classification models?
• What are the optimal segmentation units?
  • IPU, turn, question responses, question chunks
• Which feature sets are most useful?
Machine Learning Experiments

• What are the best classification models?
• What are the optimal segmentation units?
• Which feature sets are most useful?
  • Text-based: n-grams, psycholinguistic, LIWC, word embeddings
  • Speech-based: openSMILE IS09
Machine Learning Experiments

• 60% train, 30% test
• Random subsampling -> balanced T and F classes
• Classifiers:
  • Logistic Regression, N-grams (Text)
  • Random Forest, IS09 (Speech)
  • Random Forest, N-grams + IS09 (Text+Speech)
• Segmentation:
  • IPU
  • Turn
  • Question – first turn
• Evaluation: F1-score
Segmentation: IPU, Turn, Question

Mendels, Levitan et al. 2017, “Hybrid acoustic lexical deep learning approach for deception detection”
Deep Learning Approaches

• BLSTM-lexical
• DNN-openSMILE
• Hybrid: BLSTM-lexical + DNN-openSMILE

Mendels, Levitan et al. 2017, “Hybrid acoustic lexical deep learning approach for deception detection”
Mendels, Levitan et al. 2017, “Hybrid acoustic lexical deep learning approach for deception detection”
Machine Learning Experiments

• 60% train, 30% test
• Random subsampling -> balanced T and F classes
• Random Forest Classifier

Segmentation:
  • Question – first turn
  • Question – chunk

Features:
  • LIWC – psycholinguistic word dimensions
  • Lexical – based on deception literature
  • Speech – IS09

Evaluation: F1-score
Question-Level Results
## Analysis – acoustic features

<table>
<thead>
<tr>
<th>Feature</th>
<th>t</th>
<th>p</th>
<th>sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duration</td>
<td>-0.63</td>
<td>0.53</td>
<td></td>
</tr>
<tr>
<td>Pitch Max</td>
<td>4.37</td>
<td>1.28E-05</td>
<td>*</td>
</tr>
<tr>
<td>Pitch Mean</td>
<td>0.56</td>
<td>0.58</td>
<td></td>
</tr>
<tr>
<td>Intensity Max</td>
<td>3.45</td>
<td>0.0006</td>
<td>*</td>
</tr>
<tr>
<td>Intensity Mean</td>
<td>1.33</td>
<td>0.18</td>
<td></td>
</tr>
<tr>
<td>Speaking Rate</td>
<td>-1.69</td>
<td>0.09</td>
<td></td>
</tr>
<tr>
<td>Jitter</td>
<td>-1.31</td>
<td>0.19</td>
<td></td>
</tr>
<tr>
<td>Shimmer</td>
<td>-1.39</td>
<td>0.17</td>
<td></td>
</tr>
<tr>
<td>NHR</td>
<td>0.35</td>
<td>0.73</td>
<td></td>
</tr>
</tbody>
</table>
## Analysis – lexical features

<table>
<thead>
<tr>
<th>Truth</th>
<th>Deception</th>
<th>Neutral</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negation</td>
<td>Clout</td>
<td>Laughter</td>
</tr>
<tr>
<td>Function words</td>
<td>Informal</td>
<td>Comparison</td>
</tr>
<tr>
<td>Certain</td>
<td>Word count</td>
<td>Anger</td>
</tr>
<tr>
<td>Cognitive processes</td>
<td>Words per second</td>
<td>Power</td>
</tr>
<tr>
<td></td>
<td>Past tense</td>
<td>Present tense</td>
</tr>
<tr>
<td></td>
<td>Specificity</td>
<td>Future tense</td>
</tr>
<tr>
<td></td>
<td>Hedges</td>
<td>Complexity</td>
</tr>
<tr>
<td></td>
<td>Imaginary</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3rd person pronouns</td>
<td></td>
</tr>
</tbody>
</table>
Contributions

• Large-scale corpus of deceptive speech
• Deception classification experiments
  • >.70 F1-Score when using question segmentation
  • Deep learning approaches for IPU segmentation
• Empirical analysis of acoustic-prosodic and linguistic indicators of deception
Outline

• Introduction
• Deception detection from text and speech
• **Individual differences in deceptive behavior**
• Acoustic-prosodic indicators of trust
Individual Differences

- **Extroversion** was correlated with success at deception, for **English male** speakers
- **Native English** speakers performed better at deception when paired with a **native Chinese** interviewer

Levitan et al. 2015, “Cross-Cultural Production and Detection of Deception from Speech”
Levitan et al. 2015, “Individual Differences in Deception and Deception Detection”
# Group-specific Deception Indicators

<table>
<thead>
<tr>
<th>Male</th>
<th>Female</th>
<th>English</th>
<th>Chinese</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive emotion (T)</td>
<td>Jitter (T)</td>
<td>Intensity mean (F)</td>
<td>Speaking rate (T)</td>
</tr>
<tr>
<td>Interrogatives (F)</td>
<td>Perceptual processes (F)</td>
<td>Swear (F)</td>
<td>Certainty (T)</td>
</tr>
<tr>
<td></td>
<td>Future tense (F)</td>
<td></td>
<td>Feel (F)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Causation (F)</td>
</tr>
</tbody>
</table>

Levitan et al. 2018, “Linguistic indicators of deception and perceived deception in spoken dialogue”
Classification with Individual Traits
Conclusions

• There are gender and cultural differences in deceptive behavior
• Accounting for these differences can improve deception classification
Ongoing work

• Clustering speakers
• Tradeoffs of homogenization vs. using speaker information
• Speaker-dependent features
Outline

• Motivation
• Deception classification from text and speech
• Individual differences in deceptive behavior
• Acoustic-prosodic indicators of trust
Trust in HCI
## Perceived Deception

<table>
<thead>
<tr>
<th>Feature</th>
<th>t</th>
<th>p</th>
<th>sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pitch Max</td>
<td>2.35</td>
<td>0.02</td>
<td>*</td>
</tr>
<tr>
<td>Pitch Mean</td>
<td>1.65</td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td>Intensity Max</td>
<td>2.625</td>
<td>0.009</td>
<td>*</td>
</tr>
<tr>
<td>Intensity Mean</td>
<td>-0.785</td>
<td>0.43</td>
<td></td>
</tr>
<tr>
<td>Speaking Rate</td>
<td>-3.785</td>
<td>0.0002</td>
<td>+</td>
</tr>
<tr>
<td>Jitter</td>
<td>-1.815</td>
<td>0.07</td>
<td></td>
</tr>
<tr>
<td>Shimmer</td>
<td>-1.905</td>
<td>0.06</td>
<td></td>
</tr>
<tr>
<td>NHR</td>
<td>0.58</td>
<td>0.56</td>
<td></td>
</tr>
</tbody>
</table>
## Group-specific “Trust” Indicators - Speaker

<table>
<thead>
<tr>
<th>Feature</th>
<th>Male</th>
<th>Female</th>
<th>English</th>
<th>Chinese</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pitch Max</td>
<td>🟥</td>
<td>🟥</td>
<td></td>
<td></td>
<td>🟥</td>
</tr>
<tr>
<td>Pitch Mean</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intensity Max</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>🟥</td>
</tr>
<tr>
<td>Intensity Mean</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Speaking Rate</td>
<td>🟢</td>
<td>🟢</td>
<td></td>
<td></td>
<td>🟢</td>
</tr>
<tr>
<td>Jitter</td>
<td>🟢</td>
<td>🟢</td>
<td></td>
<td></td>
<td>🟢</td>
</tr>
<tr>
<td>Shimmer</td>
<td>🟢</td>
<td>🟢</td>
<td></td>
<td></td>
<td>🟢</td>
</tr>
<tr>
<td>NHR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>🟥</td>
</tr>
</tbody>
</table>
### Group-specific “Trust” Indicators - Listener

<table>
<thead>
<tr>
<th>Feature</th>
<th>Male</th>
<th>Female</th>
<th>English</th>
<th>Chinese</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pitch Max</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pitch Mean</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intensity Max</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intensity Mean</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Speaking Rate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jitter</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shimmer</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NHR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Questions

• Can we use this information to:
  • Improve *human* deception detection?
  • Create trustworthy synthesized speech?
Speech Corpus Annotation

• Experts
• Crowdsourcing
• Games with a purpose (GWAP)
Games With A Purpose (GWAP)

Idea: Motivate people to solve computational problems by presenting the problem as a series of simple steps in an enjoyable game format.
Games with a purpose

Levitan et al. 2018, “LieCatcher: Game framework for collecting human judgments of deceptive speech”
Thank you!

Julia Hirschberg
Michelle Levine
Andrew Rosenberg
Guozhen An
Bingyan Hu
Gideon Mendels
Angel Maredia
Jessica Xiang
Kai-Zhan Lee
Zoe Baker-Peng

James Shin
Ivy Chen
Meredith Cox
Gauri Narayan
Mandi Wang
Leighanne Hsu
Yocheved Levitan
Grace Ulinski
Arthur Shen
Nishi Cestero

Jennifer Senior
Yvonne Missry
Molly Scott