What makes a voice sound trustworthy?

Understanding and modeling deception and trustworthiness in human and machine speech

Sarah Ita Levitan

Guest Lecture: Advanced Topics in SLP

Columbia University

Dec 2, 2025

What can we convey and perceive from speech?

- Gender
- Age
- Native language
- Ethnicity
- Personality
- Physical health
- Mental health
- Charisma
- Likeability
- Emotion
- Sarcasm
- Humor
- Deception
- Trust

What can we **automatically** learn about speaker states and traits from their speech?

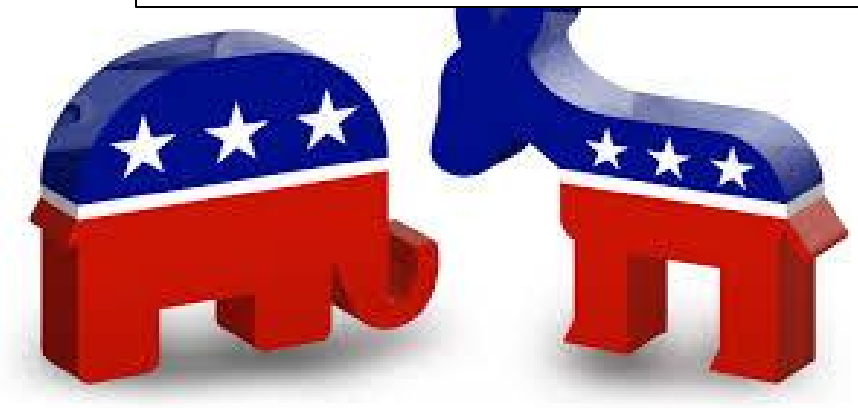
And how can we leverage this information to improve human-computer interactions?

Deception

Trust

What are the characteristics of **deceptive** and **truthful** speech?
What makes humans **perceive** speech as truthful, or **trust** speech?
Can we automatically detect **deceptive** and **trustworthy** speech?

Human performance at deception detection is about 50% -> random chance.



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)

Challenges

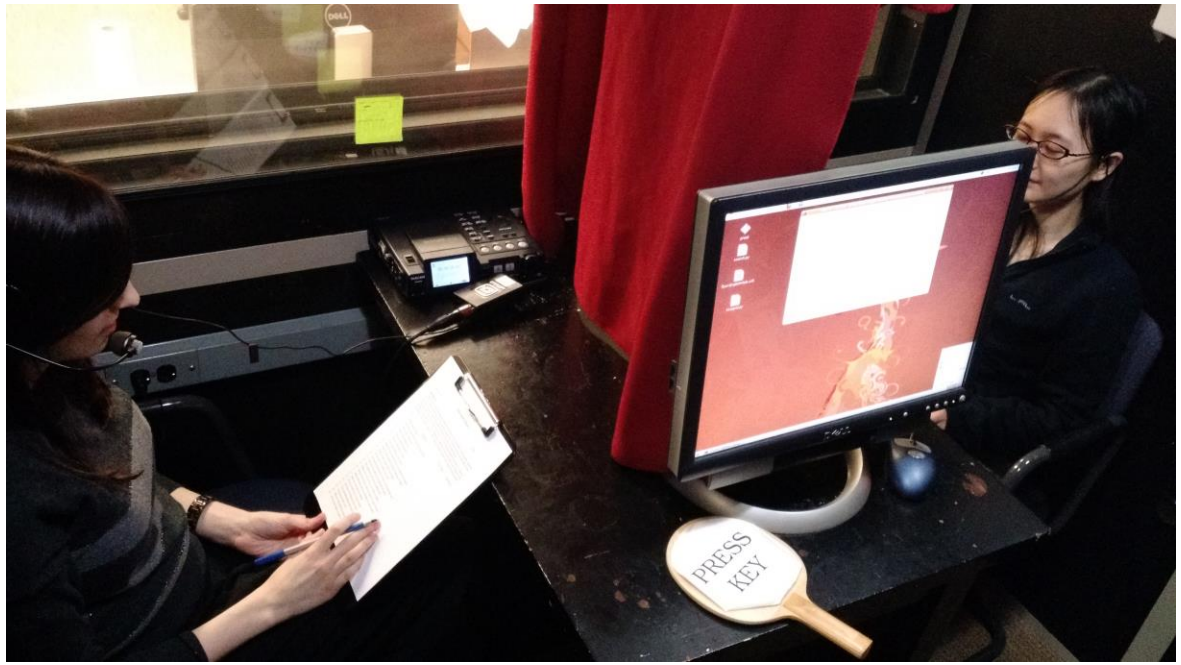
Data

Ground truth annotation

Laboratory vs. real-world deception

Individual and cultural differences

Columbia X-Cultural Deception Corpus



>120 hours of subject speech
340 subjects
Cross-cultural
Fake resume paradigm
NEO-FFI personality scores
Baseline sample
Financial incentive
Lie production/perception
Global/local deception labels

Units of analysis

IPU Pause-free segment of speech from a single speaker

Turn Sequence of speech from one speaker without intervening speech from the other speaker

Question response Interviewee turn following an interviewer biographical question

Question chunk Set of interviewee turns responding to an interviewer biographical question and subsequent follow-up questions

Units of analysis

Unit	Interviewer	Interviewee	Total
IPU	81536	111428	192964
Turn	41768	43673	85459
Question Response	8092	8092	16184
Question Chunk	8092	8092	16184

“Have you ever tweeted?”

Acoustic-prosodic and Lexical Features (152)

Acoustic-prosodic (8) pitch {max, mean}, intensity {max, mean}, speaking rate, jitter, shimmer NHR

LDI (28) hedge words, filled pauses, contractions, denials, laughter, DAL (Dictionary of Affect in Language; Whissel et al., 1986), specificity (Li & Nenkova, 2015)

LIWC (93) word counts for semantic classes – linguistic, markers of psychological processes, punctuation, formality

Complexity (23) measures of syntactic complexity (e.g. clauses per sentence, coordinate phrases per clause)

Summary: acoustic-prosodic and linguistic characteristics of deception and truth

Deception

Increased pitch & intensity max

Poor speech planning

Descriptive, detailed

Complex

Hedge

Entrainment

Truth

Negation

Cue phrases

Cognitive process

Function words

Automatic deception detection

Four units of analysis:

IPU, turn, question response, question chunk

Four statistical classifiers: Random Forest, Logistic Regression, SVM, Naïve Bayes

Three neural network classifiers: DNN, LSTM, Hybrid

Three feature sets: Acoustic (A), Lexical (L), Syntactic (S)

Evaluation metric:

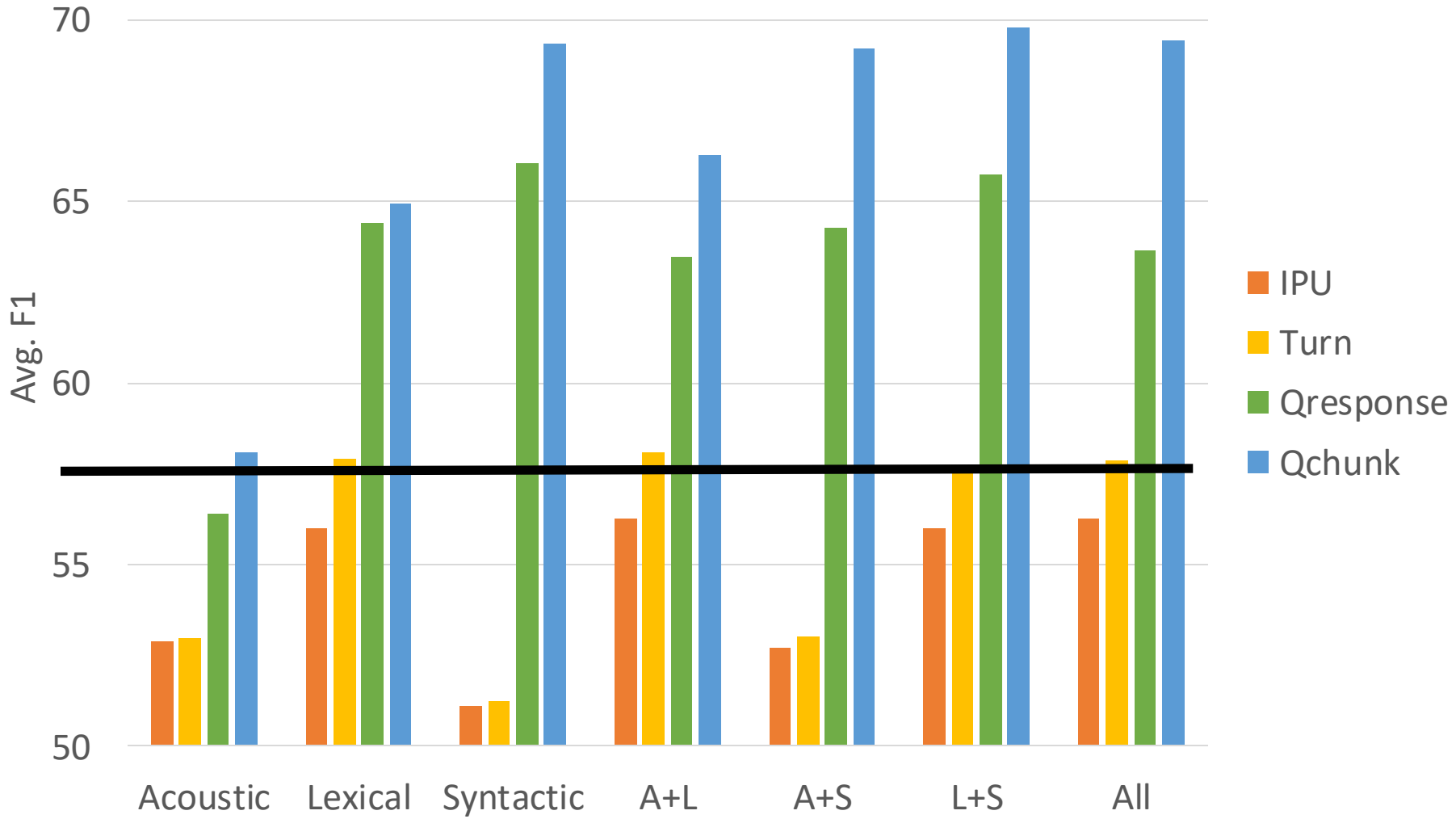
$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Baselines:

Random: 50% accuracy

Human: 56.75% accuracy (question chunk units)

Deception classification

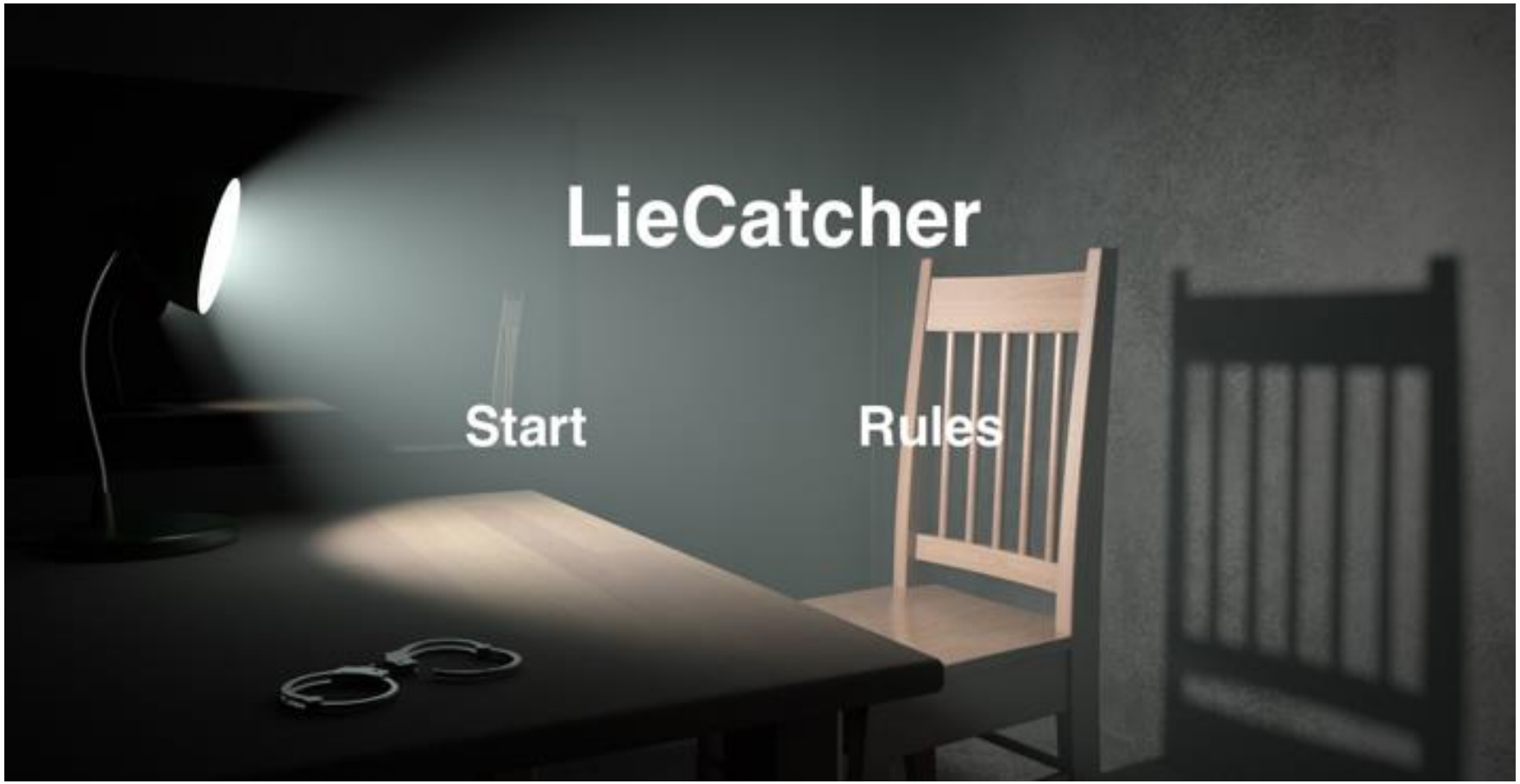


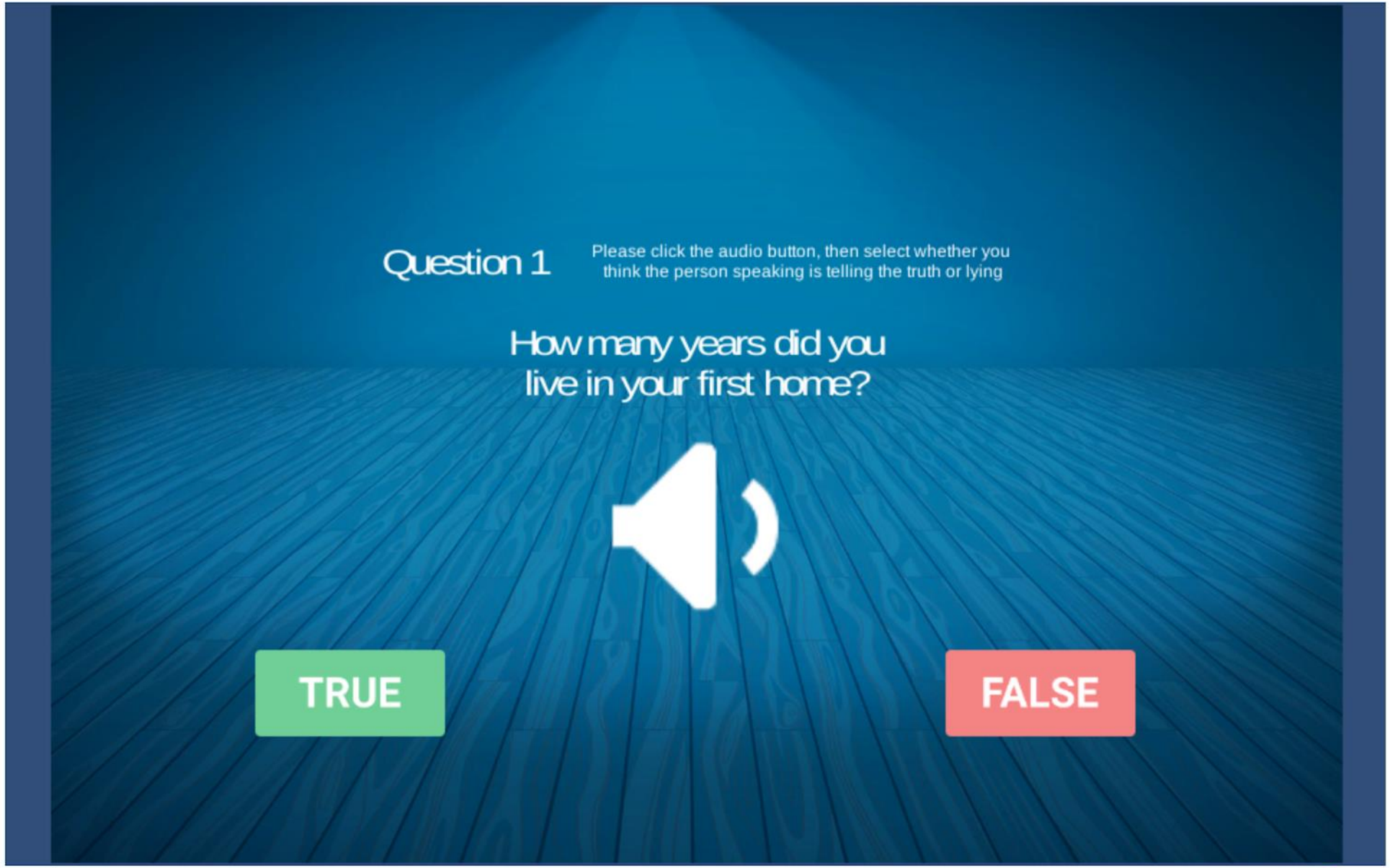
Deception

Trust

What are the characteristics of **deceptive** and **truthful** speech?
What makes humans **perceive** speech as truthful, or **trust** speech?
Can we automatically detect **deceptive** and **trustworthy** speech?

LieCatcher





unity WebGL

speech-game

Crowdsourcing Study

- 5,340 utterances
- 3 judgments per utterance
- 431 unique annotators
- 38.9% male, 59.1% female, 2.1% unreported

Lie Detection Ability

- Overall accuracy = 49.93%
- Fleiss' kappa: 0.135
- Truth bias – 65% trusted
  - Truth Default Theory (T.R. Levine, 2014)

Disfluency

Features	Trust	Deception
Has filled pause	↓↓↓↓	↑↑↑↑
# filled pause	↓↓↓↓	↑↑↑↑
Response latency	↓↓↓↓	
Repetition	↓↓↓↓	↑
False start	↓↓↓	↑↑

↓ indicates negative relationship; ↑ indicates positive relationship
↓: <.05, ↓↓ : <.01, ↓↓↓ : <.001, ↓↓↓↓ : <.0001

Prosody

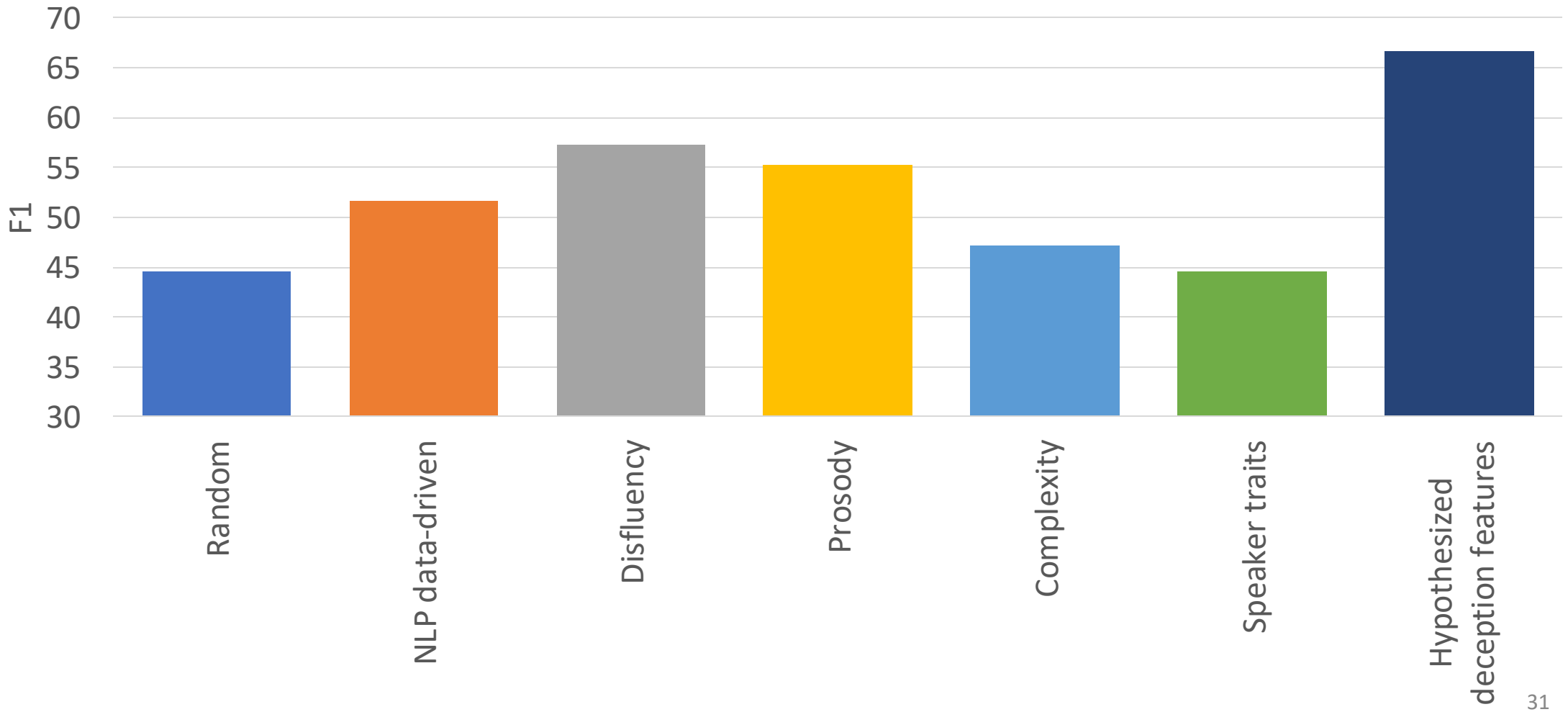
Features	Trust	Deception
Speaking rate	↑↑↑↑	
Pitch max	↑↑↑↑	↑↑↑↑
Pitch mean	↑↑	
Pitch std	↑↑	↑↑
Intensity max		↑↑↑
Intensity mean	↑↑↑↑	
Intensity std	↓↓↓↓	↑
Jitter, shimmer, nhr	↑↑↑↑	

↓ indicates negative relationship; ↑ indicates positive relationship
↓: <.05, ↓↓ : <.01, ↓↓↓ : <.001, ↓↓↓↓ : <.0001

Can we predict trusted speech?

- 5-fold cross validation, speaker independent
- Low agreement task -> only classify utterances with consensus
- Logistic regression
- Evaluate with macro-F1
- Baseline (random): 44.62 F1

Trust classification results



Contributions

- Large-scale corpus of deceptive dialogues
- Acoustic-prosodic and linguistic cues to deception
- Automatic deception classification: ~70% accuracy
- Crowdsourced study of deception perception
- Identified characteristics of trusted speech
- Predictive models for trusted speech detection

What can we **automatically** learn about speaker states and traits from their speech?

And how can we leverage this information to improve human-computer interactions?

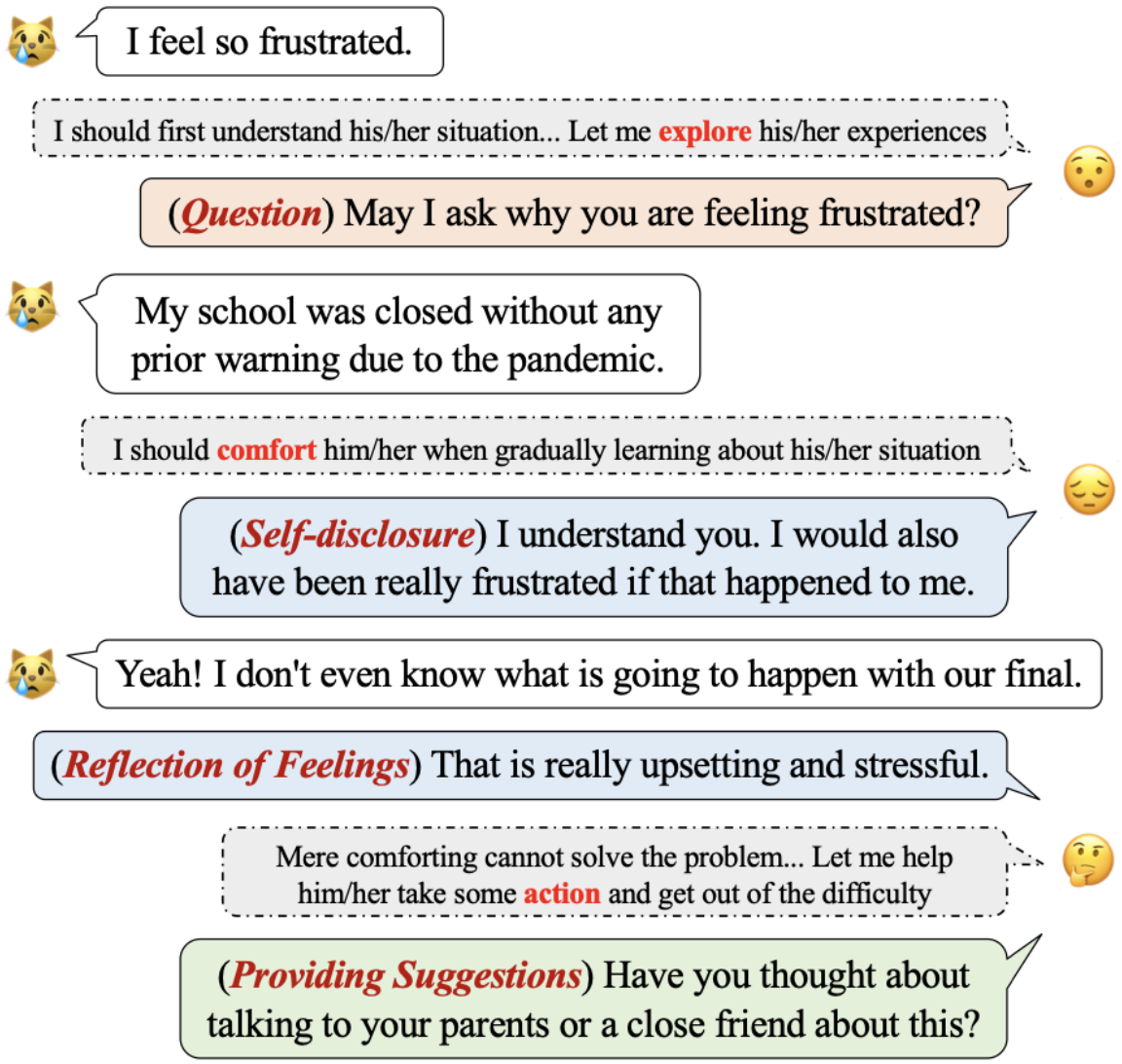
Motivation

- 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?

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

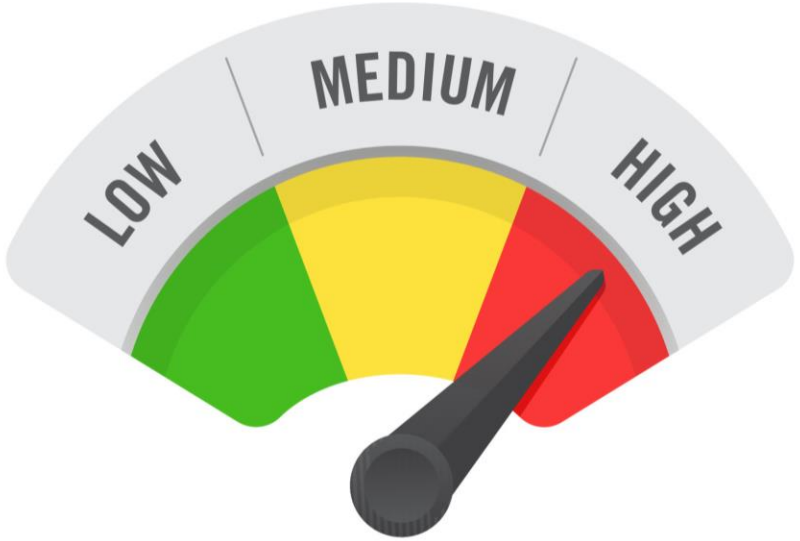
- State-of-the-art, commercial TTS system
- Integrated with dialogue systems and conversational robots
- Supports voice alterations using SSML
- Pre-trained male and female voices

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

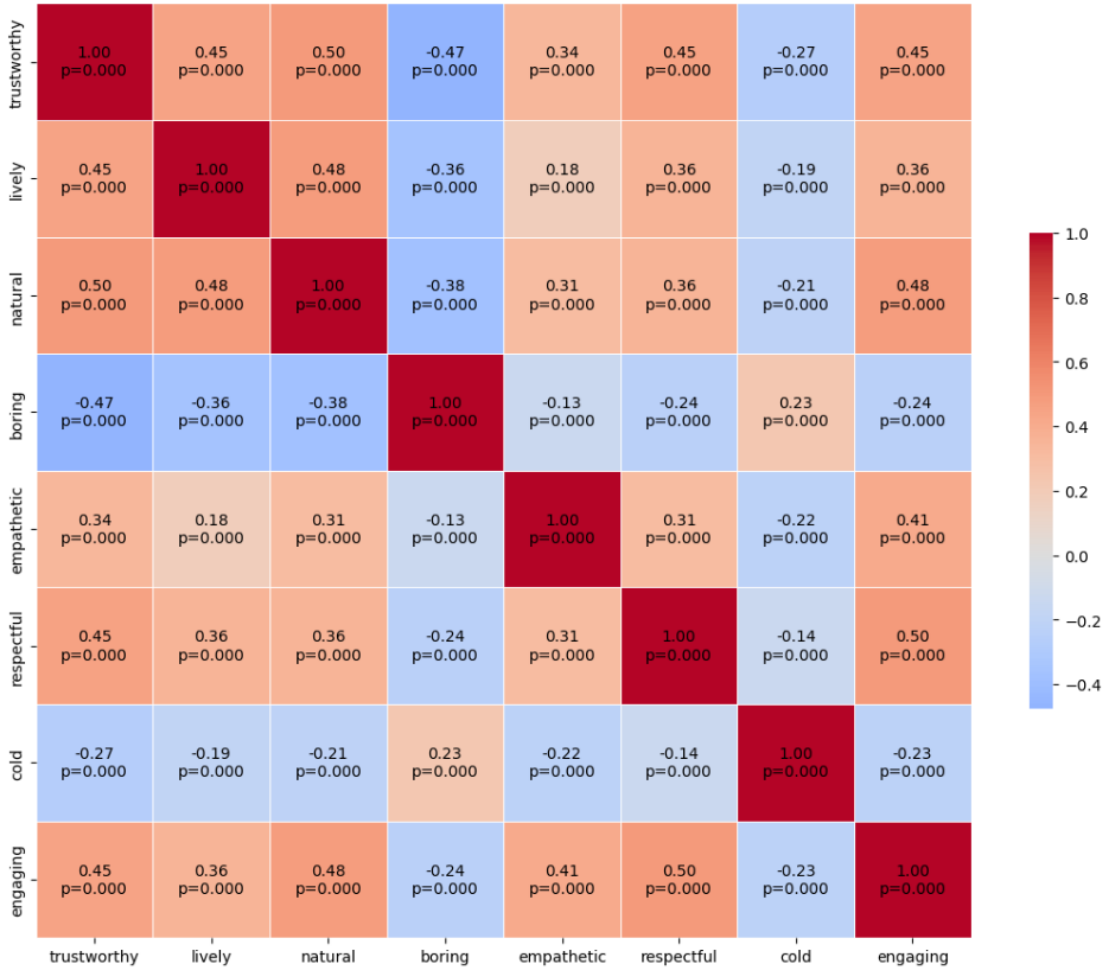
Crowdsourced Perception Study

- 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

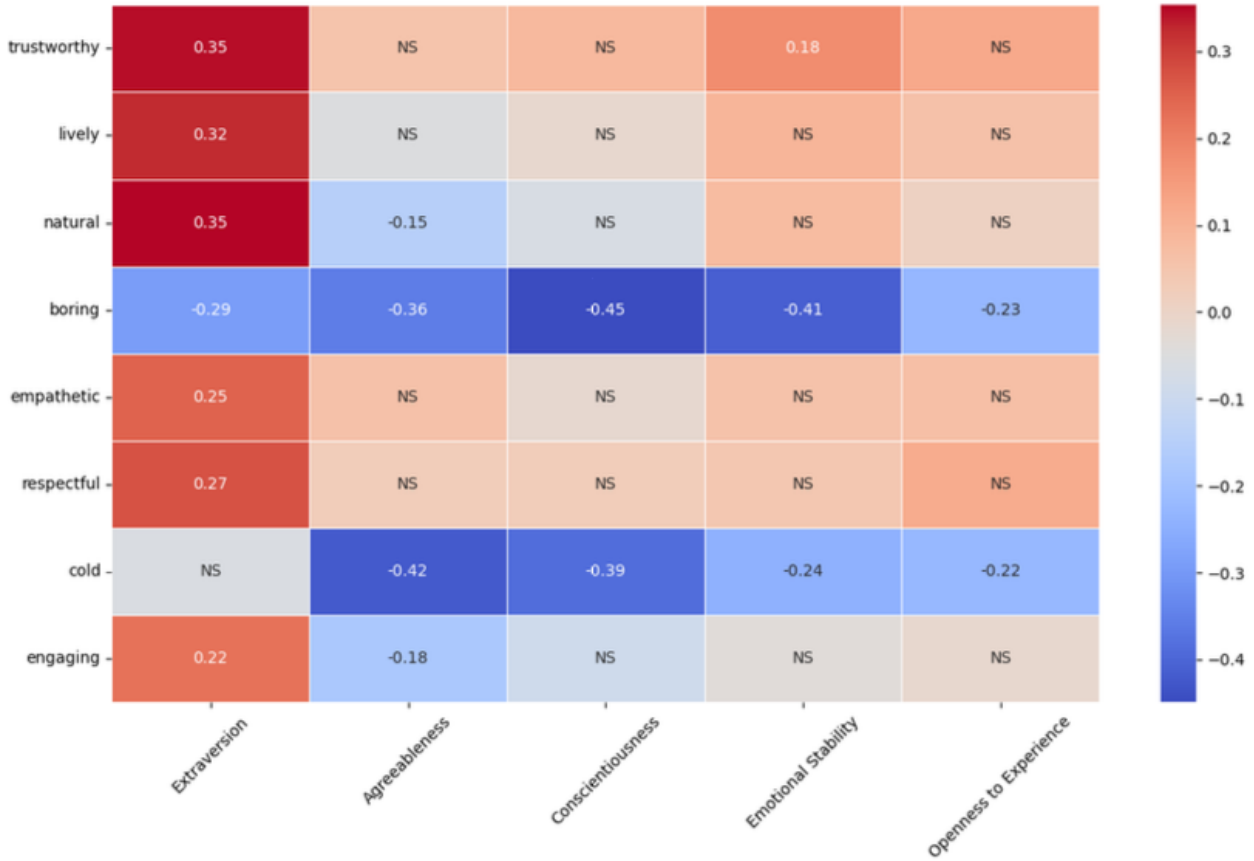
Correlation analysis of speaker attributes



Regression Analysis

Features	trustworthy	
	r	p
intensity low	-0.13	0
intensity medium	0.31	0
intensity high	-0.17	0
pitch low	-0.17	0
pitch medium		
pitch high	0.29	0
speaking rate low	0.3	0
speaking rate medium	0.4	0
speaking rate high		

How does listener personality affect their perception?



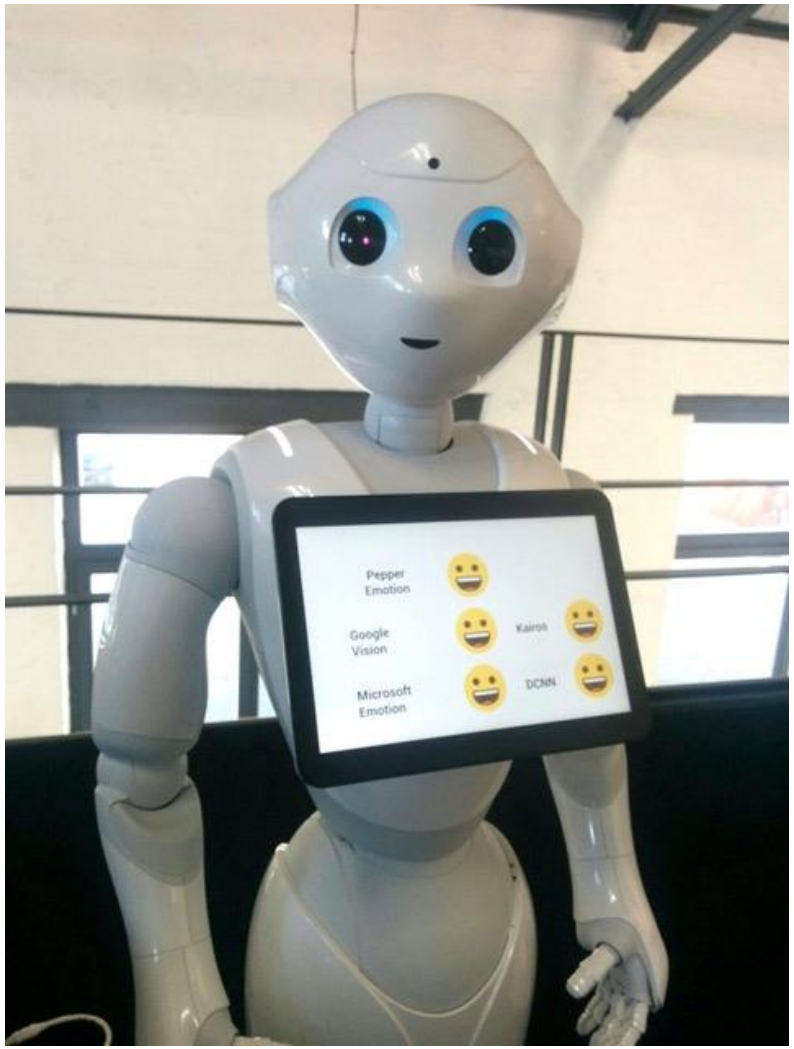
Summary

- 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

What can we **automatically** learn about speaker states and traits from their speech?

And how can we leverage this information to improve human-computer interactions?

Modeling Speaker States and Traits



Thank you!

- Julia Hirschberg, Andrew Rosenberg, Michelle Levine
- Yuwen Yu: PhD student, CUNY Graduate Center, Computer Science
- Natasha Tyulina: PhD student, CUNY Graduate Center, Linguistics
- Funding: NSF EAGER, NSF AI Institute

Publications

Understanding linguistic and visual factors that affect human trust perception of virtual agents. N. Tyulina, Y. Yu, S. I. Levitan. CUI 2024.

What makes a conversational agent sound trustworthy? Exploring the role of acoustic-prosodic factors. Y. Yu, S. I. Levitan. Speech Prosody 2024.

Believe it or not: Acoustic-prosodic cues to trust and mistrust in spoken dialogue. S. I. Levitan, J. Hirschberg. Speech Prosody 2022.

Acoustic-prosodic and lexical cues to deception and trust: Deciphering how people detect lies. X. Chen, S.I. Levitan, M. Levine, M. Mandic, J. Hirschberg. TACL 2020.

Acoustic-prosodic indicators of deception and trust in interview dialogues. S.I. Levitan, A. Maredia, J. Hirschberg. Interspeech 2018.

Acoustic-prosodic and lexical entrainment in deceptive dialogue. S.I. Levitan, J. Xiang, J. Hirschberg. Speech Prosody 2018.

Publications

Linguistic cues to deception and perceived deception in interview dialogues. S.I. Levitan, A. Maredia, J. Hirschberg. NAACL 2018.

LieCatcher: game framework for collecting human judgments of deceptive speech. S.I. Levitan, J. Shin, I. Chen, J. Hirschberg. Games4NLP, LREC workshop 2018.

Hybrid acoustic-lexical deep learning approach for deception detection. G. Mendels, S.I. Levitan, K.Z. Lee, J. Hirschberg. Interspeech 2017.

Combining acoustic-prosodic, lexical, and phonotactic features for automatic deception classification. S.I. Levitan, G. An, M. Ma, R. Levitan, A. Rosenberg, J. Hirschberg. Interspeech 2016.

Questions?