Homeland Defense & Security
Information Analysis Center

Truth or Lie?

Spoken Features of Trusted and Mistrusted Speech

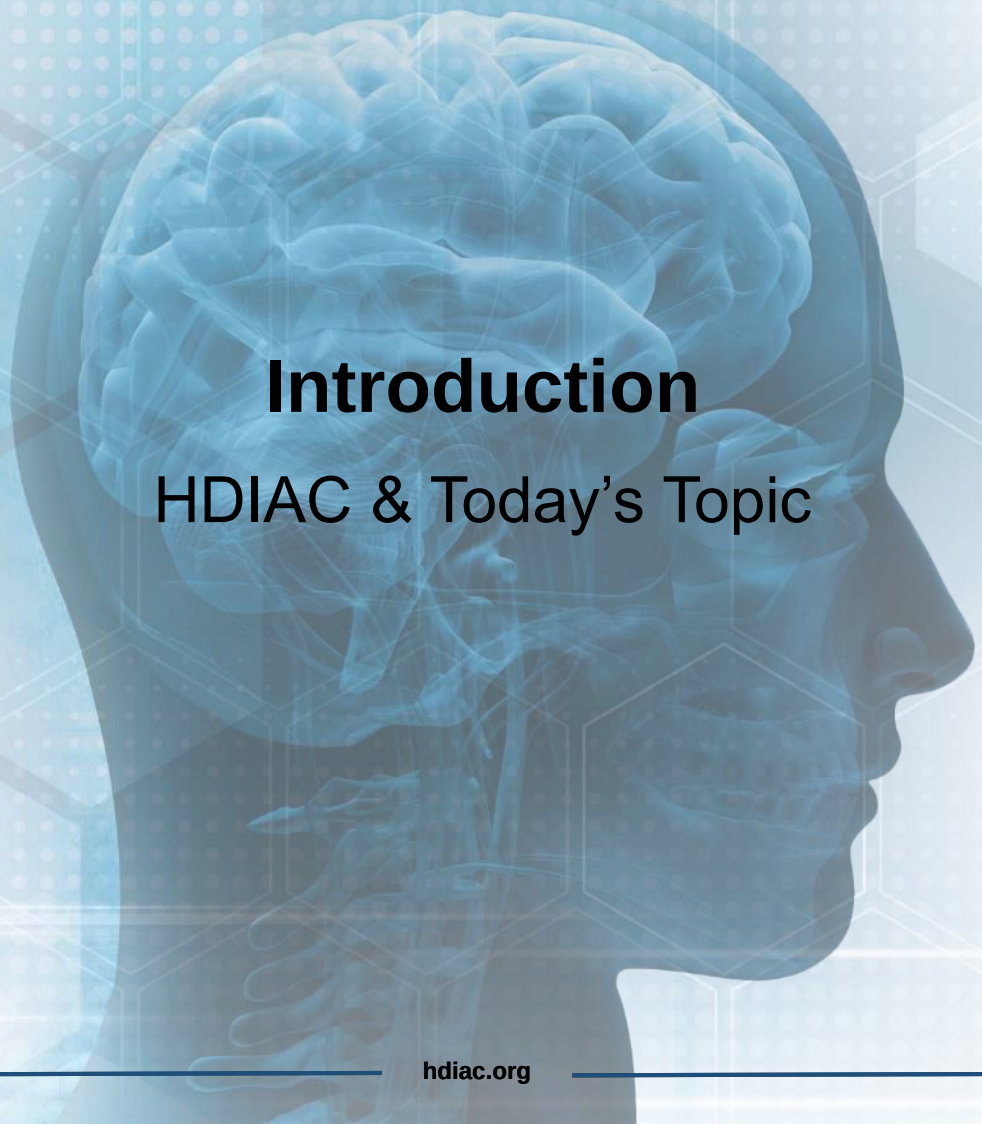
Sarah Ita Levitan, Ph.D.
Julia Hirschberg, Ph.D.

Columbia University

July 25, 2019

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

The views presented are those of the speaker and do not necessarily represent the views of DoD or its components.



Introduction

HDIAC & Today's Topic

HDIAC Overview

What is the Homeland Defense & Security Information Analysis Center (HDIAC)?

One of three Department of Defense Information Analysis Centers

Responsible for acquiring, analyzing, and disseminating relevant scientific and technical information, in each of its eight focus areas, in support of the DoD and U.S. government R&D activities

HDIAC's Mission

Our mission is to be the go-to R&D/S&T and RDT&E leader within the homeland defense and security (HDS) community, by providing timely and relevant information, superior technical solutions, and quality products to the DoD and HDS Communities of Interest/Communities of Practice.

HDIAC Overview

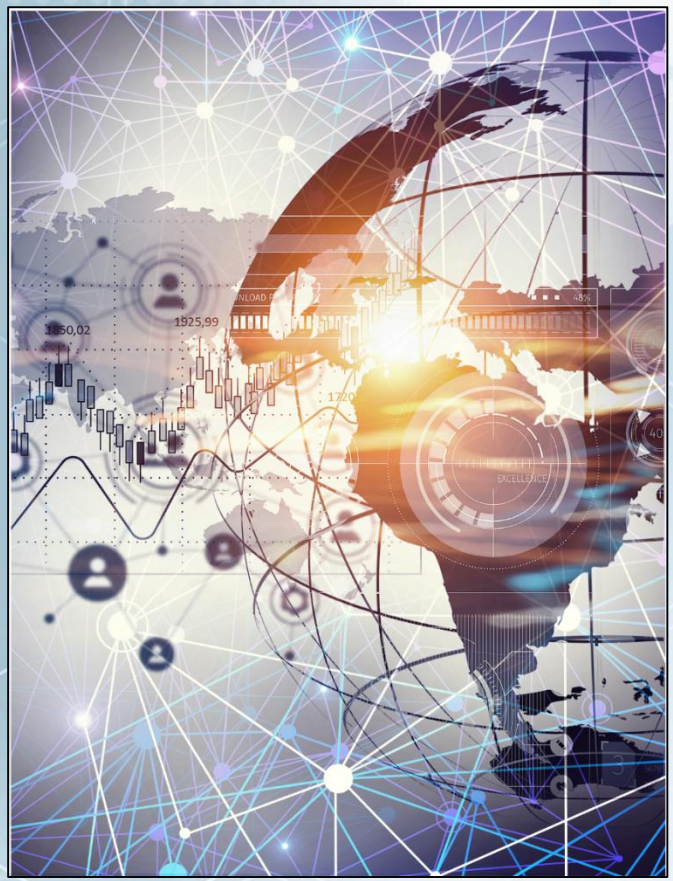
HDIAC Subject Matter Expert (SME) Network

HDIAC SMEs are experts in their field(s), and, typically, have been published in technical journals and publications.

SMEs are involved in a variety of HDIAC activities

- Authoring HDIAC Journal articles
- Answering HDIAC Technical Inquiries
- Engaging in active discussions in the HDIAC community
- Assisting with Core Analysis Tasks
- Presenting webinars

If you are interested in applying to become a SME, please visit HDIAC.org or email info@hdiac.org.



Presenters

Julia Hirschberg is Percy K. and Vida L. W. Hudson Professor of Computer Science at Columbia University. She was department chair from 2012-2018. She previously worked at Bell Laboratories and AT&T Labs where she created the HCI Research Department. She has been editor of *Computational Linguistics* and *Speech Communication*, is a fellow of AAAI, ISCA, ACL, ACM, and IEEE, and a member of the National Academy of Engineering and the American Academy of Arts and Sciences. She received the IEEE James L. Flanagan Speech and Audio Processing Award and the ISCA Medal for Scientific Achievement. She currently serves on the IEEE Speech and Language Processing Technical Committee, is co-chair of the CRA-W Board, and has worked for diversity for many years at AT&T and Columbia. She works on spoken language processing and NLP, studying text-to-speech synthesis, spoken dialogue systems, entrainment in conversation, detection of deceptive and emotional speech, hedging behavior, and linguistic code-switching (language mixing).

Sarah Ita Levitan is a postdoctoral Research Scientist in the Department of Computer Science at Columbia University. Her research interests are in spoken language processing, and she is currently working on identifying acoustic-prosodic and linguistic indicators of trustworthy speech, as well as identifying linguistic characteristics of trustworthy news. She received her PhD in Computer Science at Columbia University, advised by Dr. Julia Hirschberg, and her dissertation addressed the problem of automatic deception detection from speech. Sarah Ita was a 2018 Knight News Innovation Fellow and a recipient of the NSF Graduate Research Fellowship and the NSF IGERT From Data to Solutions fellowship. She previously worked as a graduate research summer intern at Google and at Interactions LLC.

Overview

Outline

Introduction

Deceptive speech

- Corpus collection, annotation, feature extraction
- Automatic deception detection
- Individual differences in production and perception of lies

Trusted vs. mistrusted speech

- Crowd-sourced ratings of our deception data
- Comparing mistrusted speech with actual lies
- Automatic classification of trust and mistrust

Deceptive Speech

Deliberate choice to mislead

- ***Without*** prior notification
- To gain some ***advantage*** or to avoid some ***penalty***

***Deception does not include*:**

- Self-deception, delusion, pathological behavior
- Theater
- Falsehoods due to ignorance/error

Everyday (White) Lies **very hard to detect**

But *Serious Lies may* be easier…

Why might Serious Lies be easier to detect?

***Hypotheses* in research and among practitioners:**

- Our ***cognitive load*** is increased when we lie because…
  - We must keep our story straight
  - We must remember what we ***have*** and ***have not*** said
- Our ***fear of detection*** is increased if…
  - We believe our target is difficult to fool
  - Stakes are high: serious rewards and/or punishments

All this makes it hard for us to control potential indicators of deception

Humans are Very Poor at Detecting Lies

(Aamodt & Mitchell 2004 Meta-Study)

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
Cops	8	511	55.16
Federal officers	4	341	54.54
Students	122	8,876	54.20
Detectives	5	341	51.16
Parole officers	1	32	40.42

Current Approaches to Deception Detection

'Automatic' methods **(polygraph, commercial products) no better than chance**

Human training**: e.g., John Reid & Associates**

- Behavioral Analysis: Interview/Interrogation: no empirical support, e.g.
- Truth*: I didn't take the money* vs. Lie: *I did not take the money* **(but non-native speakers use contractions less….)**

Laboratory studies**: Production and perception (facial expression, body posture/gesture, statement analysis, brain activation, odor,…)**

Our Goal

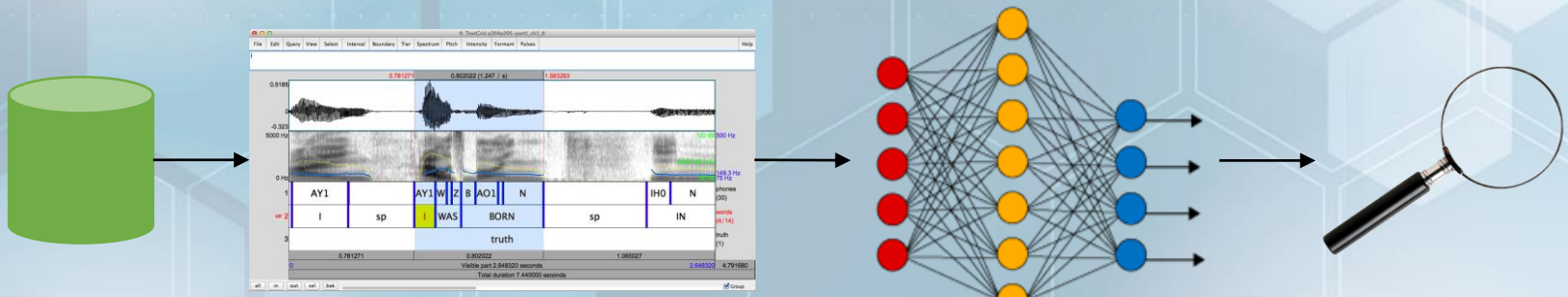
Conduct objective experiments on human subjects **to identify** ***spoken language*** **cues to deception**

Collect speech data and extract ***acoustic-prosodic*****, and** ***lexical cues*** **automatically**

Examine ***Individual Differences*****: Take** ***gender, ethnicity, culture, and personality factors*** **into account as features in classification**

Use ***Machine Learning*** **techniques to train models to classify deceptive vs. non-deceptive speech and** ***use these to improve deception detection*** **by humans by creating better methods of identifying the** ***subtle cues humans may miss*** **and** ***training humans*** **as well:** ***Collaborative AI***

Deception Detection from Spoken Language



Corpus collection

Annotation, feature extraction

Classification

Analysis

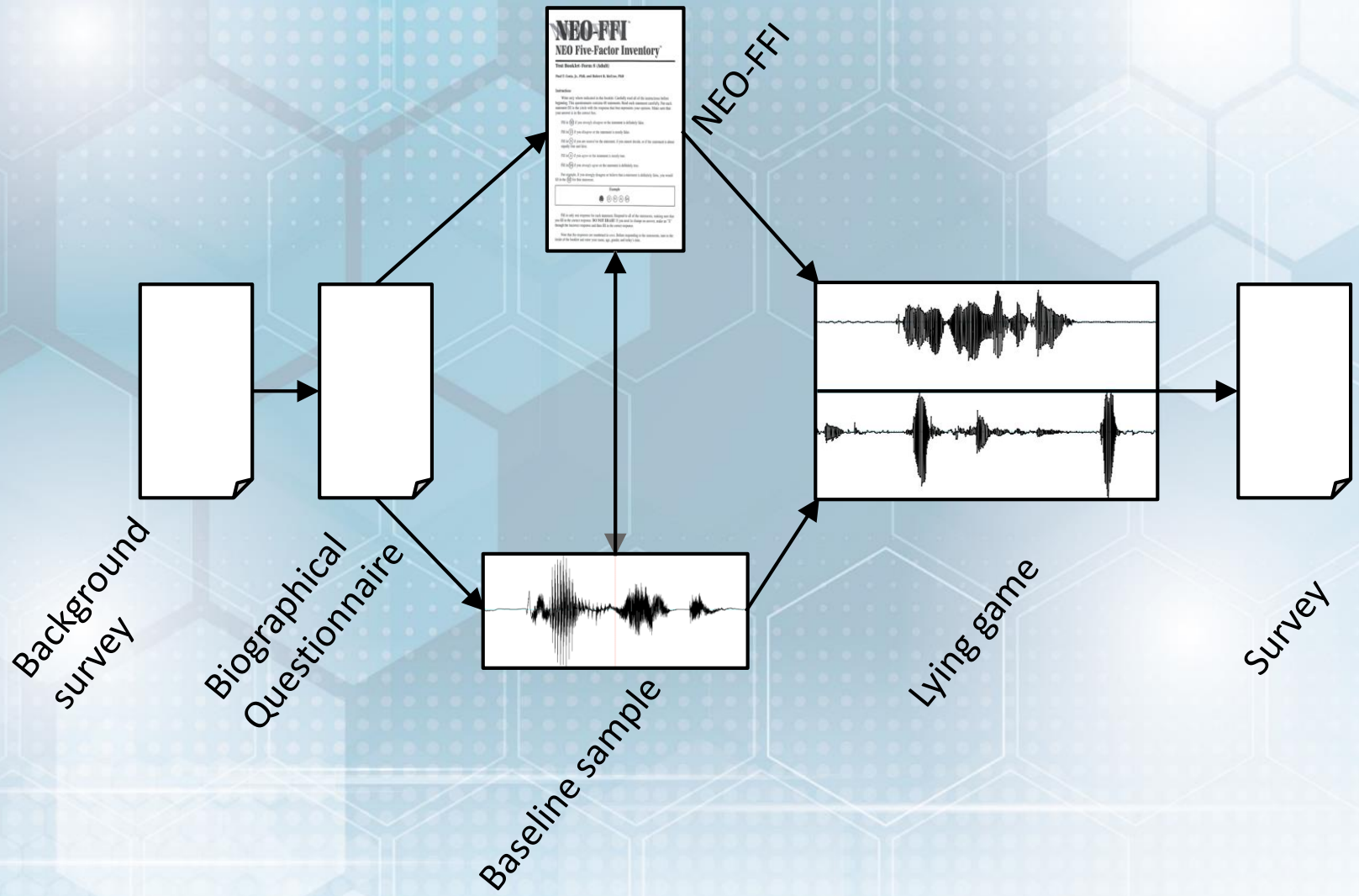
Columbia Cross-Cultural Deception Corpus (CXD)

Pair native speakers of SAE with native speakers of Mandarin Chinese, all speaking English, interviewing each other

Include

- ***Gender and personality information for all subjects***
- ***Compare subjects with different cultural and language backgrounds***

Our CXD Experiment



The Big Five NEO-FFI (Costa & McCrae, 1992)

- Openness to Experience: "I have a lot of intellectual curiosity."
- Conscientiousness: "I strive for excellence in everything I do."
- Extraversion: "I like to have a lot of people around me."
- Neuroticism: "I often feel inferior to others."
- Agreeableness: "I would rather cooperate with others than compete with them."

Our CXD Experiment

Motivation and Scoring

Monetary motivation

- **Success for interviewer**:
  - Add $1 for every correct judgment, truth or lie
  - Lose $1 for every incorrect judgement
- **Success for interviewee:**
  - Add $1 for every lie interviewer thinks is true
  - Lose $1 for every lie interviewers thinks is a lie

Good liars tell the truth as much as possible **when lying, so how do we know what's true or false for follow-up questions?**

- **Interviewees press T/F keys after every phrase**

Columbia X-Cultural Deception Corpus

- ***340 subjects, balanced by gender and native language (American English, Mandarin Chinese):*** 122 hours of speech

***Crowdsourced transcription*, automatic speech alignment (hand-corrected)**

Interviewee speech segmented into

- **Inter-pausal units (IPUs):** 111,479
- **Speaker turns:** 43,706
- **Question/answer sequences** (Q/1st Response and Q/Resp+follow-up): 7,418

“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?”

Features Extracted

Text-based: **n-grams, psycholinguistic, Linguistic Inquiry and Word Count (LIWC) (Pennybaker et al), word embeddings (GloVe trained on 2B tweets)**

Speech-based: **openSMILE IS09 (e.g. f0, intensity, speaking rate, voice quality)(386)**

Gender, native language, NEO-FFIs personality scores and clusters

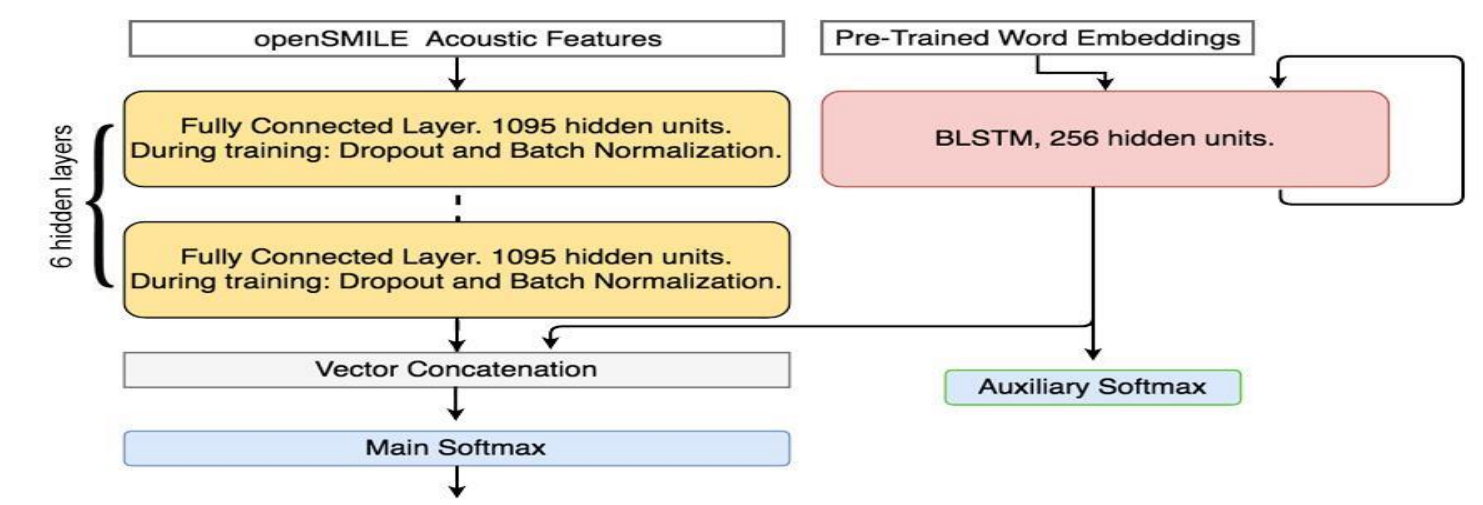
Syntactic features (complexity), entrainment, regional origin

Summary: Acoustic-prosodic and Linguistic Characteristics of Human Deception and Truth

Deception	Truth
Increased pitch & intensity max	Negation
Poor speech planning	Cue phrases
Descriptive, detailed	Cognitive process
Complex	Function words
Hedge	
Entrainment	

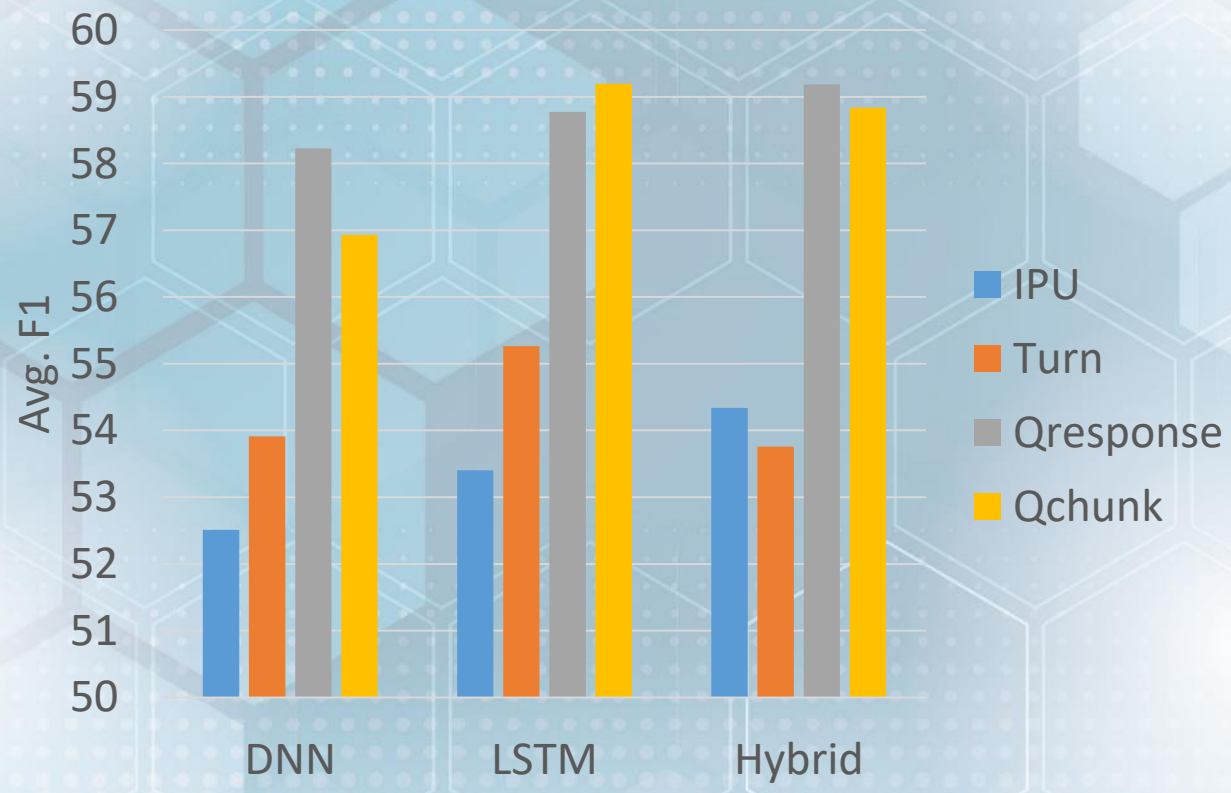
Deep Learning on Word Embeddings and openSmile Acoustic Features

- **BLSTM-word embeddings**
- **DNN-openSMILE**
- ***Hybrid: BLSTM-lexical + DNN-openSMILE***

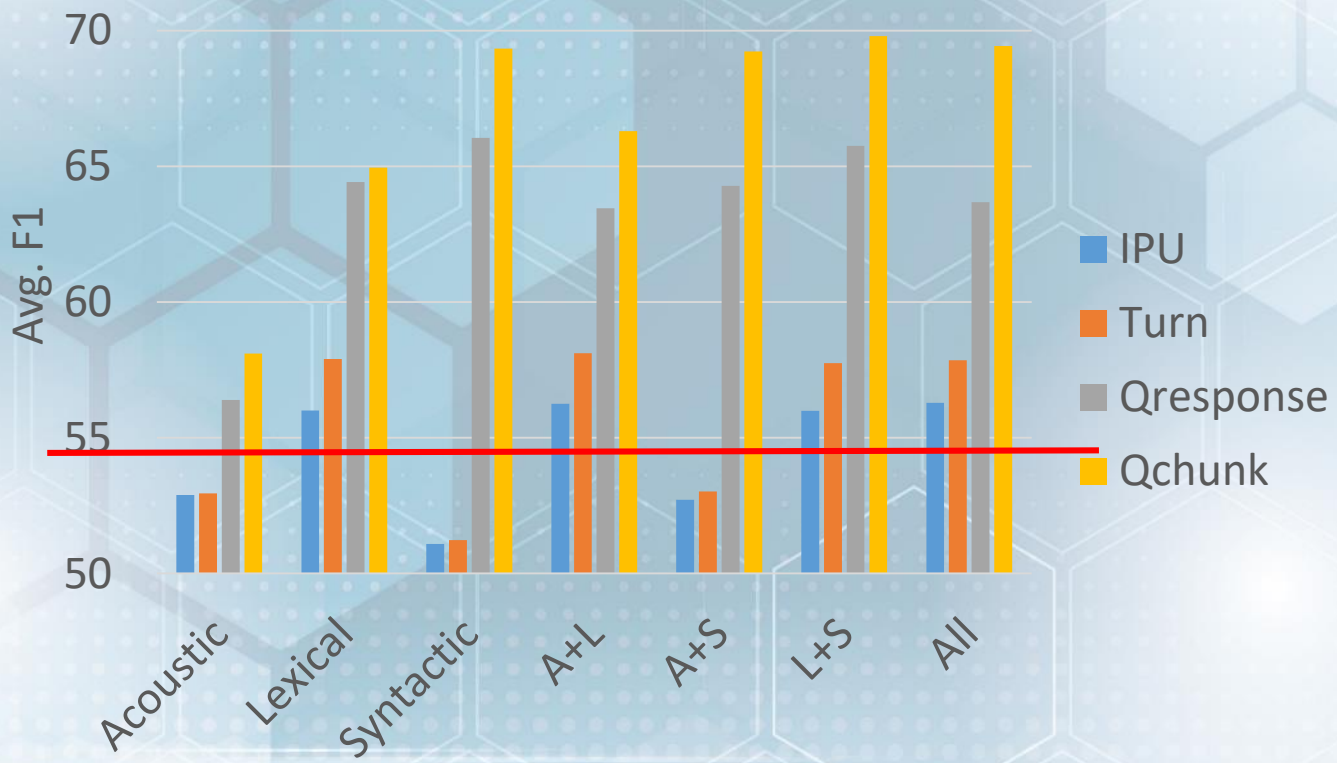


Mendels, Levitan et al. 2017, “Hybrid acoustic lexical deep learning approach for deception detection,” Interspeech, Stockholm.

Neural Network Models



Machine Learning Using Additional Features



What Next Can We Learn from Gender and Native Language?

Extract *simple acoustic/prosodic features* from question responses

***Compare distributions of features* over all and by gender and native language**

- When interview***ees*** lie vs. tell the truth
- When interview***ees*** are trusted (believed) or are not
- When interview***ers*** trust (believe) an interview***ee*** or do not

Perform *paired t-tests* to compare feature means

- Tests for significance correct for family-wise Type I error by controlling the false discovery rate at α=0.05. (Parentheses indicate an uncorrected p <=0.05.)

Individual Differences in Deceptive vs. Truthful Interviewee Speech by Gender and Native Language

Feature	Male	Female	English	Chinese	All
Pitch Max	F			F	F
Pitch Mean					
Intensity Max	F	(F)	F		F
Intensity Mean			(F)		
Speaking Rate				T	
Jitter		(T)			
Shimmer					
NHR					

Deceptive True

Gender and Native Language: Analysis of Interviewee Traits

Mistrusted **Trusted**

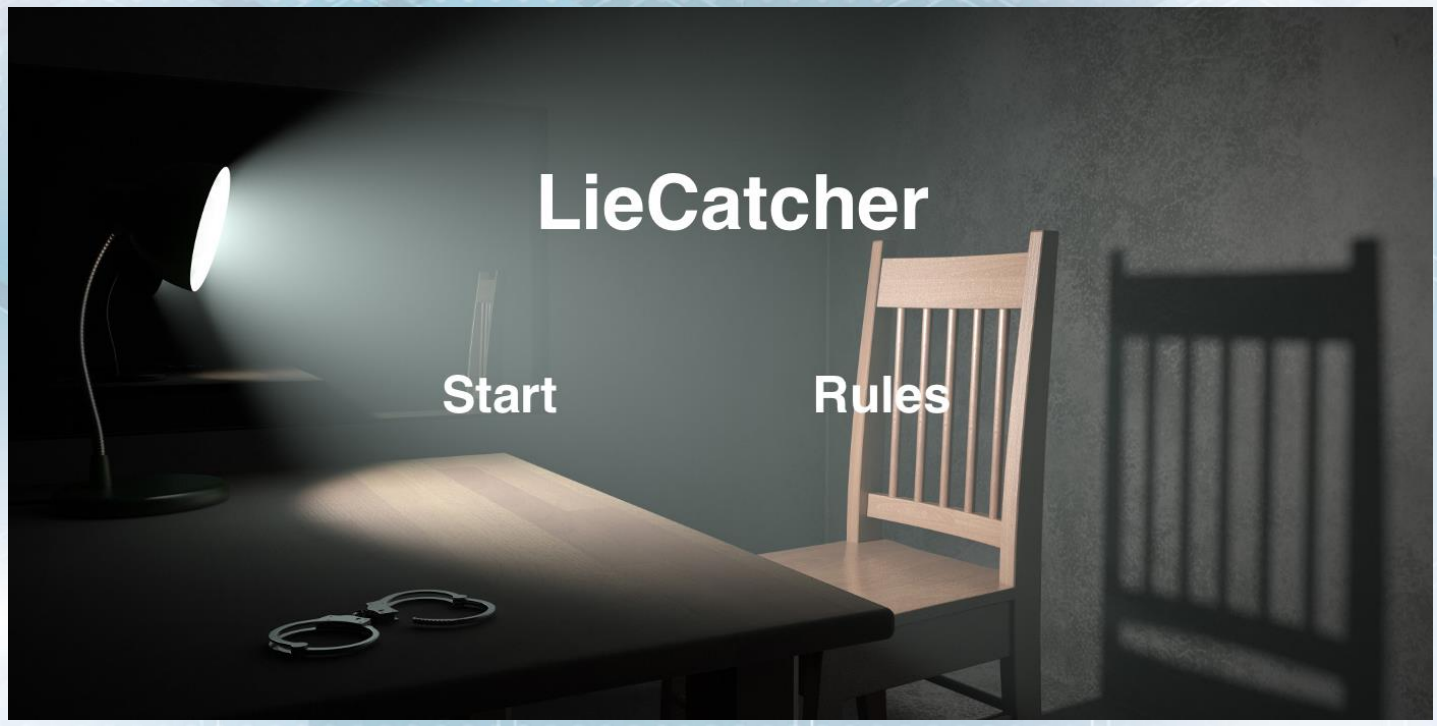
Feature	Male	Female	English	Chinese	All
Pitch Max	(F)			(F)	(F)
Pitch Mean				F	
Intensity Max				F	F
Intensity Mean					
Speaking Rate	(T)	(T)		T	T
Jitter		(T)	(T)		
Shimmer		(T)	T		
NHR				F	

Individual Differences Between *Interviewers'* Judgments by Interviewer Gender and Native Language

Feature	Male	Female	English	Chinese	All
Pitch Max			F		(F)
Pitch Mean	(F)				
Intensity Max	(F)		(F)	(F)	F
Intensity Mean					
Speaking Rate	T		T	(T)	T
Jitter		F			
Shimmer		(F)			
NHR					

Mistrusted Trusted

Games with a Purpose



Levitan et al. 2018, “LieCatcher: Game framework for collecting human judgments of deceptive speech,” LREC 2018, Miyazaki.

Crowd-sourcing Study

5,340 utterances

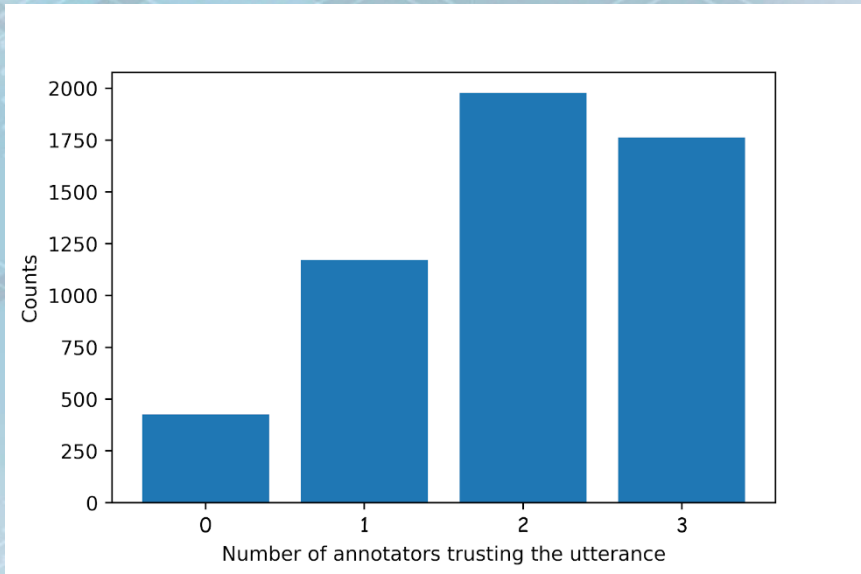
3 judgments per utterance

431 unique annotators

38.9% male, 59.1% female, 2.1% unreported

Inter-annotator Agreement

Number of annotators trusting utterances



Fleiss' kappa: 0.135

Truth bias – 65% trusted

Truth Default Theory (T.R. Levine, 2014)

Lie Detection Ability

Overall accuracy = 49.93% - below random chance!

Females are better, and take longer to judge

People with jobs related to lie detection do not perform better, and take longer to judge

Characteristics of Trusted Speakers

Gender Female speaker were trusted more than male speakers ($X^2(1)=5.1$, N=5340, $p<0.05$)

Native language Native English speakers were trusted more than native Chinese speakers ($X^2(1)=30.22$, N=5340, $p<0.00001$)

Personality

- Low Conscientiousness is most trusted
- High Openness to Experience is most trusted
- High Neuroticism is most trusted!

Why are people so poor at lie detection?

Compare features of raters' trusted/mistrusted speech with features of actual deceptive/truthful speech

Features Examined

Disfluency: “um…er”

Complexity: more words, more detailed

Affect: sentiment

Uncertainty: “sort of”, “probably”

Creativity: difference from “standard” responses for same question

Prosody: pitch, speaking rate, loudness

Disfluency

Theory: lie-telling is more cognitively demanding than truth-telling

Features
Has filled pause
filled pause
Response latency
Repetition
False start

Disfluency

Theory: lie-telling is more cognitively demanding than truth-telling

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

Disfluency

Theory: lie-telling is more cognitively demanding than truth-telling

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

Prosody

Features
Duration
Speaking rate
Pitch max
Pitch min
Intensity max, mean
Intensity min
Intensity std
Jitter, shimmer, nhr

Prosody

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

Prosody

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

How to Tell a Believable Lie

Features	Successful Lie
Duration	↓↓↓↓
Speaking rate	↑↑↑↑
Response latency	↓↓↓↓
Intensity mean	↑↑↑↑
Repetition	↓↓↓↓
Filled pauses	↓↓↓↓

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

Can We Predict Trusted Speech?

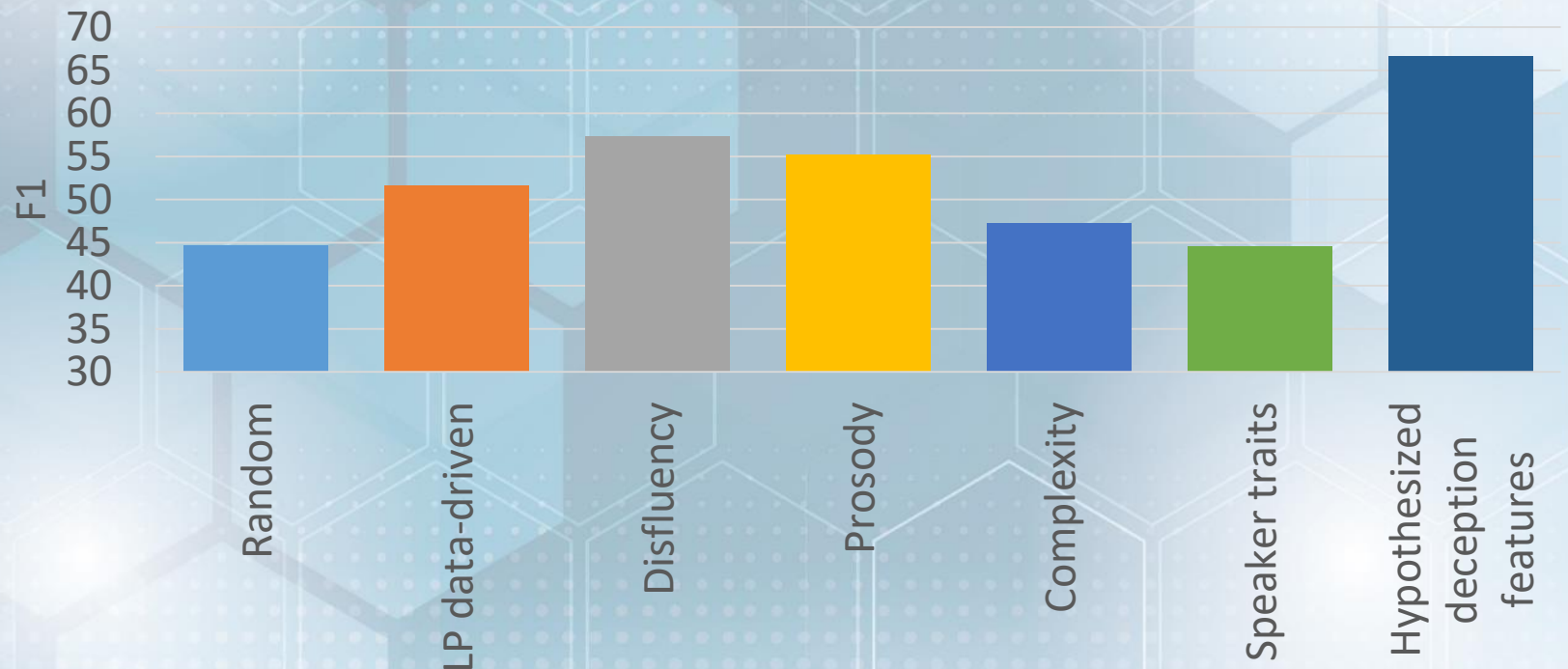
NLP Data-driven features

- GloVe embeddings
- Dependency parse n-grams
- Word n-grams

Hypothesized deception/trust features

- Disfluency
- Complexity
- Prosody
- Speaker traits

Can We Predict Trusted Speech?



Summary: Trusted Speech

Subjective task

Characteristics of trust vs. deception

Individual differences

Trust classification: 66.62 F1

Why people are bad at lie detection:

- Mismatch between features of trusted and truthful speech

Conclusion

We can automatically identify deception using acoustic-prosodic, lexical, personality and demographic features – much better than humans

We can also identify speech that humans trust and mistrust and understand the reasons for the mismatch between perceived deception and actual lies

Future research:

- Generating trusted speech automatically
- Developing techniques and software to train humans in identifying lies

“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?”

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

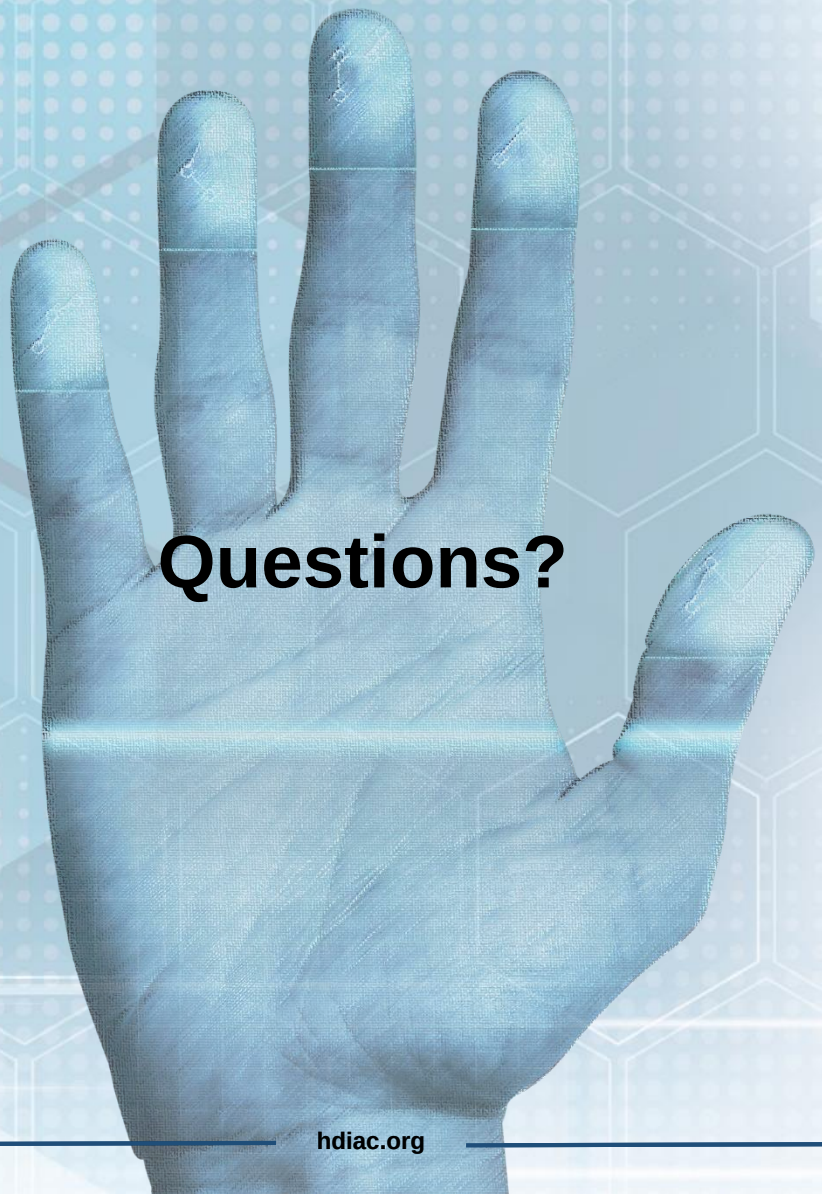
True or False?

“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?”



Questions?