What makes a conversational agent sound trustworthy?
Exploring the role of acoustic-prosodic factors.

Sarah Ita Levitan
Guest Lecture: Advanced Topics in SLP
Columbia University
February 20, 2024
Hello!

• **Current (2020 – present)**
  Assistant Prof. of Computer Science, Hunter College
  Doctoral faculty of CS and Linguistics, CUNY Graduate Center

• **Previous (2013-2019)**
  PhD & Postdoc at Columbia University
  Columbia Speech Lab: PI Julia Hirschberg
Outline

• Motivation
• Related work
• Data collection
  • Speech stimuli
  • Crowdsourcing experiment
• Acoustic-prosodic characteristics of trustworthy TTS
Motivation
Motivation

• Trust is essential for effective communication and collaboration
• In human-human interaction AND human-computer interaction
• We understand a great deal about signals of trust in human speech
• But have a limited understanding of how humans perceive trustworthiness in synthesized speech

What makes a conversational agent sound trustworthy?
Previous work

• Acoustic-Prosodic and Lexical Cues to Deception and Trust: Deciphering How People Detect Lies (Chen & Levitan et al. 2020)

• The sound of trustworthiness: Acoustic-based modulation of perceived voice personality (Belin et al. 2017)
The sound of trustworthiness: Acoustic-based modulation of perceived voice personality (Belin et al. 2017)

• How should one say “hello” to be perceived as trustworthy by new listeners?

• Generate high and low trustworthiness speech stimuli

• Evaluate perception of trustworthiness with online study
Synthesis of Trustworthy/Untrustworthy stimuli

• STRAIGHT toolkit in Matlab
• Decompose natural speech stimulus into 5 parameters:
  • F0, frequency, spectro-temporal density, aperiodicity
• Manipulate and combine parameters
• Synthesize into a novel voice stimulus
Stimuli

Trust continuum stimuli

Control stimuli
Correlation between acoustics and trust ratings ($r=0.99$, $p=0$)
Intonation and perceived trustworthiness
Current Study: Data Collection

• Speech stimuli preparation
• Crowdsourced perception study
Text selection

- Emotional Support Conversations Dataset (Liu et al. 2021)
- 1300 crowdsourced conversations between human help-seeker and virtual supporter
- Application that requires trust and vulnerability from the user
- We select sentences labeled as supporter questions
Amazon Polly Neural TTS

- State-of-the-art, commercial TTS system
- Integrated with dialogue systems and conversational robots
- Supports voice alterations using SSML
- Pre-trained male and female voices
Speech Synthesis Markup Language (SSML)

```
<speak>
<voice name="Joanna"><lang xml:lang="en-US">
<prosody pitch="-27%" rate="95%" volume="+0dB">
Call me Ishmael. <break time="300ms"/> Some years ago <break time="300ms"/> never mind how long precisely <break time="300ms"/> having little or no money in my purse, and nothing particular to interest me on shore, I thought I would sail about a little <break time="100ms"/> and see the watery part of the world.
</prosody></lang></voice>
</speak>
```
Acoustic-prosodic features

- Pitch
- Intensity
- Speaking rate
Total speech stimuli

• 27 prosodic profiles
  • 3 features (pitch, intensity, rate) x 3 settings (low, medium, high)
• 2 voices
  • 1 male ("Matthew"), 1 female ("Joanna")
• 10 question utterances

• Total: 540 speech samples
Mean feature values

<table>
<thead>
<tr>
<th>Level</th>
<th>Intensity</th>
<th>Pitch</th>
<th>Speaking rate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F</td>
<td>M</td>
<td>F</td>
</tr>
<tr>
<td>High</td>
<td>57.2</td>
<td>55</td>
<td>203</td>
</tr>
<tr>
<td>Med</td>
<td>52</td>
<td>51</td>
<td>163</td>
</tr>
<tr>
<td>Low</td>
<td>49</td>
<td>47</td>
<td>131</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>294</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>236</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>192</td>
</tr>
</tbody>
</table>
Examples

• Low pitch, intensity, speaking rate

• Medium pitch, intensity, speaking rate

• High pitch, intensity, speaking rate
Crowdsourced Perception Study

• Listen to 20 audio clips
• Rate speaker traits with 5-point Likert scale
  • Trustworthy, lively, empathetic, respectful, cold, engaging
• Quality control: transcription task
• Listener traits:
  • Ten Item Personality Inventory (TIPI)
  • Gender
Crowdsourced Perception Study

- 135 participants (71 F, 63 M)
- Each audio sample was rated by 5 unique raters
- 2700 judgments of 540 speech stimuli
- All judgments are z-normalized by rater
## Inter-Annotator Agreement

- Krippendorff’s alpha

<table>
<thead>
<tr>
<th></th>
<th>trustworthy</th>
<th>lively</th>
<th>natural</th>
<th>boring</th>
<th>empathetic</th>
<th>respectful</th>
<th>cold</th>
<th>engaging</th>
</tr>
</thead>
<tbody>
<tr>
<td>all raters</td>
<td>0.21</td>
<td>0.18</td>
<td>0.17</td>
<td>0.2</td>
<td>0.2</td>
<td>0.18</td>
<td>0.22</td>
<td>0.2</td>
</tr>
<tr>
<td>F raters</td>
<td>0.13</td>
<td>0.08</td>
<td>0.08</td>
<td>0.07</td>
<td>0.1</td>
<td>0.1</td>
<td>0.12</td>
<td>0.1</td>
</tr>
<tr>
<td>M raters</td>
<td>0.17</td>
<td>0.13</td>
<td>0.14</td>
<td>0.17</td>
<td>0.14</td>
<td>0.15</td>
<td>0.16</td>
<td>0.13</td>
</tr>
</tbody>
</table>
### Average ratings per trait

<table>
<thead>
<tr>
<th></th>
<th>trustworthy</th>
<th>lively</th>
<th>natural</th>
<th>boring</th>
<th>empathetic</th>
<th>respectful</th>
<th>cold</th>
<th>engaging</th>
</tr>
</thead>
<tbody>
<tr>
<td>all raters</td>
<td>3.66</td>
<td>3.66</td>
<td>3.48</td>
<td>2.6</td>
<td>3.34</td>
<td>3.7</td>
<td>3</td>
<td>3.6</td>
</tr>
<tr>
<td>F raters</td>
<td>3.69</td>
<td>3.7</td>
<td>3.6</td>
<td>2.5</td>
<td>3.3</td>
<td>3.6</td>
<td>3</td>
<td>3.6</td>
</tr>
<tr>
<td>M raters</td>
<td>3.66</td>
<td>3.6</td>
<td>3.4</td>
<td>2.6</td>
<td>3.3</td>
<td>3.7</td>
<td>3</td>
<td>3.5</td>
</tr>
</tbody>
</table>
Questions

• How do raters define trustworthiness in terms of other speaker traits?
  • Lively, empathetic, respectful, cold, engaging

• What are the acoustic-prosodic characteristics of trustworthy speech?
  • And other speaker traits?

• How do listener characteristics (gender, personality) affect their perception of trustworthiness and other speaker traits?
Correlation analysis of speaker attributes
Acoustic-prosodic characteristics of trustworthy TTS

<table>
<thead>
<tr>
<th>Intensity</th>
<th>Pitch</th>
<th>Rate</th>
<th>Gender</th>
<th>Avg Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>medium</td>
<td>medium</td>
<td>medium</td>
<td>M</td>
<td>4.1</td>
</tr>
<tr>
<td>high</td>
<td>high</td>
<td>medium</td>
<td>F</td>
<td>4</td>
</tr>
<tr>
<td>low</td>
<td>high</td>
<td>low</td>
<td>M</td>
<td>3.95</td>
</tr>
<tr>
<td>medium</td>
<td>high</td>
<td>medium</td>
<td>M</td>
<td>3.91</td>
</tr>
<tr>
<td>medium</td>
<td>high</td>
<td>low</td>
<td>F</td>
<td>3.9</td>
</tr>
<tr>
<td>low</td>
<td>low</td>
<td>high</td>
<td>M</td>
<td>3.2</td>
</tr>
<tr>
<td>high</td>
<td>low</td>
<td>high</td>
<td>M</td>
<td>3.15</td>
</tr>
<tr>
<td>medium</td>
<td>medium</td>
<td>high</td>
<td>M</td>
<td>3.15</td>
</tr>
<tr>
<td>high</td>
<td>low</td>
<td>high</td>
<td>F</td>
<td>3</td>
</tr>
<tr>
<td>high</td>
<td>medium</td>
<td>high</td>
<td>M</td>
<td>2.9</td>
</tr>
</tbody>
</table>
GLS Regression Analysis

<table>
<thead>
<tr>
<th>Features</th>
<th>trustworthy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>r</td>
</tr>
<tr>
<td>intensity low</td>
<td>-0.13</td>
</tr>
<tr>
<td>intensity medium</td>
<td>0.31</td>
</tr>
<tr>
<td>intensity high</td>
<td>-0.17</td>
</tr>
<tr>
<td>pitch low</td>
<td>-0.17</td>
</tr>
<tr>
<td>pitch medium</td>
<td>0.29</td>
</tr>
<tr>
<td>pitch high</td>
<td></td>
</tr>
<tr>
<td>speaking rate low</td>
<td>0.3</td>
</tr>
<tr>
<td>speaking rate medium</td>
<td>0.4</td>
</tr>
<tr>
<td>speaking rate high</td>
<td></td>
</tr>
</tbody>
</table>
How does listener gender affect their perception?

- Female listeners were more likely to perceive speakers as natural ($r=-0.18$) and empathetic ($r=-0.07$)
- Male listeners were more likely to perceive speakers as boring ($r=0.11$)
How does listener personality affect their perception?
Summary

• Crowdsourced perception study of trustworthy synthesized speech
• Identified specific patterns of synthesized speech associated with perceived trustworthiness
• Listener gender and personality traits may affect perception

• Next steps: explore the role of lexical factors
  • Dialogue act
  • Politeness
  • Complexity
Thank you!

• Yuwen Yu: PhD student, CUNY Graduate Center
• Ghazanfar Shahbaz: Previous undergraduate student, Hunter College
• Funding: NSF EAGER
Questions?