# Detecting Deceptive Speech: Requirements, Resources and Evaluation

***Julia Hirschberg***
***Columbia University***

**LREC 2008**
**29 May 2008**

# Collaborators

- Stefan Benus, Jason Brenner, Robin Cautin, Frank Enos, Sarah Friedman, Sarah Gilman, Cynthia Girand, Martin Graciarena, Andreas Kathol, Laura Michaelis,  Bryan Pellom, Liz Shriberg, Andreas Stolcke
- Columbia University, SRI/ICSI, University of Colorado

- Ordinary people tell an average of 2 lies per day
  - *Your hair looks great.*
  - *I'd love to go but my parents are in town.*
  - *I'm sorry I missed your talk but my alarm clock didn't go off.*
- Even trained professionals are very poor at detecting deception
- In many cultures 'white lies' are ***more*** acceptable than the truth
  - Likelihood of being caught is low
  - Rewards also low but outweigh consequences of being caught
- But what about more 'serious' lies? Are they easier to detect?

# What is Deception?

- Deliberate choice to mislead
  - *Without prior notification*
  - To gain some ***advantage*** or to avoid some ***penalty***
- ***Deception is Not***:
  - Self-deception, delusion, pathological behavior
  - Theater
  - Falsehoods due to ignorance/error

# Who Studies Deception?

- Students of human behavior – especially psychologists
- Law enforcement personnel
- Corporate security officers
- Social services workers
- Mental health professionals

# Is it Easy to Deceive?

- ***No…***
  - Deceivers' cognitive load is increased because…
    - They must keep story straight
    - Remember what they've said ***and*** what they haven't said
  - Deceivers' fear of detection is increased if…
    - Target believed to be hard to fool
    - Target believed to be suspicious
    - Stakes are high: serious rewards and/or punishments
  - Hard to control indicators of deception

# Where do We Look for Signs of Deception?

- 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…
- Variation in ***what*** is said and ***how*** (Adams ‘96, Pennebaker et al ‘01, Streeter et al ‘77)
  - Contractions, lack of pronominalization, disfluencies, slower response, mumbled words, increased or decreased pitch range, less coherent,…

# Potential Spoken Cues to Deception (DePaulo et al. ’03)

- Liars less forthcoming?
  - \- Talking time
  - \- Details
  - \+ Presses lips
- Liars less compelling?
  - \- Plausibility
  - \- Logical Structure
  - \- Discrepant, ambivalent
  - \- Verbal, vocal involvement
  - \- Illustrators
  - \- Verbal, vocal immediacy
  - \+ Verbal, vocal uncertainty
  - \+ Chin raise
  - \+ Word, phrase repetitions
- **Liars less positive, pleasant?**
  - \- Cooperative
  - \+ Negative, complaining
  - \- Facial pleasantness
- **Liars more tense?**
  - \+ Nervous, tense overall
  - \+ Vocal tension
  - \+ F0
  - \+ Pupil dilation
  - \+ Fidgeting
- **Fewer ordinary imperfections?**
  - \- Spontaneous corrections
  - \- Admitted lack of memory
  - \+ Peripheral details

# Current Approaches to Deception Detection

- Training Humans
  - John Reid & Associates
    - Behavioral Analysis: Interview and Interrogation
- `Automatic' methods
  - Polygraph
  - Voice Stress Analysis
    - Microtremors 8-12Hz
  - Nemesysco and the Love Detector
  - *No objective evidence that any of these work*

# Exploring Corpus-Based Methods for Deception Detection

- **Goal**: Identify a set of acoustic, prosodic, and lexical features that distinguish between deceptive and non-deceptive speech
  - As well or better than human judges
  - Using automatic feature-extraction
  - Using Machine Learning techniques to identify best-performing features and create automatic predictors

# Major Obstacles

- Corpus-based approaches require large amounts of training data – difficult to obtain for deception
  - Differences between real world and laboratory lies
    - Motivation and potential consequences
    - Recording conditions
    - Identifying ground truth
- Ethical issues
  - Privacy
  - Subject rights and Institutional Review Boards

# Our Approach

- Record a new corpus of deceptive/non-deceptive speech and transcribe it
- Use automatic speech recognition (ASR) technology to perform forced alignment on transcripts
- Extract acoustic, prosodic, and lexical features based on previous literature and our work in emotional speech and speaker id
- Use statistical Machine Learning techniques to train models to distinguish deceptive from non-deceptive speech
  - Rule induction (Ripper), CART trees, SVMs

# Columbia/SRI/Colorado Deception Corpus (CSC)

- Deceptive and non-deceptive speech
  - Within subject (32 adult native speakers)
  - 25-50m interviews
- Design:
  - Subjects told goal was to find *“people similar to the ‘25 top entrepreneurs of America’”*
  - Given tests in 6 categories (e.g. knowledge of food and wine, survival skills, NYC geography, civics, music), e.g.
    - *“What should you do if you are bitten by a poisonous snake out in the wilderness?”*
    - *“Sing Casta Diva.”*
    - *“What are the 3 branches of government?”*

  - Questions manipulated so scores always differed from a (fake) entrepreneur target in 4/6 categories
  - Subjects then told real goal was to compare those who actually possess knowledge and ability vs. those who can “talk a good game”
  - Subjects given another chance at $100 lottery if they could convince an interviewer they match target completely
- Recorded interviews
  - Interviewer asks about overall performance on each test with follow-up questions (e.g. *“How did you do on the survival skills test?”*)
  - Subjects also indicate whether each statement T or F by pressing pedals hidden from interviewer

# The Data

- 15.2 hrs. of interviews; 7 hrs subject speech
- Lexically transcribed & automatically aligned
- Truth conditions aligned with transcripts: Global / Local
- Segmentations (Local Truth/Local Lie):
  - Words (31,200/47,188)
  - Slash units (5709/3782)
  - Prosodic phrases (11,612/7108)
  - Turns (2230/1573)
- 250+ features
  - Acoustic/prosodic features extracted from ASR transcripts
  - Lexical and subject-dependent features extracted from orthographic transcripts

# Limitations

- Samples (segments) not independent
- Pedal may introduce additional cognitive load
  - Equally for truth and lie
  - Only one subject reported any difficulty
- Stakes not the highest
  - No fear of punishment
  - Self-presentation and financial reward

# Acoustic/Prosodic Features

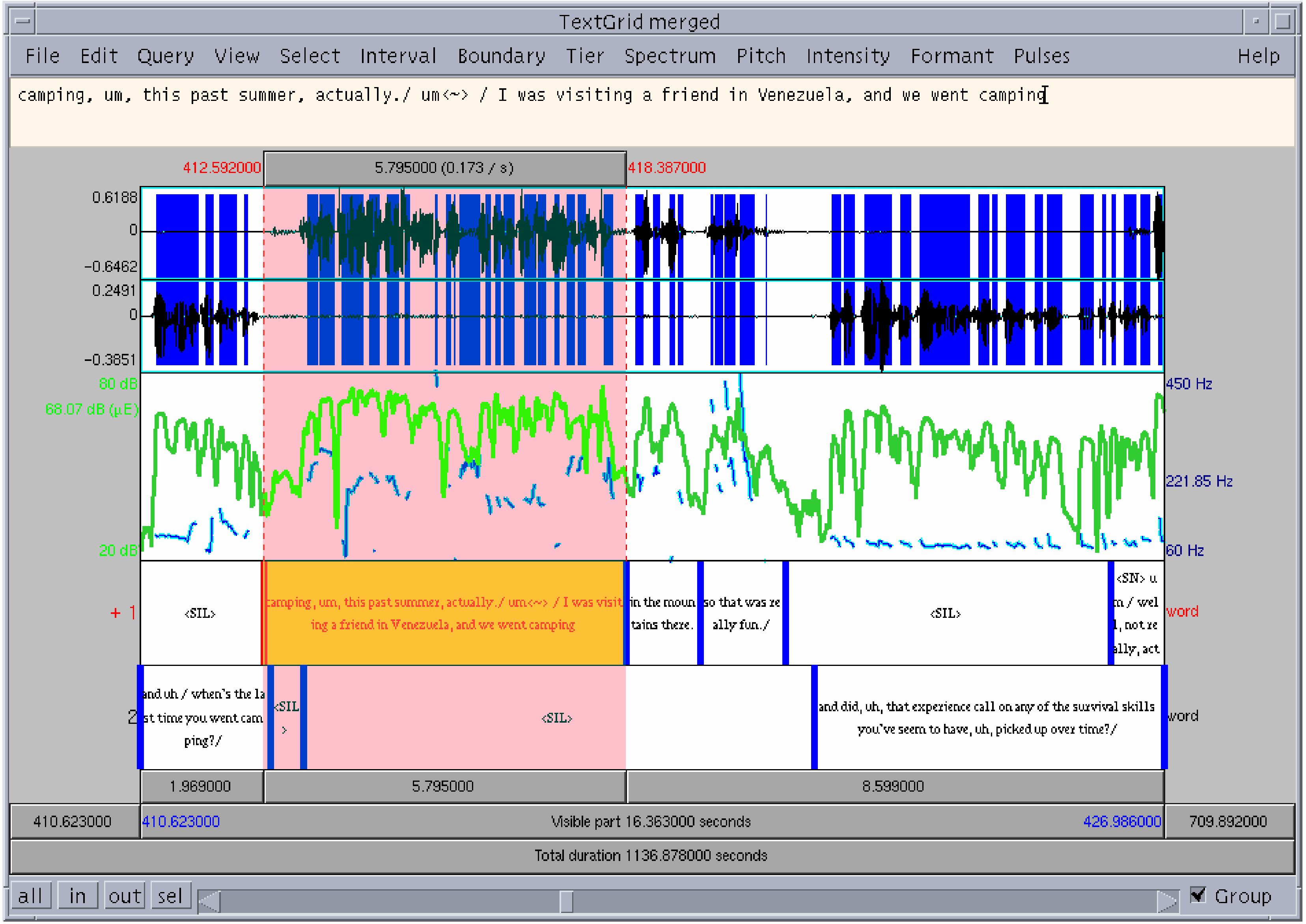
- Duration features
  - Phone / Vowel / Syllable Durations
  - Normalized by Phone/Vowel Means, Speaker
- Speaking rate features (vowels/time)
- Pause features (cf Benus et al ‘06)
  - Speech to pause ratio, number of long pauses
  - Maximum pause length
- Energy features (RMS energy)
- Pitch features
  - 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: Calibrating Truthful Behavior

- % units with cue phrases
- % units with filled pauses
- % units with laughter
- Ratio lies with filled pauses/truths with filled pauses
- Ratio lies with cue phrases/truths with filled pauses
- Ratio lies with laughter / truths with laughter
- 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
<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
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

# CSC Corpus: Objective Evalution

- Classification via Ripper rule induction, randomized 5-fold xval)
  - Slash Units / Local Lies — Baseline 60.2%
    - Lexical & acoustic: 62.8 %; + subject dependent: 66.4%
  - Intonational Phrases / Local Lies — Baseline 59.9%
    - Lexical & acoustic 61.1%; + subject dependent: 67.1%
- Other correlations
  - Positive emotion words → deception (LIWC)
  - Pleasantness → deception (DAL)
  - Filled pauses → truth
  - Some pitch correlations — varies with subject

# Evaluation: Human Deception Detection

- Most people very poor at detecting deception
  - ~50% accuracy (Ekman & O’Sullivan ‘91, Aamodt ‘06)
  - People use unreliable cues, ***even with training***

# 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** |

# Evaluating Automatic Methods by Comparing to Human Performance

- Deception detection on the CSC Corpus
- 32 Judges
  - Each judge rated 2 interviews
  - Received ‘training’ on one subject.
- Pre- and post-test questionnaires
- Personality Inventory

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.

# What Makes Some People Better?

- 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** (reversed as Emotional Stability). Characterized by traits like tense, moody, and anxious.
  - **Openness to Experience** (aka Intellect or Intellect/Imagination). Includes having wide interests, and being imaginative and insightful.

# 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<br>Agreeableness | Accuracy | 0.51<br>0.41 | 0.003<br>0.021 |
| Neuroticism<br>Agreeableness | F-measure for TRUTH | 0.37<br>0.41 | 0.035<br>0.019 |
| Openness | F-measure for LIE | 0.52 | 0.003 |

# Other Useful Findings

- ***No*** effect for training
- Judges' post-test confidence did ***not*** correlate with pre-test confidence
- Judges who claimed experience had significantly higher pre-test confidence
  - But ***not*** higher accuracy
- Many subjects reported using disfluencies as cues to deception
  - But in this corpus, disfluencies correlate with ***truth*** (Benus et al. '06)

# Future of Deception Research

- Need corpora that
  - Are collected in ‘real’ conditions
  - Provide multimodal data for corpus analysis
    - Speech and language
    - Biometric features
    - Visual information
  - Are reliably labeled for ground truth
  - Support research on individual differences in deception behavior
    - Personality data…
  - Support the study of cultural differences in deception

# THANK YOU!