Whom Do We Trust in Dialogue Systems?

Julia Hirschberg, Xi (Leslie) Chen, Michelle Levine, Columbia University, Sarah Ita Levitan, Hunter College (CUNY) and Marko Mandic (Launchpad.ai)

Collaborators in this research since 2003

- Current CxD Deception & Trust:
  - Sarah Ita Levitan, Michelle Levine, Laura Willson, Nishmar Cestero, Guozhen An, Angel Maredia, Elizabeth Petitti, Molly Scott, Yogesh Singh, Jessica Xiang, Jixuan (Gilbert) Zhang, Rivka Levitan, Andrew Rosenberg, Xi (Leslie) Chen, Rebecca Calinsky, Marko Mandic, Xinyue Tan
- Earlier CDC Deception Project:
  - Frank Enos, Stefan Benus, Jennifer Venditti-Ramprashad, Sarah Friedman, Sarah Gilman, Jared Kennedy, Max Shevyakov, Wayne Thorsen, Alan Yeung, and collaborators from SRI/ICSI and from the University of Colorado at Boulder

Background

- **Multimodal deception detection** supported after 9/11 by a new Department of Homeland Security
  - What **aspects of human behavior** are valid indicators of deception?
  - Can we create **reliable automatic deception detection models** to identify future terrorists?
- Current support from Air Force Office of Scientific Research focuses on **identifying human trust** as well
  - What **aspects of human speech are trusted** by other humans? Can we build **good trust models** to identify and, for better purposes, to generate trusted speech?

Why is Deception Detection a Problem?

Human Performance at Detecting Lies is Very Poor

(Aamodt & Mitchell, 2004; Hartwig et al., 2017)

Group	# Studies	# Subjects	Accuracy %
Criminals	1	52	65.40
Secret service	1	34	64.12
Psychologists	4	508	61.56
Judges	2	194	59.01
Police officers	8	511	55.16
Federal officers	4	341	54.54
Students	122	8,876	54.20
Detectives	5	341	51.16
Investment professionals	1	215	49.4
Parole officers	1	32	40.42

How do People Decide: Truth or Lie?

- **Language**
  - The **words** people say?
  - The **syntax** people choose?
  - How **complex** their discourse is?
  - The **acoustic content** of their speech – Pitch? Intensity? Speaking rate?
  - The **prosody** of their speech?
- Body **gestures**, **facial** features?

Modalities Explored for Deception vs. Truth Detection

- **Body posture and gestures** (Burgoon et al ‘94)
- **Facial expressions** (Ekman ‘76; Frank, ‘03)
- **Biometric factors** (Horvath, ‘73): but not polygraphs
- **Brain imaging** technologies (Bles & Haynes ‘08)
- **Language-based** features
  - **Text** (Adams ’96, Pennebaker et al ‘01)
  - **Speech** (Enos et al. ’07, Levitan et al ’18, Chen et al ‘20)

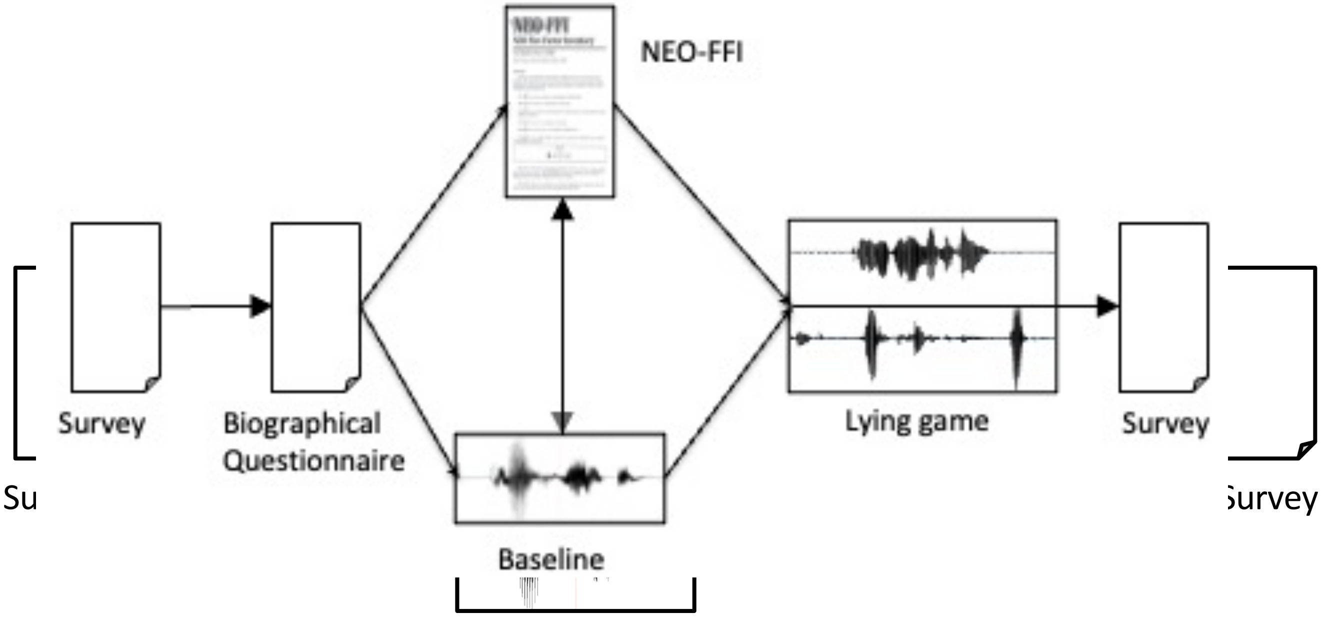
Goals of Our Research

- Identify acoustic-prosodic and linguistic characteristics of **deceptive** and **trustworthy** language: speech and transcripts
- Develop **automated methods** to detect **deceptive** language and **trusted** language – *note that* **these are *not always different***
- Today' talk: Focus on **trusted** and **mistrusted** language in the Columbia Cross-Cultural Deception (CXD) Corpus *(Xi (Leslie) Chen et al "Acoustic-Prosodic and Lexical Cues to Deception and Trust", TACL 2020)*

Columbia X-Cultural Deception (CxD) Corpus

- **340 subjects balanced for gender and native language**
  - Native speakers of **Mandarin Chinese and Standard American English**
  - **>120h** of subject speech
- **Demographic survey**
- **Part-Fake resume** paradigm
- **NEO-FFI** personality scores
- **Baseline** voice sample
- **Financial** incentives
- **Data: Deception production and perception labels**
  - **Global** (interviewer) and **local** (interviewee) **deception labels**

Columbia X-Cultural Deception Corpus



Baseline

No.	Questions
1	Where were you born?
2	How many years did you live in your first home?
3	What is your mother's job?
4	What is your father's job?
5	Have your parents divorced?
6	Have you ever broken a bone?
7	Do you have allergies to any foods?
8	Have you ever stayed overnight in a hospital as a patient?
9	Have you ever tweeted? (posted a message on twitter)
10	Have you ever bought anything on eBay?
11	Do you own an e-reader of any kind?
12	Who was the last person you were in a physical fight with?
13	Have you ever gotten into trouble with the police?
14	Who ended your last romantic relationship?
15	Whom do you love more, your mother or father?
16	What is the most you have ever spent on a pair of shoes?
17	What is the last movie you saw that you really hated?
18	Have you ever gone ice-skating?
19	Do you currently own a tennis racket?
20	How many roommates do you have?
21	If you attended college, what was your major?
22	Did you ever have a cat?
23	Have you ever watched a person or pet die?
24	Did you ever cheat on a test in high school?

Four Units of Analysis: Small to Large

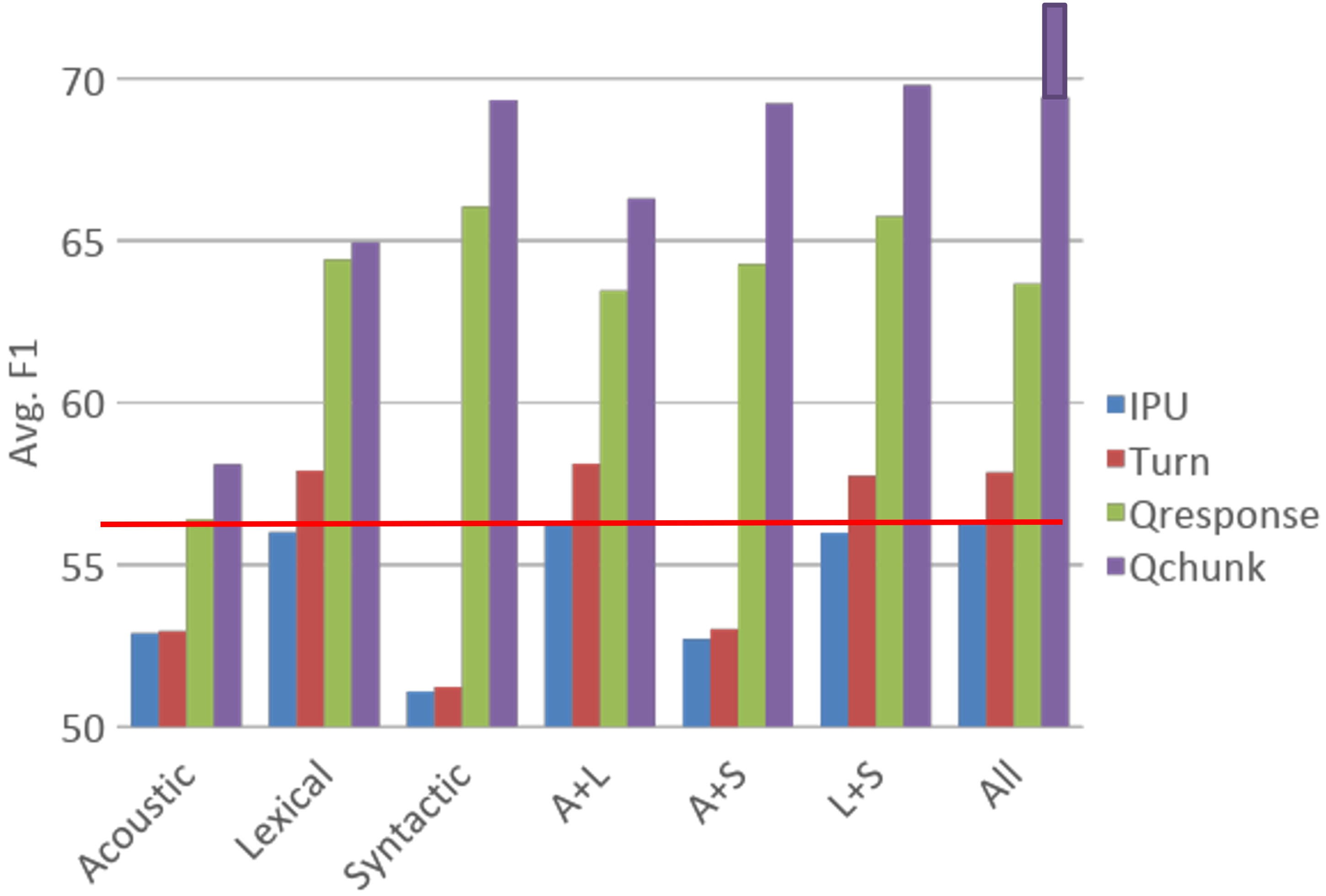
Inter-Pausal Unit (IPU) Pause-free segment of speech from a single speaker

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

Question response Interviewee first turn following an interviewer biographical question

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

Deception Classification: Our Classifiers vs. Human Interviewer Performance



“Did you ever cheat on a test in high school?”

TRUE or FALSE?

“Did you ever cheat on a test in high school?”

“Did you ever cheat on a test in high school?”

TRUE or FALSE?

“Did you ever cheat on a test in high school?”

“Who was the last person you had a physical fight with?”

True or False?

“Who was the last person you had a physical fight with?”

TRUE

“Who was the last person you had a physical fight with?”

True or False?

“Who was the last person you had a physical fight with?”

“Who was the last person you had a physical fight with?”

True or False?

“Who was the last person you had a physical fight with?”

FALSE

“Who was the last person you had a physical fight with?”

True or False?

“Who was the last person you had a physical fight with?”

How do *Humans* Decide which Answers to Trust?

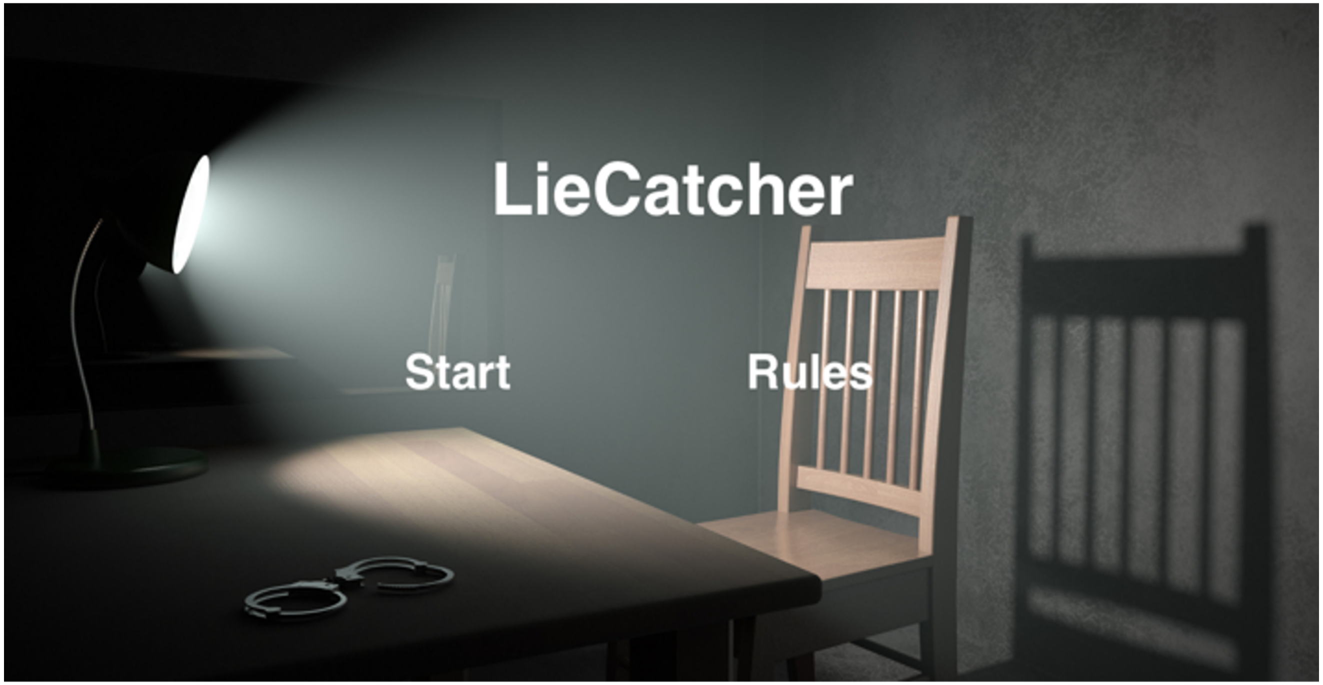
- The **words** people say?
- The **syntax** people choose?
- The **acoustic content** of their speech?
- The **prosody** of their speech?
- **Why do humans believe lies are true?**
  - How do **human decisions** compare with the features our **classifiers** use more successfully to detect lies and truth?
  - How does each compare with the **actual characteristics** of **trustworthy** (true) vs. **untrustworthy** (lying) speech?

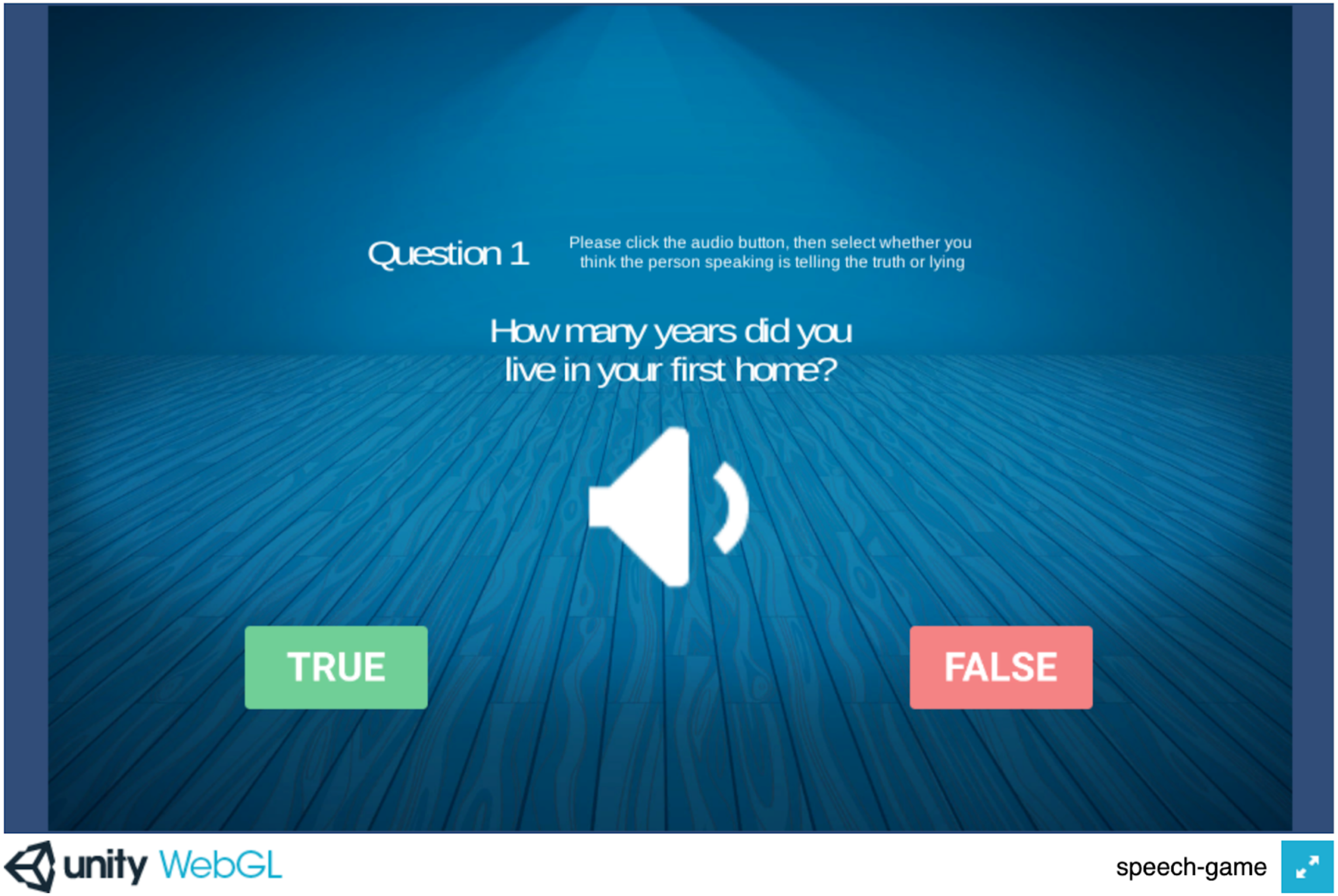
Approach

- Created a game, **LieCatcher**, to collect additional human judgments via AMT tasks
  - Each task included 12 of the 24 **questions (text)** and **first responses (speech)** of interviewees from our corpus, one at a time, plus a check question to ensure attention to game
  - Audio samples **balanced by gender, native language, question, and speaker**
  - **Balanced also for truth/ lie**: Half true, half false responses
- Collected **human judgments** with their **demographic** and **personality** (TIPI) information

- Restricted raters to **fluent speakers of English**
- Asked them about **prior experience in law enforcement**
- Must listen to **full response** before answering
- Also collected **time interval** before decision
- **Success rate** provided to raters only at end of task of each 13 question task – not for each question they rated
- Each turker limited to **10 tasks max**

LieCatcher: Game-with-a-Purpose





Question 1

Please click the audio button, then select whether you think the person speaking is telling the truth or lying

How many years did you live in your first home?

Crowdsourcing Study Procedures

- **5,340** utterances w/ **3 rater judgments** per utterance
  - **431** unique turkers: 38.9% male, 59.1% female, 2.1% unreported
  - Only 4.8% reported **previous experience in law enforcement**
  - Raters also reported at the end of the game the **features they thought indicated deception or truth**
  - Then they were told their **overall score**

amazon mechanicalturk
beta

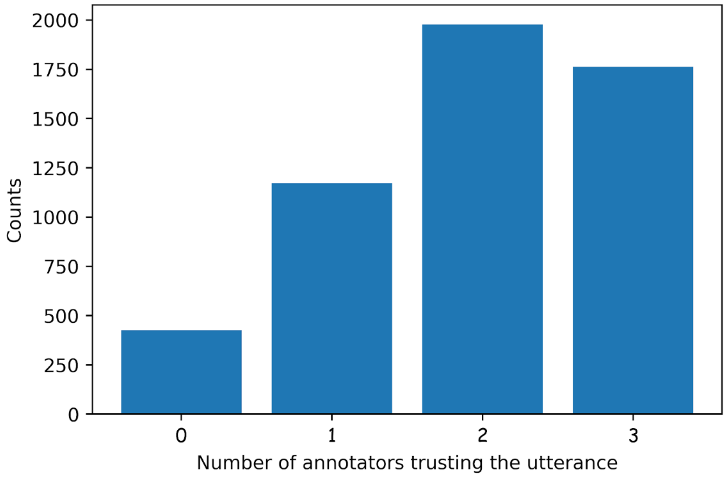
Lie Detection Ability

- Used majority vote: **Overall accuracy = 49.93%**
  - Fleiss' kappa: 0.135 (slight agreement)
- Where **all agreed**: 50.75% accuracy
  - **Agreement** did ***not*** correlate with **gender** or **native language** of speakers or with **utterance length** (with slightly lower agreement on longer responses)
- But:
  - **Female speakers *were trusted* more** (71.5%) than male speakers (68.6%)
  - **Native English speakers *were trusted* more** (73.5%) than native Chinese speakers (66.6%)

- In terms of **interviewee personality scores**: Speakers with **low conscientiousness** scores (71.9%) or **high openness to experience** scores (71.6%) or **high neuroticism** scores (71.8%) ***were trusted more*** than their opposites; **no high scores** for speech from speakers high in **extraversion** or ***agreeableness***, which was surprising

Inter-annotator Agreement

- Number of annotators agreeing on trusted utterances



- Fleiss' kappa low: 0.135
- **Truth bias** – 65% agreement on trust over-all
- **Truth Default Theory** (Timothy R. Levine, 2014)

Features of Responses Examined

- **Disfluency** “um…er”
- **Prosody/Acoustics:** pitch, speaking rate, loudness
- **Complexity:** more words, more detail
- **Affect:** positive/negative/neutral
- **Uncertainty:** e.g. hedging “sort of”, “probably”
- **Creativity:** measured in difference from other interviewees’ responses to the same question

Disfluency Features: Humans Do Fairly Well

Features	**Trusted Responses**	**Actual Deceptions**
Has filled pause	↓↓↓↓	↑↑↑↑
# filled pause	↓↓↓↓	↑↑↑↑
Response latency	↓↓↓↓	
Repetitions	↓↓↓↓	↑
False start	↓↓↓	↑↑

↓ indicates negative relationship of feature with response; ↑ indicates positive relationship

↓: <.05, ↓↓ : <.01, ↓↓↓ : <.001, ↓↓↓↓ : <.0001

Green is correct human judgment of truth or lie; red is incorrect (e.g. filled pause was correctly mistrusted but response latency was not an indicator of lies)

Acoustic/Prosodic Features: Humans Very Poor

Features	Trusted	Actual Deception
Speaking rate faster	↑↑↑↑	
Pitch max higher	↑↑↑↑	↑↑↑↑
Pitch mean higher	↑↑	
Pitch min higher	↑↑↑↑	↓↓
Pitch stdev lgr	↑↑	↑↑
Intensity max		↑↑↑
Intensity mean higher	↑↑↑↑	
Intensity min higher	↑↑↑↑	↓
Intensity std smlr	↓↓↓↓	↑
Jitter, shimmer, nhr higher	↑↑↑↑	

Humans very poor at judging lies from ***how*** people speak: **only 3 correct judgments** (higher min pitch and min intensity; smaller std of intensity)

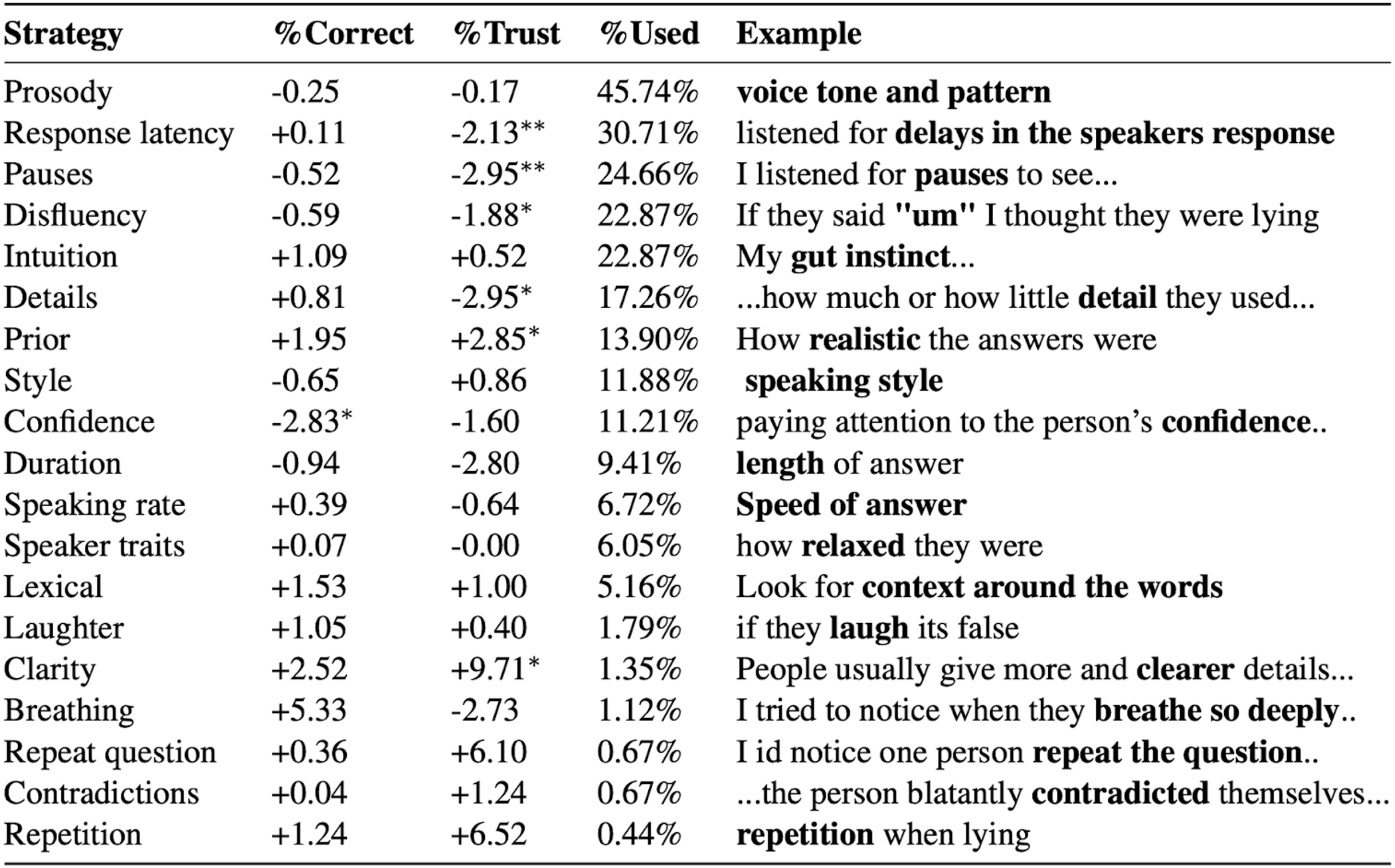
How Did Humans Do?

- **Disfluency**
  - Raters mistrusted disfluency, which was correct but also mistrusted **response latency**, which was ***not*** a cue to deception
- **Prosody**
  - Raters ***trusted*** **higher, louder and faster speech** with **higher degrees of jitter, shimmer and NHR** – but these features were ***not*** significant indicators of truth: **higher and louder speech** were in fact ***strong cues to deception***
- **Complexity**
  - Multiple measures showed raters **correctly mistrusted more complex utterances**, which, contra prior belief, ***were*** signs of deception in our data

- **Affect**
  - While raters trusted **more pleasant utterances** this was ***not*** a useful indicator of truthfulness
- **Uncertainty**
  - Raters did correctly mistrust utterances containing **hedge terms** (*possibly, sort of*) but ***did not correctly trust*** utterances indicating **certainty** (*always, never*)
- **Creativity**
  - While people were **more creative when lying** than when telling the truth, raters’ apparently ***did not recognize*** this

Analysis of Rater Strategies

- Asked raters to provide **strategies** that they used
- **Annotated** strategies by category
- **Which strategies were useful and which were not?**
  - Compared raters' **stated strategies** with their **performance** on the tasks
  - ***None of the strategies*** reported by a labeler were associated with **higher performance** than raters who did ***not*** report using those strategies in their tasks
  - **Speaker confidence** using *("the speaker sounded confident") as a cue to deception* was in fact ***negatively*** correlated with task success



Strategy	**%Correct**	**%Trust**	**%Used**	**Example**
Prosody	-0.25	-0.17	45.74%	**voice tone and pattern**
Response latency	+0.11	-2.13**	30.71%	listened for **delays in the speakers response**
Pauses	-0.52	-2.95**	24.66%	I listened for **pauses** to see...
Disfluency	-0.59	-1.88*	22.87%	If they said **"um"** I thought they were lying
Intuition	+1.09	+0.52	22.87%	My **gut instinct**...
Details	+0.81	-2.95*	17.26%	...how much or how little **detail** they used...
Prior	+1.95	+2.85*	13.90%	How **realistic** the answers were
Style	-0.65	+0.86	11.88%	**speaking style**
Confidence	-2.83*	-1.60	11.21%	paying attention to the person's **confidence**..
Duration	-0.94	-2.80	9.41%	**length** of answer
Speaking rate	+0.39	-0.64	6.72%	**Speed of answer**
Speaker traits	+0.07	-0.00	6.05%	how **relaxed** they were
Lexical	+1.53	+1.00	5.16%	Look for **context around the words**
Laughter	+1.05	+0.40	1.79%	if they **laugh** its false
Clarity	+2.52	+9.71*	1.35%	People usually give more and **clearer** details...
Breathing	+5.33	-2.73	1.12%	I tried to notice when they **breathe so deeply**..
Repeat question	+0.36	+6.10	0.67%	I id notice one person **repeat the question**..
Contradictions	+0.04	+1.24	0.67%	...the person blatantly **contradicted** themselves...
Repetition	+1.24	+6.52	0.44%	**repetition** when lying

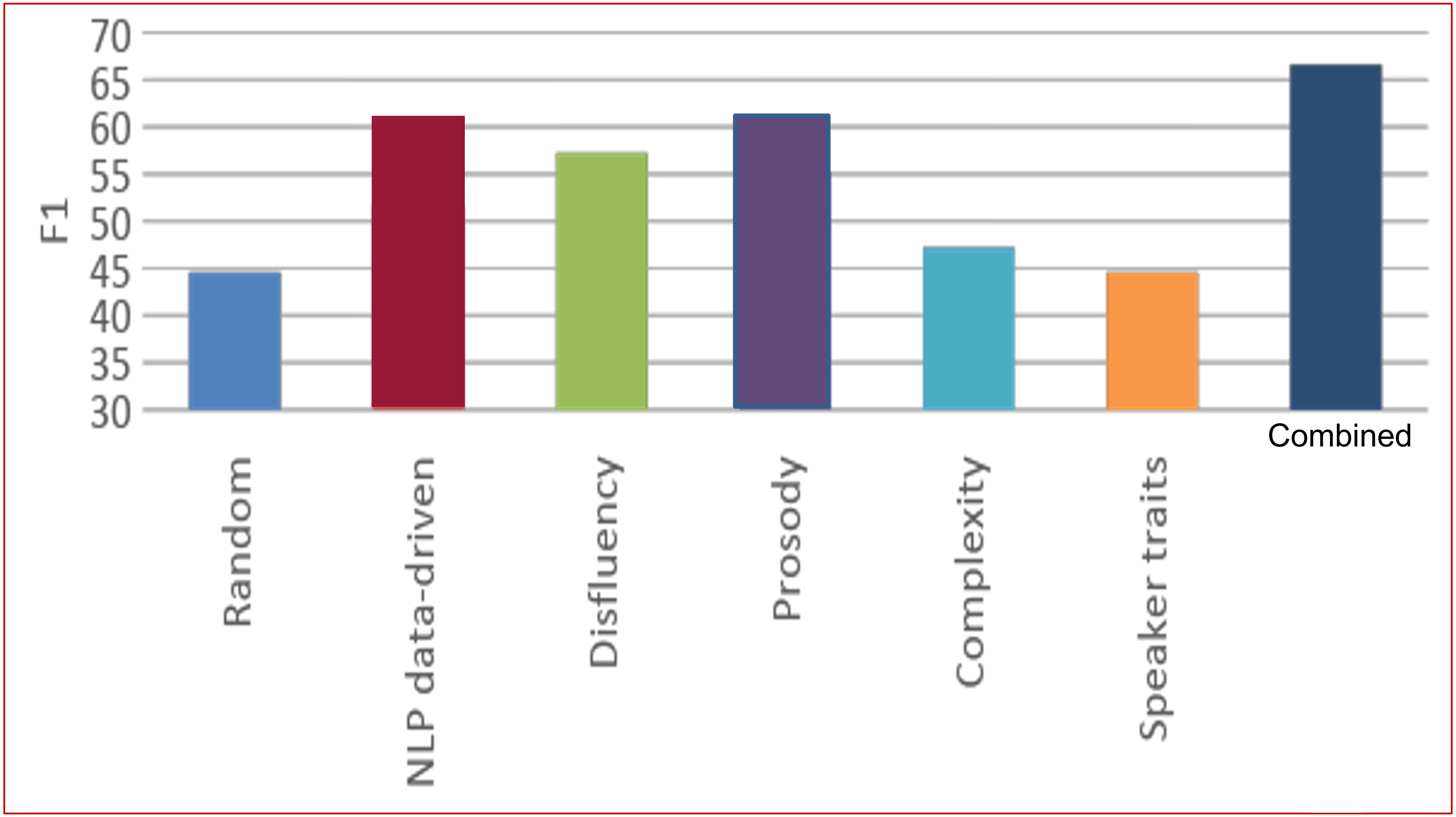
So can we Predict and Produce the Speech People Trust?

- Many **reasons to produce trustworthy speech**
  - **Dialogue systems**, spoken **information systems** and **robots**, public **travel announcements** and other broadcasts
  - **Training** for actors and newscasters and salespeople…
- Procedure to predict trusted speech
  - 5-fold cross validation, speaker independent
  - Low agreement task -> only classify utterances with **rater consensus** (of 5340: only 1762 trusted by all raters; 427 mistrusted by all raters)
  - Logistic regression; Evaluated with macro-F1 for balance
  - Baseline (random): 44.97 macro F1

What Features Can Identify Trusted Speech?

- **NLP** data-driven features
  - **GloVe *embeddings***
  - **Dependency parse n-grams**
  - **Word n-grams**
- **Our findings** on human deception/ trust differences
  - **Disfluencies**
  - **Complexity scores**
  - **Prosodic features:** openSMILE 2013 6373 features
  - **Speaker traits**: gender, native language, personality

Trust Classification Results: Feature Sets Alone and Best Combined Model



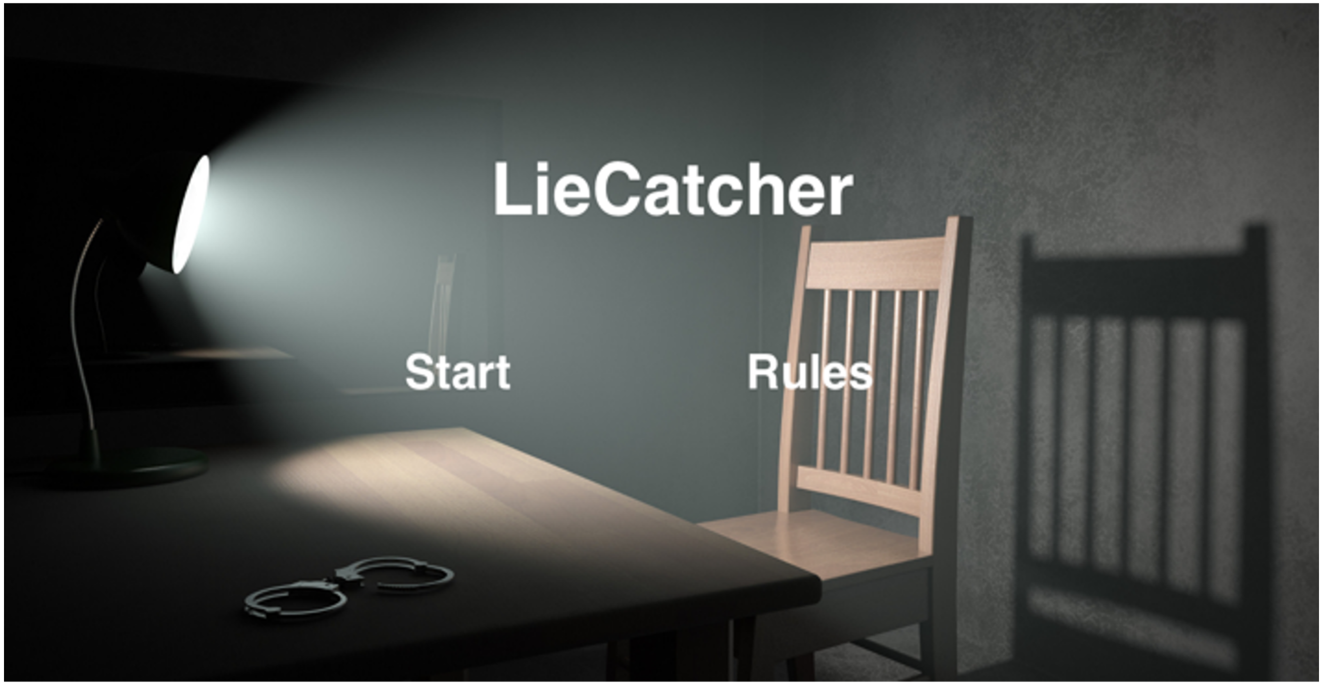
Characteristics of Successful Lies: Human and Machine Learning Models

- **Deceiving our raters**: shorter, louder, faster, fewer filled pauses, fewer repetitions, less variation in intensity and harsher (unstable amplitude) in VQ
- **Deceiving our ML models**: creativity, fewer filled pauses, less specific, shorter, lower pitch, less "concrete" (referring to perceptual entity)
- **Features that *fooled both***: Successful lies had fewer filled pauses and were shorter in duration and number of sentences
  - ***Different types of lies were successful at deceiving humans vs. ML models: So perhaps we need both?***

Why are Humans So Poor at Detecting Lies?

- Prior work on our full CXD Corpus found that **even with more data** (full responses to their questions)
  - Humans achieved only **55.33 F1** (on full Q/A conversations)
  - But our best ML models on these achieved **73.67 F1**
- **Why** do ML-trained systems do better?
  - **Mismatch for humans** between **language they trust** and **actual truthfulness**
  - Ineffective reported strategies – especially in the acoustic/prosodic domain
  - ***Can we train humans to do better?***

Current Research: LieCatcher for Lie Detection *Training* on Our Data *with Help*



But… Unless People *can* Learn to Detect Deception Better…

- What sort of speech should our ***Dialogue Systems*** employ?
- Do we want to produce speech that we know humans ***do trust***, whether correctly or not?
- Or do we want to produce speech that actually represents speech that we ***should*** trust?
- This remains a real challenge for the dialogue community and for Ethical AI…

Thank you!

Questions?