Gender, Ethnicity, and Personality Factors in Deceptive Speech Detection

Sarah Ita Levitan, Julia Hirschberg

Computer Science

Columbia University

5 February 2016

Collaborators

- Guozhen An, Michelle Levine, Andrew Rosenberg
- And thanks to: Zoe Baker-Peng, Lingshi Huang, Leighanne Hsu, Bingyan Hu, Melissa Kaufman-Gomez, Yocheved Levitan, Gideon Mendels, Yvonne Missry, Elizabeth Petitti, Sarah Roth, Molly Scott, Jennifer Senior, Grace Ulinski, Christine Wang, Mandi Wang

Outline

- Motivation and previous work
  - Defining deception
  - Previous studies
  - Cues to deception
  - Human ability to detect deception
- Current cross-cultural studies The experiment
  - Role of ethnicity, gender, personality factors
  - Classification studies
  - Future research

Our Definition of Deception

- Deliberate choice to mislead
  - *Without prior notification*
  - To gain some ***advantage*** or to avoid some ***penalty***
- ***Not***:
  - Self-deception, delusion, pathological behavior
  - Theater
  - Falsehoods due to ignorance/error
- Everyday (white) Lies hard to detect
- But Serious Lies?

Why are 'serious' lies difficult?

- Hypotheses:
  - Our cognitive load is increased when lying because…
    - Must keep story straight
    - Must remember what we've said ***and*** what we ***haven't*** said
  - Our fear of detection is increased if…
    - We believe our target is hard to fool or suspicious
    - Stakes are high: serious rewards and/or punishments
- Makes it hard for us to control ***indicators*** of deception

Cues to Deception: Current Proposals

- Body posture and gestures (Burgoon et al ‘94)
  - Complete shifts in posture, touching one’s face,…
- Microexpressions (Ekman ‘76, Frank ‘03)
  - Fleeting traces of fear, elation,…
- Biometric factors (Horvath ‘73)
  - Increased blood pressure, perspiration, respiration… other correlates of stress
  - Odor
- Changes in brain activity: true vs. false stories
- Variation in ***what*** is said and ***how*** (Adams ‘96, Pennebaker et al ‘01, Streeter et al ‘77)

Current Approaches to Deception Detection

- Training humans
  - John Reid & Associates
    - Behavioral Analysis: Interview and Interrogation
- Laboratory studies: Production and Perception
- `Automatic' methods
  - Polygraph
  - Nemesysco and the Love Detector
  - *No evidence that any of these work….*
    but publishing this statement can be dangerous!
    *(*Anders Eriksson and Francisco La Cerda)

What's Missing?

- ***More objective, experimentally verified*** studies of cues to deception ***which predict better than humans or polygraphs***
- ***Method:***
  - Identify acoustic, prosodic, and lexical cues that can be extracted automatically as well as simple personality features
  - Examine statistical correlations with deception
  - Use Machine Learning techniques to train models to classify deceptive vs. non-deceptive speech

Our Previous Work on Deception

- Created Columbia/SRI/Colorado Deception Corpus
  - Within-subject recordings of deceptive and non-deceptive speech
    - 32 adult native American English speakers
    - 25-50m interviews by trained interviewer (Reid technique): 15.2h of speech, 7h from subjects
    - Subjects given tasks and incentivized to lie about performance
    - Ground truth identified by subjects

Acoustic/Prosodic Features

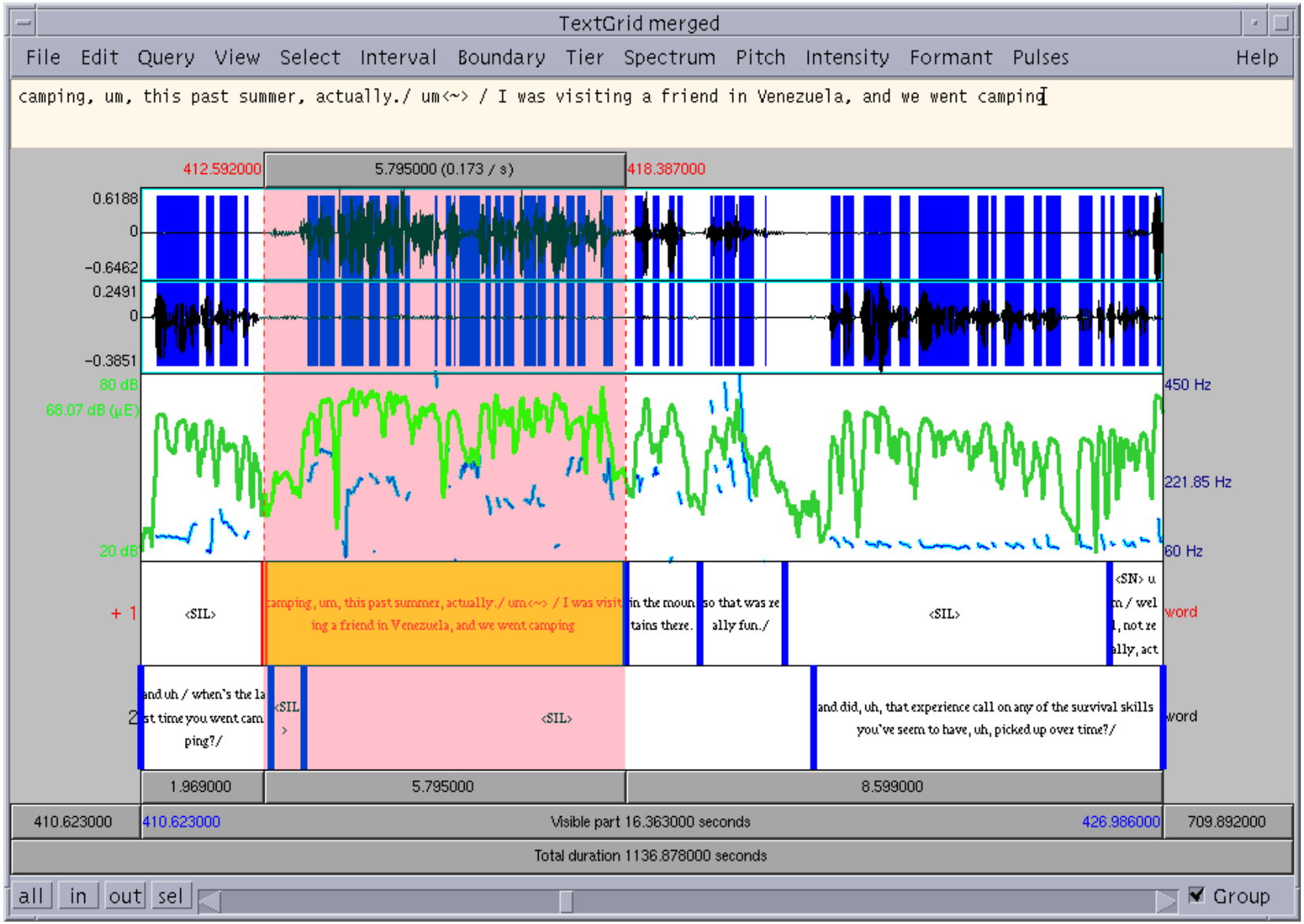
- Duration features
  - Phone / Vowel / Syllable Durations
  - Normalized by Phone/Vowel Means, Speaker
- Speaking rate features (vowels/time)
- Pause features (cf Benus et al ‘06)
  - Speech to pause ratio, number of long pauses
  - Maximum pause length
- Energy features (RMS energy)
- Pitch features
  - Pitch stylization (Sonmez et al. ‘98)
  - Model of F0 to estimate speaker range
  - Pitch ranges, slopes, locations of interest
- Spectral tilt features

Lexical Features

- Presence and # of filled pauses
- Is this a question? A question following a question
- Presence of pronouns (by person, case and number)
- A specific denial?
- Presence and # of cue phrases
- Presence of self repairs
- Presence of contractions
- Presence of positive/negative emotion words
- Verb tense
- Presence of 'yes', 'no', 'not', negative contractions
- Presence of 'absolutely', 'really'
- Presence of hedges
- Complexity: syls/words
- Number of repeated words
- Punctuation type
- Length of unit (in sec and words)
- # words/unit length
- # of laughs
- # of audible breaths
- # of other speaker noise
- # of mispronounced words
- # of unintelligible words

Subject-Dependent Features

- % units with cue phrases
- % units with filled pauses
- % units with laughter
- Lies/truths with filled pauses ratio
- Lies/truths with cue phrases ratio
- Lies/truths with laughter ratio
- Gender


TextGrid merged
File Edit Query View Select Interval Boundary Tier Spectrum Pitch Intensity Formant Pulses Help
camping, um, this past summer, actually./ um<~> / I was visiting a friend in Venezuela, and we went camping
412.592000
5.795000 (0.173 / s)
418.387000
0.6188
0
-0.6462
0.2491
0
-0.3851
80 dB
68.07 dB (µE)
20 dB
450 Hz
221.85 Hz
60 Hz
+ 1
<SIL>
camping, um, this past summer, actually./ um<~> / I was visiting a friend in Venezuela, and we went camping
in the mountains there.
so that was really fun./
<SIL>
<SN> um / well, not really, act
word
2
and uh / when's the last time you went camping?/
<SIL>
<SIL>
and did, uh, that experience call on any of the survival skills you've seem to have, uh, picked up over time?/
word
1.969000
5.795000
8.599000
410.623000
410.623000
Visible part 16.363000 seconds
426.986000
709.892000
Total duration 1136.878000 seconds
all
in
out
sel
Group

Results

- 88 features, normalized within-speaker
  - Discrete: Lexical, discourse, pause
  - Continuous features: Acoustic, prosodic, paralinguistic, lexical
- Best Performance: Best 39 features + c4.5 ML
  - **Accuracy: 70.00%**
  - **TRUTH F-measure: 75.78**
  - Lexical, subject-dependent & speaker-normalized features best predictors
  - ***Interesting individual differences: how to predict?***

Evaluation: Compared to Human Deception Detection

- Most people are very poor at detecting deception
  - ~50% accuracy (Ekman & O'Sullivan '91, Aamodt '06)
  - People use unreliable cues, ***even with training***
- Our study
  - 32 Judges, rating 2 interviews
  - Received 'training' on one subject.
- Pre- and post-test questionnaires
- Personality Inventory

A Meta-Study of Human Deception Detection

(Aamodt & Mitchell 2004)

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**

What Makes Some People Better Judges?

- Costa & McCrae (1992) NEO-FFI Personality Measures
  - **Extroversion** (Surgency). Includes traits such as talkative, energetic, and assertive.
  - **Agreeableness.** Includes traits like sympathetic, kind, and affectionate.
  - **Conscientiousness.** Tendency to be organized, thorough, and planful.
  - **Neuroticism** (opp. of Emotional Stability). Characterized by traits like tense, moody, and anxious.
  - **Openness to Experience** (aka Intellect or Intellect/Imagination). Includes having wide

Table 1: *Judges' aggregate performance classifying* **TRUTH / LIE**.

Lie Category	Chance Baseline	Mean[a]	Median	Std. Dev.	Min.	Max.
Local	63.87 [b]	58.23	57.42	7.51	40.64	71.48
Global	63.64 [c]	47.76	50.00	14.82	16.67	75.00

[a]Each judge's score is his or her average over two interviews; as percentages.
[b]Guessing **TRUTH** each time.
[c]Guessing **LIE** each time.

By Judge
58.2% Acc.

By Interviewee
58.2% Acc.

Table 1: *Aggregate performance by interviewee.*

Lie Type	Mean[a]	Median	Std. Dev.	Min.	Max.
Local	58.23	58.58	9.44	35.86	87.79
Global	44.83	45.58	17.40	10.00	81.67

[a]Each interviewee's score is the average over two judges; as percentages.

Neuroticism, Openness & Agreeableness Correlate with Judge's Performance

On Judging Global lies.

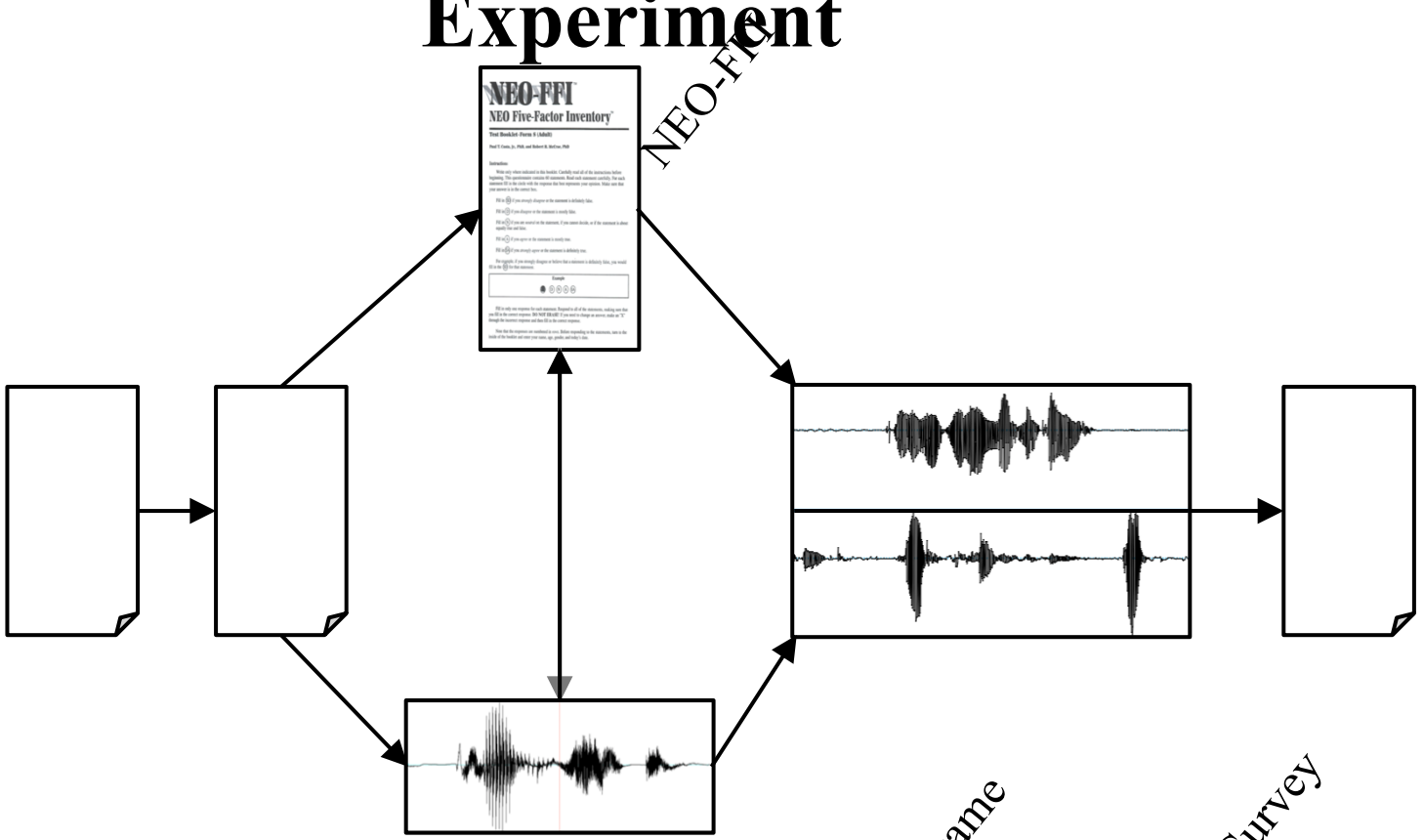
Table 1: *Correlations between personality factors and judge performance at labeling global lies.*

Factor	Measure	Pearson's corr. coef.	p-value
Neuroticism	Proportion of segments judged LIE	-0.44	0.012
Openness	Accuracy	0.51	0.003
Agreeableness		0.41	0.021
Neuroticism	F-measure for TRUTH	0.37	0.035
Agreeableness		0.41	0.019
Openness	F-measure for LIE	0.52	0.003

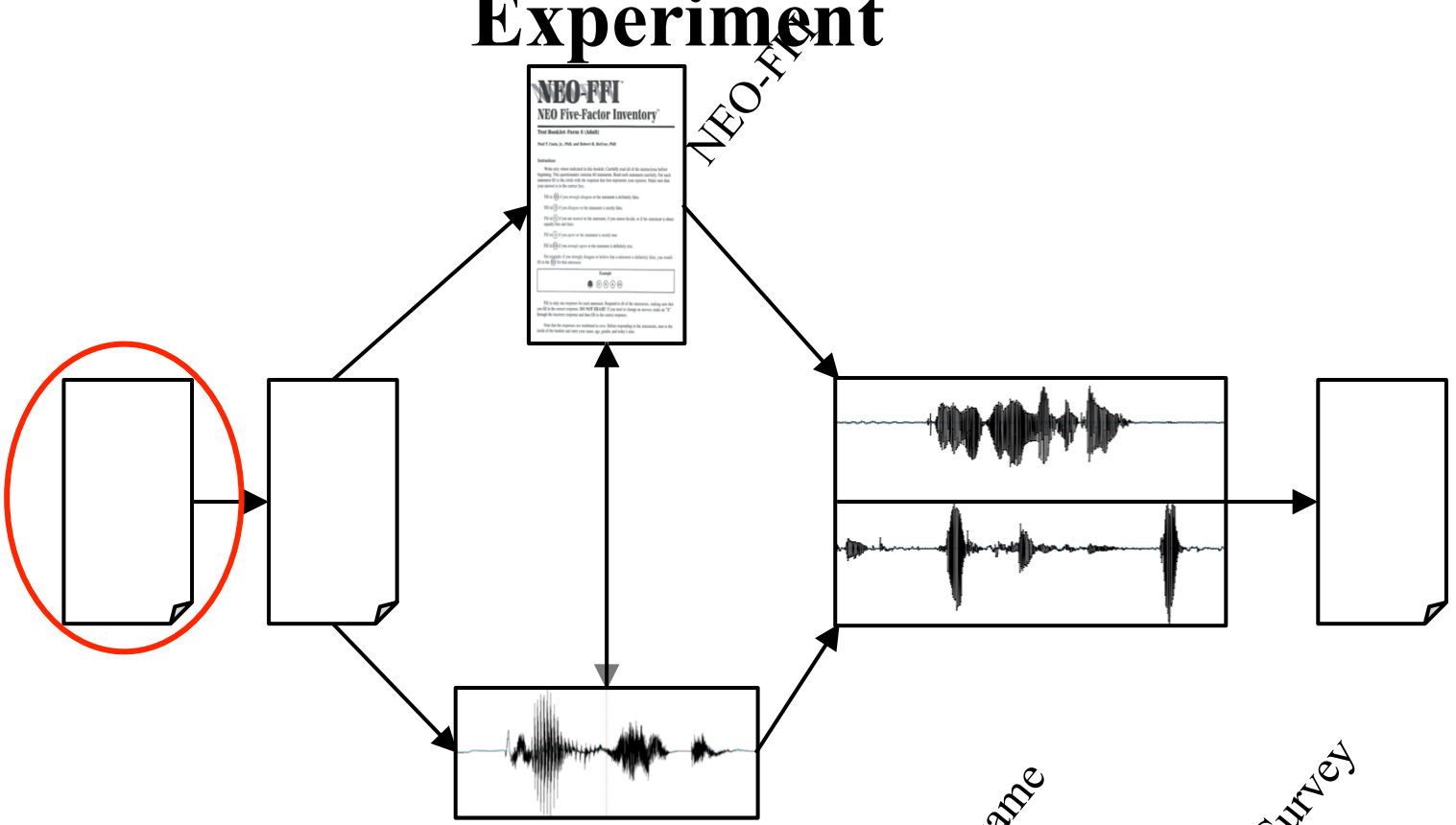
Current Study: Hypotheses

- Personality factors can help to predict differences in deceptive behavior
- Subjects who deceive better can also detect deception better
- Cultural differences and gender also play a role in deceptive behavior and in deception detection abilities
- New task: Studies of pairs of American English and Mandarin Chinese native speakers, speaking English, interviewing each other

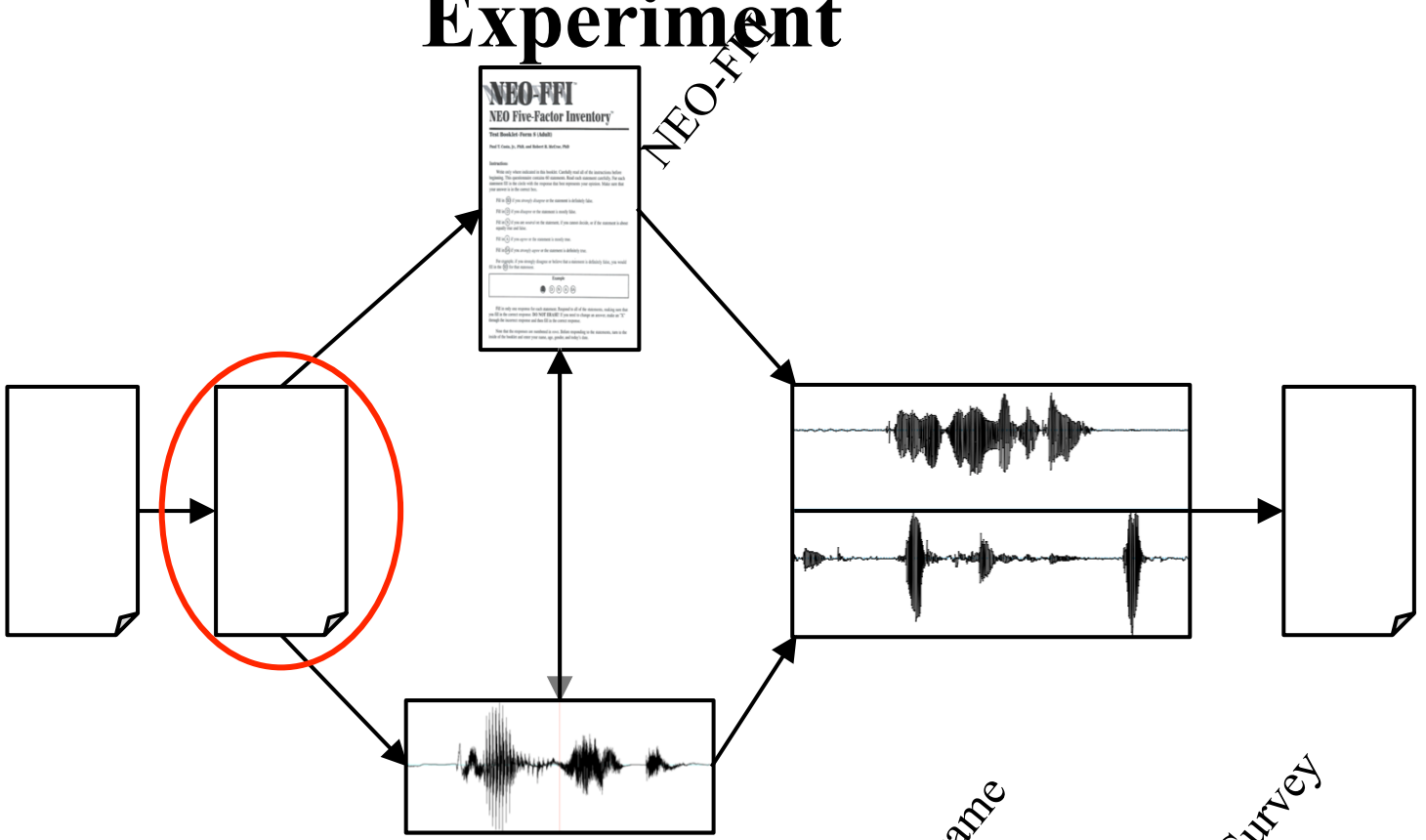
Experiment


NEO-FFI
NEO Five-Factor Inventory
NEO-FFI
Background survey
Biographical Questionnaire
Baseline sample
Lying game
Survey

Experiment


NEO-FFI
NEO Five-Factor Inventory
NEO-FFI
Background survey
Biographical Questionnaire
Baseline sample
Lying game
Survey

Experiment


NEO-FFI
NEO Five-Factor Inventory
NEO-FFI
Background survey
Biographical Questionnaire
Baseline sample
Lying game
Survey

Biographical Questionnaire

Participant No. _____ Date _____

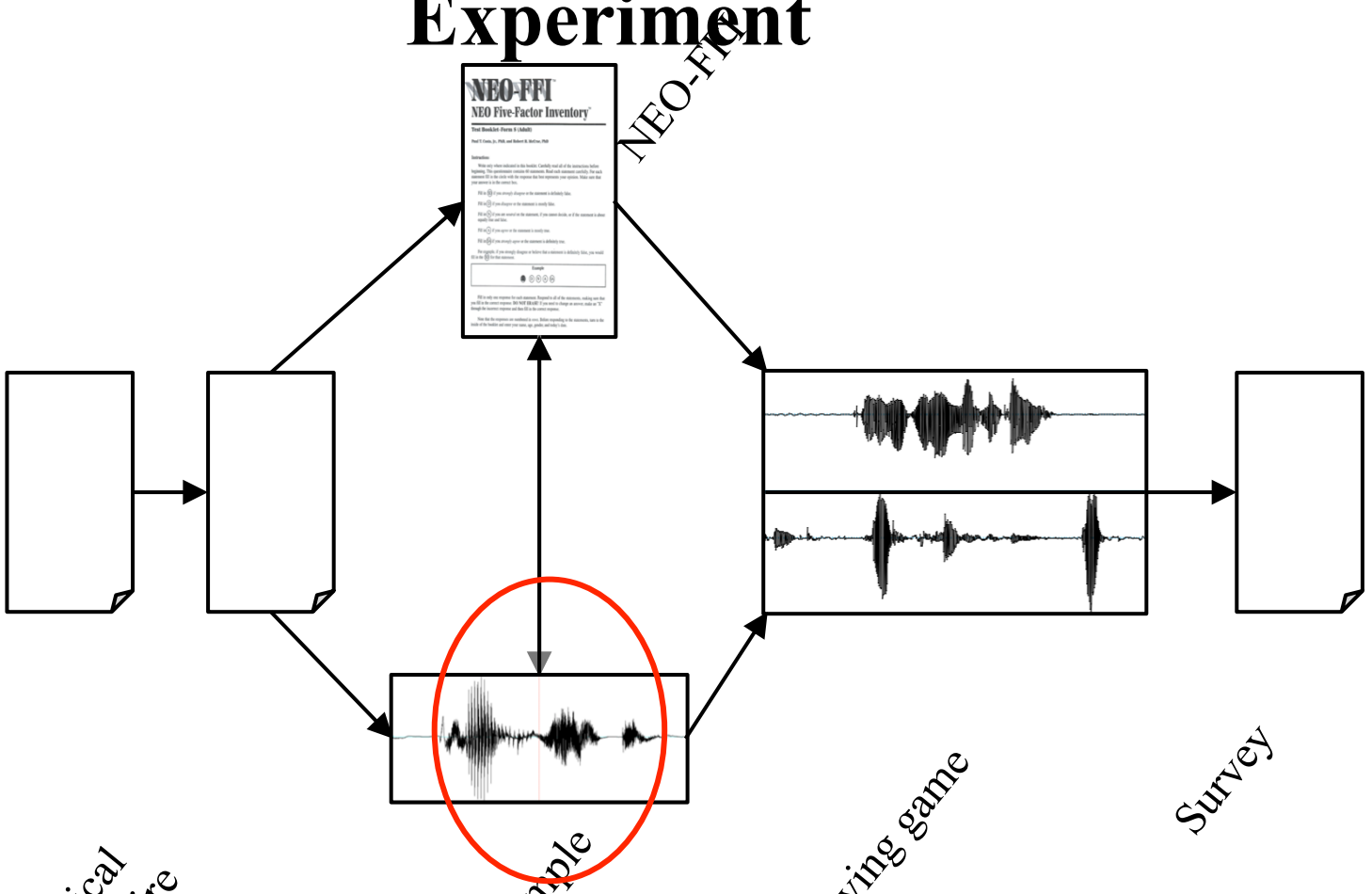
Instructions

Please carefully look through the questions. Write down the true answer to each question in the "True Answer" column. When you have finished that, for all the questions that have don't have "X"s in the "False Answer" column, make up an answer. Consult the additional sheet you have been given. You want to choose a lie that you are not as familiar with as the true answer.

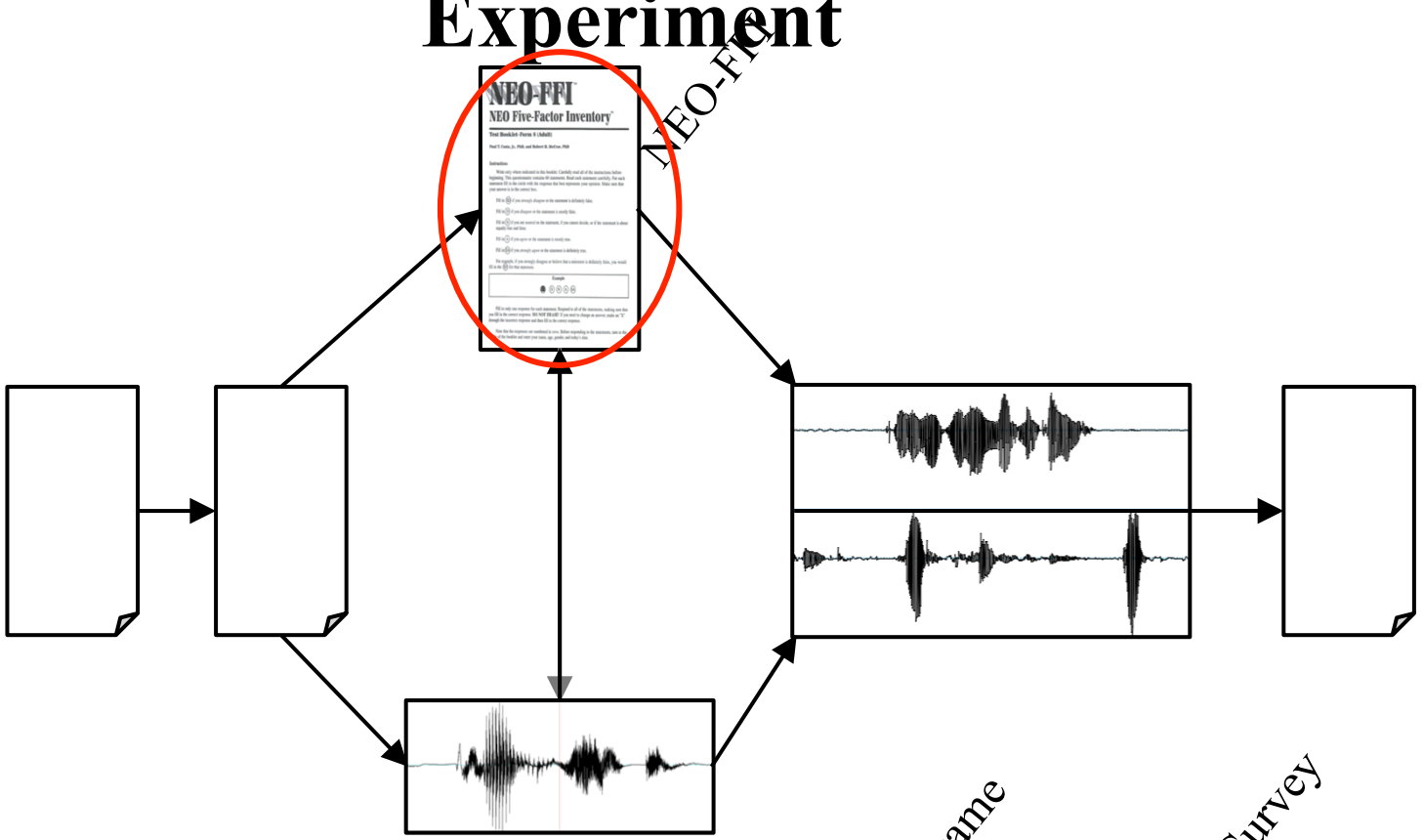
This experiment is completely anonymous- your name will never be linked to the data.

No.	Questions	True Answer	False Answer
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?		[redacted]
5	Have your parents divorced?		
6	Have you ever broken a bone?		[redacted]
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?		[redacted]
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?		[redacted]
17	What is the last movie you saw that you really hated?		

Experiment


NEO-FFI
NEO Five-Factor Inventory
NEO-FFI
Background survey
Biographical Questionnaire
Baseline sample
Lying game
Survey

Experiment


NEO-FFI
NEO Five-Factor Inventory
NEO-FFI
Background survey
Biographical Questionnaire
Baseline sample
Lying game
Survey

NEO-FFI

O C E A N

openness to experience

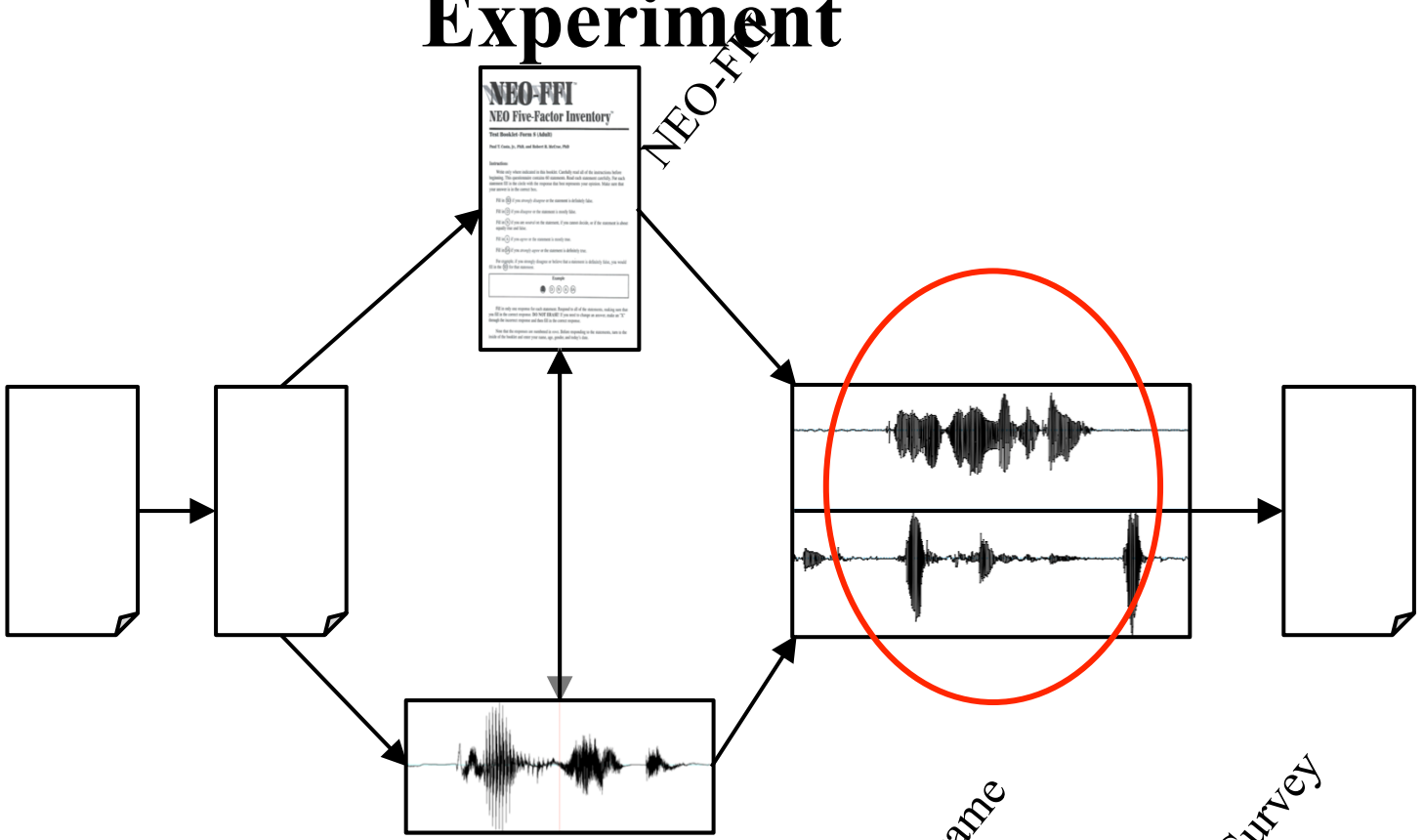
conscientiousness

extraversion

agreeableness

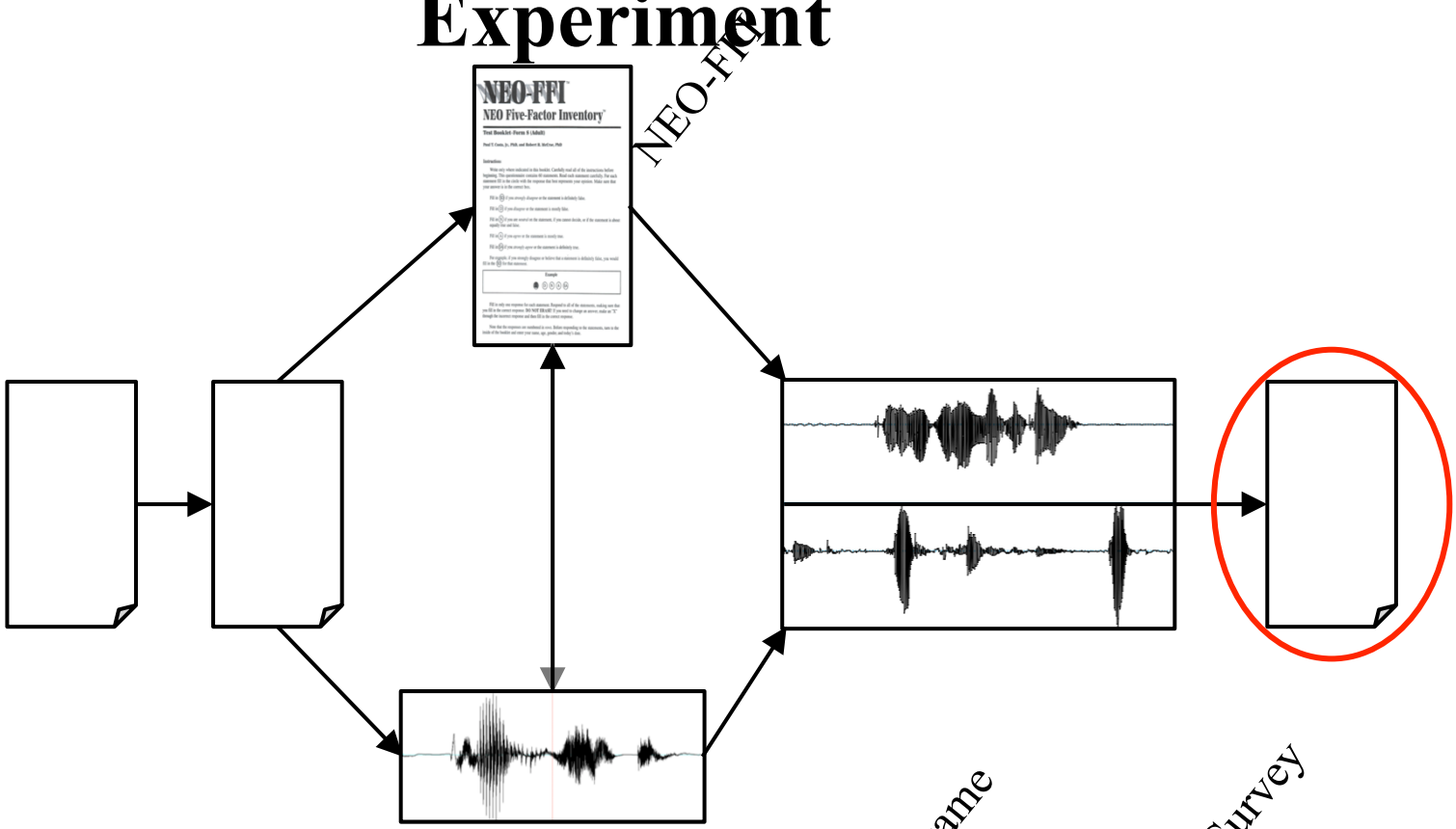
neuroticism

Experiment


NEO-FFI
NEO Five-Factor Inventory
NEO-FFI
Background survey
Biographical Questionnaire
Baseline sample
Lying game
Survey

Experiment

Experiment


NEO-FFI
NEO Five-Factor Inventory
NEO-FFI
Background survey
Biographical Questionnaire
Baseline sample
Lying game
Survey

Scoring and Motivation

- Success
  - Ability to lie -> as interviewee, number of lies believed true by interviewer
  - Ability to detect lies -> as interviewer, number of correct guesses for truth and lie
- Note: $1 added or subtracted for each right or wrong decision

Example: “Where were you born?”

True or False?

Example: “Where were you born?”

False!

Annotation

- Transcribed using Amazon Mechanical Turk
  - 5 Turkers per utterance, combined using Rover techniques
- Automatically segmented using Praat into Inter-Pausal Units (IPUs) at 50ms silence
- Automatically aligned with speech and truth/lie labels using aligner built with Kaldi

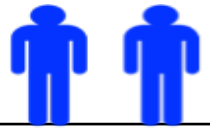
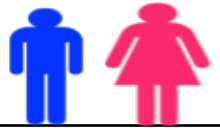
Rover Example

1	its	really	fun	um	I	go	like	to	a	place
2	its	really	fun		i	go	like		a	place
3	it's	really	fun	um	I	go	like	to	a	place
Rover output	its	really	fun	um	I	go	like	to	a	place
score	1	1	1	2/3	2/3	1	1	2/3	1	1

ROVER Score = (1+1+1+2/3+2/3+1+1+2/3+1+1)/10=0.9

Balanced Corpus

- 140 subject pairs
- ~112 hours of speech
- Pair types:

	English	Chinese	English/ Chinese
	14	14	14
	14	14	14
	14	14	28

Statistical Results: Deception Detection

- People's ability to detect deception correlates with their ability to deceive $r(252) = 0.13, p = 0.04$
- Holds across all subjects but
  - Strongest for **females** $r(126) = 0.26, p = 0.003$
  - No difference between English and Chinese females
- Subjects who are better at detecting deception are more likely to predict their partners have lied and vice versa

Gender, Ethnicity, Personality & Ability to Deceive

- No effect of gender or ethnicity across subjects but
  - ***Extraversion*** is significantly ***negatively*** correlated
    - English/Male $r(68) = -0.25$, $p = 0.04$
- Tendencies:
  - Chinese/female extraversion ***positively*** correlated with ability to deceive
  - American/female conscientiousness ***negatively*** correlated with ability to deceive

Gender, Ethnicity, Personality & Deception Detection

- *No effect* of personality factors
  - Contra earlier findings for English speakers (Enos et al ’06)

Confidence in Judgments

- Ability to detect deception ***negatively*** correlates with confidence in judgments for all subjects $r(250) = -0.14$, $p = 0.03$
  - Strongest for females $r(126) = -0.24$, $p = 0.01$
- Ability to deceive negatively correlated with confidence for males $r(124) = -0.185$, $p = .04$
  - Strongest for Chinese males $r(58) = -.35$, $p = 0.007$
- Less confident interviewers may ask more follow-up questions and obtain more evidence for decisions?

- Neuroticism ***negatively*** correlates with confidence for Chinese female subjects r(68) = -0.27, p = 0.02
- Openness to experience ***negatively*** correlates with confidence for all subjects r(249) = -0.14, p = 0.03
  - Strongest for females r(126) = -.021, p = 0.02
  - Strongest for Chinese females r(68) = -0.29, p = 0.02
- Some effect of gender, ethnicity and personality factors on confidence but …

Larger Corpus

- 139 subject pairs
- 100.5 hours of speech
- Largest cleanly recorded corpus of within-subject deceptive/non-deceptive speech with known ground truth

Classification Results

- Features:
  - Acoustic features: f0, intensity, voice quality, speaking rate – raw and normalized 2 ways
  - Gender: subject and partner
  - Ethnicity: subject and partner
  - Personality scores
  - Lexical features not yet available
- Weka experiments
  - J48 decision trees
  - Random Forests
  - Bagging

Classification Results

Model	Raw	SessionNorm	BaselineNorm
J48	59.89	62.09	62.19
Bagging	58.65	61.19	61.01
RF	61.23	**63.03**	62.79

- Baseline accuracy: 59.9%

Added features

- Speaker gender
- Speaker native language
- NEO-FFI personality scores

Classification Results (SessionNorm)

Model	Acoustic/ prosodic	Acoustic/prosodic +gender,lang,NEO
J48	62.09	64.86
Bagging	61.19	63.9
RF	**63.03**	**65.86**

- Baseline accuracy: 59.9%

Accuracy Predictions (Baseline 59.9%)

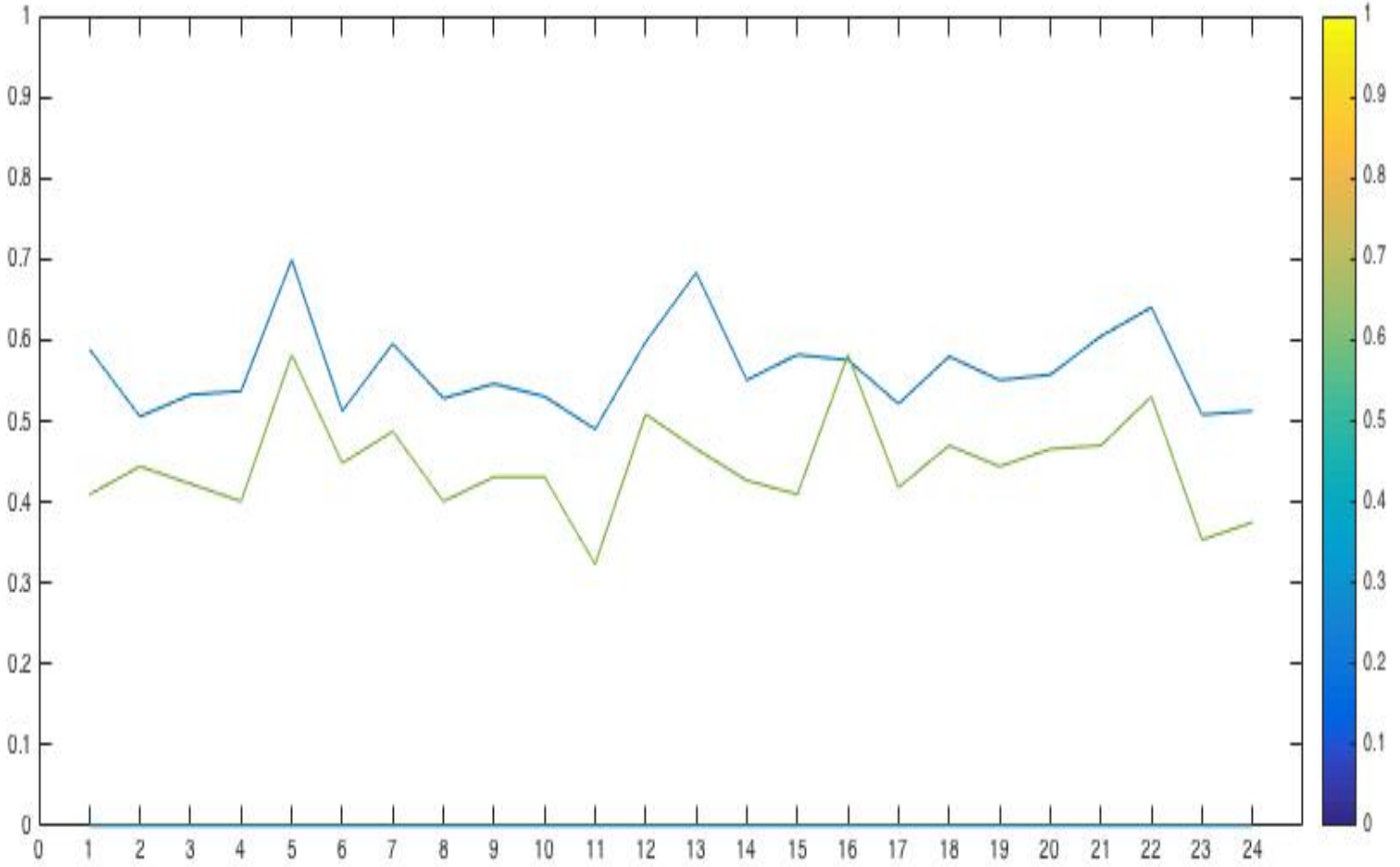
3 ML Models, Raw vs. Norm'd Acoustic Features

Model	Raw	Session Norm	Baseline Norm
J48	59.89	62.09	62.19
Bagging	58.65	61.19	61.01
RandomForest	61.23	***63.03***	62.79

All Features, (Session Norm'd Acoustic)

Model	Precision
J48	64.86
Bagging	63.9
RandomForest	***65.86***

Deception Detection by Question



Related and Future Work

- Laughter and deception studies
- More classification experiments
  - Additional features: Lexical, subject-dependent features
  - Examining entrainment as a factor:
    - Do subjects who entrain make better deceivers or deception detectors?
  - Deception detection and trust
  - Clustering subjects by gender, ethnicity, and personality features to build different models for each cluster

Publications

- 2015. S. I. Levitan, M. Levine, J. Hirschberg, N. Cestero, G. Ahn, A. Rosenberg, *"Individual Differences in Deception and Deception Detection,"* Cognitive 2015, Nice. (**Best Paper Award**)
- 2015. S. I. Levitan, G. An, M. Wang, G. Mendels, J. Hirschberg, M. Levine, A. Rosenberg, *"Cross-Cultural Production and Detection of Deception from Speech,"* ACM Workshop on Multimodal Deception Detection, ICMI 2015, Seattle.

Thank you!