Engineering for Speech Science:
From data and models to applications

Shrikanth (Shri) Narayanan
Signal Analysis and Interpretation Laboratory (SAIL)
http://sail.usc.edu

University of Southern California
Work reported represents efforts of numerous colleagues and collaborators. Too many to name, but grateful to all.

SUPPORTED BY:
NSF, NIH, ONR, ARMY, DARPA, IARPA
IBM, SIMONS FOUNDATION, GOOGLE

http://sail.usc.edu/
Talk objectives

• Review the role of engineering techniques/tools
  • for speech production data acquisition, analysis, applications
  • with a focus on “data science” approaches

• Describe specific imaging, signal processing, and machine learning possibilities using two case studies
  • Dynamics of vocal tract shaping during speech
  • Vocal expressions in children with Autism Spectrum Disorder

• Highlight opportunities for basic and applied speech communication research
Engineering for speech science

A long history and ubiquitous use of engineering tools for

• **Acquiring the right data**
  • from recording speech audio in a variety of environments to using advanced instrumental techniques for observing speech production

• **Analyzing and modeling data**
  • from spectral analysis and linear prediction to novel machine learning methods for analysis, data visualization and theory building

• **Applying and using data**
  • from facilitating clinical screening and diagnostics to supporting interventions

*data and data sciences providing the glue*
Acquiring The Right Data: Mics & Cameras To Advanced Instrumentation

X-ray (Stevens, 1962)
http://psyc.queensu.ca/~munhallk/05_database.htm

Ultrasound (Stone, 1980)
http://www.speech.umaryland.edu

Electropalatography
(courtesy: UCLA Phonetics Lab)

Electromagnetometry

upper lip
lower lip
tongue
velum
teeth

Zemlin-Narayanan-October-2017 - November 9, 2017
Acquiring the data

In laboratory and clinics to naturalistic “in the wild” settings

• Instrumented environments: arrays of sensors—microphones, cameras, depth sensors, mocap,..
  • Sense people: identity/location; speech, visual activity, interaction,..
  • and the environment: lab, classroom, clinic, home, playground,..

• Instrumented people: body sensing, ambulatory settings
  • Sense the user state, activity, environmental context
  • Wireless smart phone based digital activity, location, EMA,..
Analyzing and modeling data

A variety of audio, speech, image processing tools available for analyzing/understanding verbal and nonverbal behavior

- **Classic techniques**: speech activity detection, pitch estimation, spectral analysis, linear prediction modeling, voice quality measures,..

- **Novel advances**: computational methods for characterizing cognitive and emotional states, detecting atypicality/pathology, interaction mechanisms,..

---

Dialogue
Recording
Digital format
Capture Audio
Audio Waveform
Spectrogram
Information extraction
A whole range of speech/language technology possibilities

- Voice Activity Detection
- Audio Segmentation
- Alignment
- Transcription
- Keyword Spotting
- Prosody Modeling: Intonation, Phrasing, Prominence
- Voice Quality
- Natural Language Processing of Text/Transcripts
- Dialog Act Tagging
- Interaction Modeling: Turn Taking Dynamics, Entrainment
- Speaker/Verification Identification
- Affective Computing from Speech and Language
- Speaker State and Trait Characterization
- Joint Speech and Visual Cue Processing

With Varying Degrees of Technology Maturity
Who spoke when, for how long, and about what?

Detected Speech Regions

Speakers in the conversation

Transcription

Voice Activity Detection

Speech Spectrogram

Hello!

How are you?

I am good.

We should go.

Come on!

Ok!
What more can we infer beyond words?

<table>
<thead>
<tr>
<th>Words:</th>
<th>So ‘n’ your chest pains have been going on just for two days is that right</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speaker:</td>
<td>spkr1 (Doctor)</td>
</tr>
<tr>
<td>Gender:</td>
<td>Male</td>
</tr>
<tr>
<td>Age:</td>
<td>Adult</td>
</tr>
<tr>
<td>Prominent words:</td>
<td>So ‘n’ your chest pains have been going on just for <strong>two days</strong> is that <strong>right</strong></td>
</tr>
<tr>
<td>Prosodic phrasing:</td>
<td>[So ‘n’ your chest pains] [have been going on just for two days] [is that right]</td>
</tr>
<tr>
<td>Speech act:</td>
<td>Yes-No Question</td>
</tr>
<tr>
<td>Affect:</td>
<td>Neutral</td>
</tr>
<tr>
<td>Attitude:</td>
<td>Polite</td>
</tr>
</tbody>
</table>

Rich information beyond words
Applying and using data

For supporting scientific research and discovery
• tools for illuminating mechanisms e.g., individual differences in articulatory strategies from production data

For supporting technology development
• biologically-informed speech synthesis, biometrics

For supporting clinical translation
• objective tools for screening, diagnostics, intervention support including delivery and behavior change modeling
TALK OUTLINE: two case studies

Dynamics of vocal tract shaping during speech and beyond

• Acquisition
  • real-time MRI for vocal tract imaging

• Analysis and modeling
  • audio & image processing, signal & systems modeling

• Applications
  • scientific: illuminating individual variability
  • clinical: speech in post-glossectomy, apraxia, aglossia
Behavioral phenotyping in Autism Spectrum Disorder

**TALK OUTLINE: two case studies**

- **Acquisition**
  - ecologically-valid behavioral recordings: clinics, homes,..

- **Analysis and modeling**
  - behavioral signal processing, machine learning

- **Applications**
  - objective understanding of atypical prosody
  - supporting diagnostics, behavior change modeling
Vocal tract imaging
Speech Production Studies: Data Is Integral

- Observe, measure, visualize speech articulation
- Rich and long history of speech instrumentation & imaging
- Number of techniques, each with strengths and limitations
  - Spatial and temporal resolution
  - User safety
  - Flexibility, ease of use, portability
  - Data interpretability
  - Specific research and application needs
Newer Possibilities:
MRI for structural vocal tract imaging

**Capable of 3D imaging of the hydrogen concentration in human body**

**Number of advantages:**
- Non-invasive, no ionizing radiation
- Arbitrary scan plane: Information on complete vocal tract geometry
- Excellent, flexible structural differentiation: Good soft tissue contrast, SNR
- Amenable to computerized 3D modeling: reconstruction and visualization
- Quantitative information: area function and acoustic relations
- Variability analyses

**Limitations/Challenges**
- Slow: Spatial & Temporal resolution tradeoffs, optimizing to a given application
- Noisy images: Susceptibility, blurring artifacts
- Imaging teeth
- Interaction with other physiological activities: respiration, swallowing, other movement
- Clean, Synchronized audio (and other modalities, as needed)
- Ease of experimentation, including cost and portability
MRI: Static Vocal tract Information

Information on 3D vocal tract structure/shape, inter-speaker variations
Accurate vocal tract measurements: area functions, length
Detailed studies on vowels and a number of continuant speech sounds
Facilitated new acoustic modeling studies

- Vowels, Nasals, Fricatives, Liquids: English and other languages

Midsagittal vowel images from Haskins (from Goldstein)

3D airway reconstruction for vowel /a/, Univ. Iowa (Story)
http://everest.radiology.uiowa.edu/nlm/app/vocal/vocal.html

Improving MRI temporal resolution

- A non 2D-FFT acquisition strategy (spiral k-space trajectory) on a GE Signa 1.5T CV/i scanner with a low-flip angle spiral gradient echo, 9-10 images/second
- Adapted pulse sequence originally developed for cardiac imaging.
- Effective reconstruction rates of 24-35 frames/second
  - sliding window reconstruction technique

First to use real-time MRI and synchronous noise-cancelled audio to understand vocal tract movements during natural speech production.

Spatial vs. Time resolution: speech MRI

- Our new system (circa 2015) enables visualization of all speech tasks

Spatial vs. Time resolution: speech MRI

- Our system enables visualization of all speech tasks

Highly accelerated RT-MRI of speech is achieved by synergistic combination of

- Novel custom upper-airway coil design
- Fast spiral readouts
- Constrained reconstruction/TT-GRAPPA

Real-time MRI at 83 fps, 2.4 mm/pixel
Vowels and diphthongs in h_d
Real-time MRI at 83 fps, 2.4 mm/pixel
Voiceless fricatives
Real-time MRI at 83 fps, 2.4 mm/pixel
A child speaker
USC-TIMIT: A MULTIMODAL ARTICULATORY DATA CORPUS FOR

- 10 American English talkers (5M, 5F).
- Real time MRI (5 speakers also with EMA) and synchronized audio.
- 460 sentences each (>20 minutes)
- Freely available for speech research.

WEB-LINK (with download info):
http://sail.usc.edu/span/usc-timit/

SAIL homepage: http://sail.usc.edu

Analysis of data

• Image analysis
• Deriving
  • morphological (structural) details, and
  • linguistically meaningful articulatory features
Quantitative imaging biomarkers

Objective, quantitative characteristic derived from MRI as an indicator of normal biological processes, pathogenic processes, or a response to a therapy

Sullivan et al. Radiology, 2015

Most biomarkers for dynamic MRI scans of the head and neck are used in only one study
Measurement of Structural Details

a) Mid-sagittal slice

b) Axial slice
First define a contour model segmentation manually: each articulator in a different color.

Now hierarchically optimize the model fit to the image in the Fourier domain using gradient descent!

Articulatory Posture & Constriction Task Variables

These feature sets are useful for modeling speech production dynamics


Tracking Constriction Variables


real-time magnetic resonance imaging


tissue-air boundary segmentation


constriction degree measurement

Analysis and Modeling

- Characterizing vocal tract morphology
- Understanding speaker specific articulatory strategy
- Inferring vocal tract structure/strategy from speech signal
- Enriching Speaker Verification with production information
**Overarching Questions**

- How are individual vocal-tract **structural differences** reflected in the speech acoustics?
- Can structural differences be **predicted** from acoustics?
- How do individuals adopt to structural differences to achieve **phonetic equivalence**?
- **What** contributes to distinguishing speakers from one another from the speech signal?

Not only try to differentiate individuals from their speech signal but understand **what** makes them different from a structure-function perspective.
Different individuals....

..each with a uniquely shaped vocal instrument
And with differing articulatory strategies during speech ...

Fifteen different individuals producing vowel /i/
Interspeaker Variability: Vocal Tract Morphology (Structure)

- Confined articulatory environment
- Variability across speakers
  - Highly articulated, layered system
- Reflected in acoustical properties
Extremes of Hard Palate Variation

Min:
- flat

Max:
- domed

Mode 1: Concavity

Mode 2: Anteriority

Mode 3: Sharpness
Why is morphological structure relevant?

Palate Shape - Principal Components

46% of Variance
Component 1: concavity

30% of Variance
Component 2: anteriority

10% of Variance
Component 3: sharpness

46% of Variance
30% of Variance
10% of Variance

**Concavity**: impact on F1 and F2; **Anteriority**: impact on F2 only; **Sharpness**: marginal

THEORETICAL IMPACT: ACOUSTIC MODELING

Formant Sensitivity to Palatal Concavity

Experiments on estimating some of these vocal tract shape details from acoustics...
Inversion: Palatal Concavity

Motor Controller

Binary Classification concave or flat palate?

Speech Articulation

Vocal Tract Shape

Speaker ID Features
- MFCC
- Open-smile
- GMM UWPP
- Inv. artic. features

Inversion Accuracy: 63% - 71%

Inversion: Vocal Tract Length

Motor Controller

Speech Articulation

Vocal Tract Structure

Vocal Tract Shape

Speech Signal

compensation limited

… speaker-specific acoustics

invert vocal tract length?
Variation in Vocal Tract Length, Acoustics

Vorperian (2009)

Age in Years

Vocal Tract Length (cm)

Frequency (Hz)

Vocal Tract Length (cm)

F1

F2

F3
Vowel Variation & Vocal Tract Length

![Graph showing F1 and F2 for different vowels](image)

- /i/:
  - F1: 2200 Hz
  - F2: 13.2 cm vt

- /u/:
  - F1: 1700 Hz
  - F2: 16.4 cm vt

- /a/:
  - F1: 1200 Hz
  - F2:
VT Length Estimation: Design

\[ \hat{L} = \frac{c}{4\hat{\Phi}} \]

\[ \hat{\Phi} = \frac{\beta_1 F_1}{1} + \frac{\beta_2 F_2}{3} + \frac{\beta_3 F_3}{5} + \ldots + \frac{\beta_m F_m}{2m - 1} \]

**Wakita (1977)**

- \( \beta_1 = 0 \)
- \( \beta_2 = 0 \)
- \( \beta_3 = 0.5 \)
- \( \beta_4 = 0.5 \)

**Fitch (1997)**

- \( \beta_1 = -0.167 \)
- \( \beta_2 = 0 \)
- \( \beta_3 = 0 \)
- \( \beta_4 = 1.167 \)

**Proposed**

Determine \( \beta \) via linear regression
Estimating VT Length From Acoustics

- 5 SPAN-TIMIT Subjects: real time MRI data
- Median estimated value (30 sec. read speech)

Could be used to account for speaker differences: new strategies for Vocal Tract Length Normalization

![MRI images](image)

\[ r = 0.98 \]
## VT Length Estimation: Results

<table>
<thead>
<tr>
<th>Simulated Data</th>
<th>Human Speech Data</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Accuracy</strong></td>
<td><strong>Accuracy</strong></td>
</tr>
<tr>
<td>• 0.631 RMSE (cm)</td>
<td>• 1.159 RMSE (cm)</td>
</tr>
<tr>
<td>• 3.8% of total length</td>
<td>• 7.6% of total length</td>
</tr>
<tr>
<td><strong>Improvement</strong></td>
<td><strong>Improvement</strong></td>
</tr>
<tr>
<td>• 8% over previous best</td>
<td>• 43% over previous best</td>
</tr>
</tbody>
</table>

* further improvements possible by refining model

Articulatory strategies
how talkers move their vocal tracts

• Vocal tract is a redundant system
• Articulators have overlapping functions
  • e.g., both jaw and lips contribute to bilabial constrictions
• Speakers have several ways to change airway shape to make a constriction
• We call these articulatory strategies

Articulatory Flexibility

• articulators have overlapping functions
• a constriction can be achieved in several different ways

examples of tradeoffs
• upper lip and lower lip
• jaw and lips
• jaw and tongue

Quantifying Individual Articulatory Strategy

Predicting how different speakers navigate this flexible system:

Estimate the forward kinematic map, which relates vocal tract shape to constriction degree
- how a speaker uses their forward kinematic map reflects their articulatory strategy

Analysis-by-synthesis approach
- estimate forward kinematic map using real time MRI data
- run dynamical system simulations
- simulations display predicted articulatory strategies

modeling goal
quantify how much the jaw, tongue, and lips contribute to a constriction of the vocal tract
Approach

- rtMRI videos
- anatomically guided factor analysis
- tissue-air boundary segmentation
- forward map estimation
- dynamical systems simulations
- articulatory strategy summary
- constriction degree measurement
Locally linear forward map

from factors of vocal tract shape (Toutios & Narayanan, 2015)

to constriction degrees (Ramanarayanan et al., 2013)

Contours first automatically extracted (Bresch & Narayanan 2009)

Articulator positions to constriction degrees using a statistical estimation technique (Lammert et al. 2013)
Articulatory strategies across speakers
insights using data from 18 speakers in task dynamic simulation (analysis by synthesis)

A model-based approach quantifies articulatory strategies by

- approximating the speaker specific forward map with rt-MRI data
- simulating vocal tract constrictions with a dynamical system, and
- interpreting the results.

Data: rtMRI of read passages from 18 speakers (9M, 9 F)

Articulatory strategies predictions

Each circle (at each place of articulation) represents a speaker
Alveolar closure

small jaw movement, speaker M3

prediction

observation
Alveolar closure

large jaw movement, speaker F9

prediction

observation
Summary

• real time MRI of the vocal tract can be used to estimate forward kinematic map
• forward kinematic map differs by speaker according to vocal tract geometry
• articulatory strategies predicted on the basis of vocal tract geometry can be compared against observed strategies

→ tool for relating vocal tract structure and function

Individual variability: Insights from Clinical Applications
Characterizing Articulation in Apraxic Speech

Components may be easily missed by traditional methods relying on acoustic data and/or phonetic transcription

- **rtMRI reveals intrusion errors in apraxic speech** – coproduction of target speech movement and erroneous speech movement
  - in word pair repetition task (“top-cop, top-cop…”) and non-repetitive speech (e.g., “statistical”, “telephone”)
- **errors sometimes acoustically imperceptible**

![Images of speech articulation](image_url)

- **rtMRI reveals frequent multiple initiation attempts** (“groping”) on motorically complex segments (e.g., /w/, /dʒ/, /s/)

rtMRI reveals covert (unphonated) articulation of entire words

“I can TYPE KNOW know…”

Clinical Takeaway: Apraxia of speech affects ability to select appropriate vocal tract movements for a target word/phrase and coordinate them in time, suppressing other movements. Errors may not always be auditorily perceptible.
Head and Neck Cancer

Head and neck cancer impairs speech and swallowing
- Cancer-associated cachexia, Peripheral nerve damage
- Radiation-induced fibrosis
- Surgical treatment effects
Tongue size effects: Glossectomy Patients

Glossectomy: surgical removal all/part of tongue (such as for treating oral cancer)

Tongue Tip

Tongue Base/Tip

G3

P1

16% 39% Normalized Tongue Size
Characterizing Post-Glossectomy Speech

What aspects of consonant and vowel production are maintained or lost following a partial glossectomy?

How do glossectomy patients compensate for reduced lingual mass and mobility, post-operatively?

Consonants
- Constriction degree (amount of occlusion) is maintained
- Constriction location is not maintained
  - Patients use articulators other than those conventionally used
    - e.g., target coronal constriction —> labiodental constriction

Vowels
- Constriction degree is not maintained
  - Exceedingly wide constrictions result in poor vowel intelligibility
- Constriction location is maintained

Clinical Takeaway: Patients would likely benefit from pre- and post-operative speech therapy focused on using jaw height modulation and lip protrusion to compensate for reduced lingual mobility during vowel production
A speaker with congenital aglossia

Production of /ata/

Non-aglossic speaker
- partial closure of lips
- full alveolar closure of tongue tip

Aglossic speaker
- no alveolar closure of tongue tip
- full closure of lips

- Bilabial closure instead of a normal lingua-alveolar closure for coronals
- However a cavity anterior to the constriction is formed, filtering transient & frication sources in a manner similar to typical alveolar production

Behavioral phenotyping in Autism Spectrum Disorder
Novel Technologies for Rich Understanding of Expressive Behavior and Interaction

Example: Parent and child creating a story together

Computational Targets
Joint Attention
Turn-taking
Shared enjoyment
Behavioral Synchrony
BEHAVIOR UNDERSTANDING CENTRAL TO MANY ENDEAVORS
..BOTH IN BASIC RESEARCH AND ACROSS APPLICATION DOMAINS

BEHAVIORAL SIGNAL PROCESSING & INFORMATICS:
COMPUTING BEHAVIORAL TRAITS & STATES FOR DECISION MAKING AND ACTION

✓ HELP US DO THINGS WE KNOW TO DO MORE EFFICIENTLY, CONSISTENTLY

✓ HELP HANDLE NEW DATA, CREATE NEW MODELS TO OFFER UNIMAGINED INSIGHTS:
CREATE TOOLS FOR DISCOVERY

✓ HELP CREATE TOOLS TO SUPPORT PERSONALIZED INTERVENTION, AND ITS TRACKING

..using bio-behavioral cues

Automatic Behavior Coding: Estimate behavioral codes from data
How is technology helping already?

- Significant advances in foundational aspects of behavior modeling: detect, classify and track
  - Audio & Video diarization: who spoke when; doing what,..
  - Speech recognition: what was spoken
  - Visual activity recognition: head pose; face/hand gestures
  - Physiological signal processing with EKG, GSR, ..

SIGNAL PROCESSING AND MACHINE LEARNING ARE KEY ENABLERS
Autism Spectrum Disorder (ASD)

Technology possibilities

- Illuminating social communication and behavioral patterns
- Behavior phenotyping with objective and adaptable metrics for screening and diagnostics in a stratified manner
- Track, quantify behavior change e.g., response to interventions
- Technologies to support measurements and intervention delivery
  - personalized, just in time, ecologically valid
Analyzing Interaction in ASD

- Assessment, Intervention, Game play/training Examples
ASD Assessment

ASD

Language and Interaction
- Echolalia
- Gestures
- Conversation
- Social Response

Stereotyped Behaviors
- Unusual Preoccupation

Prosodic Abnormalities
- Intonation
- Volume
- Rate
- Voice Quality

Danny Bone, PhD
Quantifying Atypical Prosody

Qualitative descriptions are general and contrasting

ADOS
Module 3

“slow, rapid, jerky and irregular in rhythm, odd intonation or inappropriate pitch and stress, markedly flat and toneless, or consistently abnormal volume”

Structured assessment may not capture how atypical prosody affects social functioning apart from pragmatics
Quantifying Prosody: Acoustic features

- **24 Features:** pitch (6), volume (6), rate (4), and voice quality (8)
  - **Intonation:** F0 curvature, slope, center
  - **Volume:** Intensity curvature, slope, center
  - **Rate:** Boundary (turn end word), Non boundary
  - **Voice Quality:** Jitter, Shimmer, CPP, HNR

♀ median, IQR of above
Our Case study Setup

Approach

• **Automatic measures from spontaneous speech**
  • Create generally applicable tools for discovery

• **Data**
  • N=28 children.
  • **ADOS module 3 Interviews**
    • USC CARE Corpus

Hypotheses

1. Children with ASD will demonstrate correlation between acoustic-prosodic cues and severity of ASD-related impairment

2. Psychologist’s speech is also informative of rated severity (both participant and evaluator)
Atypical Prosody & Interaction

Spearman’s Correlation between rated severity and prosodic cues dataset of ADOS 3 administration (N=28)

Child’s Prosody  ➔  Psychologist’s’s Prosody

- “Monotone”  
  \[ p < 0.01 \]
- “Abnormal volume”  
  \[ p < 0.05 \]
- “Breathy/Rough”  
  \[ p < 0.01 \]
- Slower speaking rate  
  \[ p < 0.05 \]
- Questions/affect  
  \[ p < 0.05 \]
- Variable prosody  
  \[ p < 0.01 \]
- also higher jitter  
  \[ p < 0.01 \]
- slower/then faster  
  \[ p < 0.01 \]

The psychologists may be varying their engagement strategies

ASD Severity Regression

<table>
<thead>
<tr>
<th>Descriptor’s Included</th>
<th>Child Prosody</th>
<th>Psych Prosody</th>
<th>Child and Psych Prosody</th>
<th>Underlying Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spearman’s ρ</td>
<td>0.50**</td>
<td>0.71****</td>
<td>0.50**</td>
<td>-0.14</td>
</tr>
</tbody>
</table>

Spearman’s ρ between prediction and labels. [**,****]≡a=[0.01,1e-4]. N=28.

- Multiple linear regression forward-feature selection on the 20 prosodic features, leave-one-session-out
- Psychologist’s acoustics more predictive of child’s ratings
- Using total feature set shows no advantage.

Modeling Interaction Dynamics Critical

- More data can offer further insights into prosody, and beyond, in speech communication

Language and Turn-taking Features

Global Turn-taking Measures (4 features)
- Can indicate style of interaction
  - *speech %*, *silence %*, *overlap %* (*interruption %*), and *median latency* (time between turn exchanges)

Rate (3 features)
- Also useful for characterizing interaction
  - *speaking rate* (*SR*, #-words/utt. dur.; includes pausing)
  - *per-word articulation rate* (*AR*, syl/word dur.)
  - *intra-utterance pausing duration*

Language
- Linguistic Inquiry and Word Count (LIWC) toolbox
- Percentages normalized by the total number of words spoken
  *Examine the whole session, not only the interviews*

(1) words per sentence (WPS)—a rough approximation of mean-length-of-utterance (MLU); (2) first-person, singular pronouns (*I, me, mine*); (3-5) positive emotion, negative emotion, and affect (positive or negative) language; (6-8) assents (*OK, yes*), non-fluencies (*hm, umm*), and fillers (*I mean, you know*).
**Results: Language & Turn taking**

<table>
<thead>
<tr>
<th></th>
<th>Trend w/</th>
<th>Psych</th>
<th>Sp. ρ</th>
<th>Trend w/</th>
<th>Child Feature</th>
<th>Sp. ρ</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Speech Amount</strong></td>
<td>Increased</td>
<td>Speech %</td>
<td>0.54</td>
<td>Decreased</td>
<td>Speech %</td>
<td>-0.36</td>
</tr>
<tr>
<td></td>
<td></td>
<td>WPS (MLU)</td>
<td></td>
<td></td>
<td></td>
<td>-0.42</td>
</tr>
<tr>
<td><strong>Turn-taking</strong></td>
<td>Increased</td>
<td>Articulatory</td>
<td>0.38</td>
<td>Decreased</td>
<td>Articulatory</td>
<td>-0.34</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Intra-turn sil.</td>
<td>0.32</td>
<td></td>
<td>Latency</td>
<td>0.34</td>
</tr>
<tr>
<td><strong>Pronouns</strong></td>
<td>Increased</td>
<td>Personal</td>
<td>0.38</td>
<td>Decreased</td>
<td>Personal</td>
<td>-0.40</td>
</tr>
<tr>
<td><strong>Language Use</strong></td>
<td>Decreased</td>
<td>Assent Lang.</td>
<td>-0.48</td>
<td>Decreased</td>
<td>Affect Lang.</td>
<td>-0.40</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Non-fluencies</td>
<td>-0.48</td>
<td></td>
<td>Fillers</td>
<td>-0.41</td>
</tr>
</tbody>
</table>

- Psychologist partners influence the child’s behavior.
- Psychologist may wait for a prompted response, then proceed.
- Child is interacting less with the psychologist.
- The psychologist is reacting to child’s behavior.
- Child may be reluctant to discuss themselves, and may not follow up on prompts to engage psychologist.
  - Child avoid use of the word ‘I’ [Baltaxe, 1997]
- Psychologist back-channels less, Child uses less fillers.
Summary

Objective insights from computational processing

• Prosodic, turn-taking, and language features of the interacting psychologist and child indicate the conversational quality degrades for children with greater severity of ASD symptoms
• Psychologist language features may be robust to social demand
• Need for mathematical models of interaction in ASD

Future Work

• Investigate interplay between these varied features
• Larger datasets that include TD and non-ASD DD
• Unsupervised behavioral signals e.g., arousal dynamics, entrainment
More speech processing possibilities…
understanding behavioral mechanisms and processes
Robust Vocal Arousal Estimation: A simple tool

• Simple framework: compares favorably to cross-corpus supervised classification systems
• Tool generalizes: researchers can use to investigate behavioral hypotheses

3 features: pitch, intensity, and spectral slope (HF500)

• Largely unsupervised, only requires “neutral” labels from each speaker

select features model neutral robust scoring

\[
\begin{align*}
\rho_1,j &\quad \rho_2,j \\
\rho_3,j &\quad \beta \\
\rho_j &
\end{align*}
\]

Computing Vocal Entrainment: A novel measure
“HOW MUCH DO TWO PEOPLE SYNCHRONIZE IN A CONVERSATION?”

Preprocessing
Raw Acoustic Features
Speech-text Alignment

Vocal Features Extraction

PCA Vocal Characteristics Space
(1) PCA for Each Speaking Turn (Local)
(2) PCA for Each Speaker (Global)

Representative Acoustic Features Identification and Parameterization

Compute Similarity Metric
(1) Symmetric Similarity Metric (non-directional)
(2) Directional Similarity Metric

Vocal Entrainment Measures
Predict Latent Mental State
Affective Synchrony

• Synchrony of vocal arousal during child-psychologist interactions

DANIEL BONE, CHI-CHUN LEE, ALEXANDROS POTAMIANOS AND SHRIKANTH NARAYANAN. AN INVESTIGATION OF VOCAL AROUSAL DYNAMICS IN CHILD-PSYCHOLOGIST INTERACTIONS USING SYNCHRONY MEASURES AND A CONVERSATION-BASED MODEL. IN PROCEEDINGS OF INTERSPEECH, 2014

DANIEL BONE, CHI-CHUN LEE AND SHRIKANTH S. NARAYANAN. ROBUST UNSUPERVISED AROUSAL RATING: A RULE-BASED FRAMEWORK WITH KNOWLEDGE-INSPIRED VOCAL FEATURES. IEEE TRANSACTIONS ON AFFECTIVE COMPUTING. 2014.
Child-Parent Synchrony

Overt behavior signals of children with autism may be inconsistent with their inner affective state.

**Electrodermal response (EDA): arousal (activation) levels**

**Verbal Response Latency (VRL): reflect cognitive and affective state**

**APPROACH**
- Joint representation of child and parent physiological cues with coupled HMMs modeled on physiological features
- Predict cognitive load with synchrony features

**FINDINGS**
- Language and physiology give complementary information
- Parent’s cues provide additional information about child’s behavior
- Parents tend to synchronize with their children depending on the child’s ability to engage in task


"Sound to code" system: Talk therapy example

Estimating empathic behavior directly from audio

- **82%** accuracy for *fully* automatic system (no human intervention)
- **61%** (chance), **85%** (manual transcripts), **90%** (human agreement)

Bo Xiao, Zac Imel, Panayiotis Georgiou, David Atkins and Shrikanth Narayanan. "Rate my therapist": Automated detection of empathy in drug and alcohol counseling via speech and language processing. PLoS ONE, 10(12): e0143055. 2015
Open Challenges

- **Getting the right multimodal data**
  - sensing in natural context; capturing context
  - doing it in an ecologically valid way

- **Processing the data**
  - variability, heterogeneity and uncertainty in data
  - specifying behavior representations for computing
  - reflecting multiple (diverse) perspectives & subjectivity
  - interpretable, targetable “features” for interventions
  - dealing with various levels of “imperfect” solutions
  - learning/transfer across domains

- **Using the data, closing the loop with stakeholders**
  - Data provenance, integrity, sharing, and management
  - Enabling interventions & evaluation at scale, cost, JIT
  - Choosing the right operating point: adaptivity
Rich Opportunities

- Developing and sustaining trusting and productive partnerships between various stakeholders
  - recognize and acknowledge the problem is tough
  - development takes time, and wrought with failure points: be positive with big picture in mind
  - communicate often, intellectual openness
  - empower trainees
  - learn from experiences, share
  - there’s not one magic bullet: seek flexible solutions
- Work on real use cases. Share, disseminate information, (even if results not too exciting)
- Build community
Concluding Remarks

• **Data is integral to advancing speech communication research**
  • Requires concurrent progress in sensing, processing and modeling
  • Need to gather and integrate multiple, disparate sources of information toward getting a more complete picture of human speech production

• **The problem is highly challenging**
  • Technological, computational as well as conceptual/theoretical challenges
  • Potential for scientific, technological, and clinical applications

• **An ongoing interdisciplinary scientific endeavor**
ACKNOWLEDGEMENTS

ASHA Special Interest Group 19 and the Zemlin Lecture Committee

- NIH Grants DC007124 & DC03172
- NSF, ONR, DoJ
- USC Imaging Sciences Center
- LAC-USC Hospital
- USC Center for High Performance Computing

Papers, Videos, Teaching resources

http://sail.usc.edu/span
• ALL REFERENCES BELOW ARE AVAILABLE AT sail.usc.edu/publications.php


• Ming Li, Adam Lammert, Jangwon Kim, Prasanta Kumar Ghosh and Shrikanth S. Narayanan, Automatic Classification of Palatal and Pharyngeal Wall Shape Categories from Speech Acoustics and Inverted Articulatory Signals, in: ISCA Workshop on Speech Production in Automatic Speech Recognition (SPASR), Lyon, France, 2013


• Christina Hagedorn, Adam Lammert, Mary Bassily, Yihe Zu, Uttam Sinha, Louis Goldstein, Shrikanth S. Narayanan,
• Ming Li, Jangwon Kim, Adam Lammert, Prasanta Ghosh, Vikram Ramanarayanan and Shrikanth Narayanan. Speaker verification based on the fusion of speech acoustics and inverted articulatory signals. Computer, Speech, and Language. 36: 196-211, March 2016


• Sajan Goud Lingala, Yinghua Zhu, Yongwan Lim, Asterios Toutios, Yunhua Ji, Wei-Ching Lo, Nicole Seiberlich, Shrikanth Narayanan, Krishna S. Nayak, "Feasibility of through-time spiral generalized autocalibrating partial parallel acquisition for low latency accelerated real-time MRI of speech", Magnetic Resonance in Medicine, 2017. Early view.


DATABASES/WEBSITES with MULTIMEDIA RESOURCES

- **USC TIMT CORPUS**
  
  
  [http://sail.usc.edu/span/usc-timit/](http://sail.usc.edu/span/usc-timit/)

- **USC Speech and Vocal Tract Morphology MRI Database**
  
  
  [http://sail.usc.edu/span/morphdb/index.html](http://sail.usc.edu/span/morphdb/index.html)

- **USC rtMRI IPA Chart illustration**
  
  
  [http://sail.usc.edu/span/rtmri_ipa/index.html](http://sail.usc.edu/span/rtmri_ipa/index.html)

- **USC EMO MRI CORPUS**
  

- **USC EMA CORPUS**
  
  Sungbok Lee, Serdar Yildirim, Abe Kazemzadeh and Shrikanth S. Narayanan, An articulatory study of emotional speech production, in Proceedings of Interspeech, pages 497-500, 2005
  
  [http://sail.usc.edu/ema_web/index.html](http://sail.usc.edu/ema_web/index.html)

- **MUSIC**
  
  

  [http://sail.usc.edu/span/videos/USC-Soprano-AveMaria.mov](http://sail.usc.edu/span/videos/USC-Soprano-AveMaria.mov)