USC University of Southern California

SAIL

Child-centered Multimodal Machine Intelligence

Shrikanth (Shri) Narayanan

University Professor and Niki & C. L. Max Nikias Chair in Engineering

Professor of Electrical and Computer Engineering, Computer Science,
Linguistics, Psychology, Neuroscience, Pediatrics, Otolaryngology-Head & Neck Surgery

SAIL: Signal Analysis and Interpretation Laboratory
http://sail.usc.edu
Email: shri@usc.edu

February 21, 2023
Language-Based AI Agent Interaction with Children
@IWSDS'23, Los Angeles, USA

Outline

Framing human-centered machine intelligence

- *with a spotlight on children*

Some technical highlights

- *a focus on speech, language and multimodal communication*

Application use cases

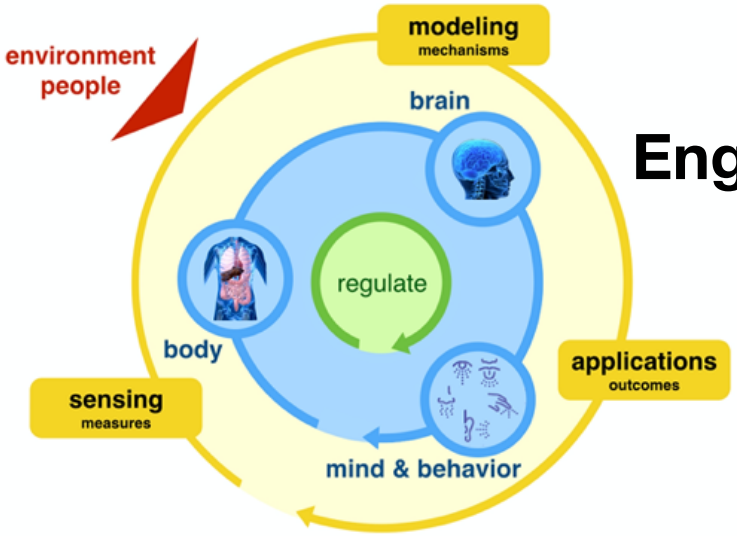
- *highlight behavior phenotyping: Autism Spectrum Disorder*

Human-centered Machine Intelligence: Promise & Possibilities

- **Exciting and accelerating converging advances**
  - *technologies*: sensing, computing, machine learning, data communication, interfaces (e.g., devices on/with/by people)
  - *people*: amazing cross-disciplinary partnerships, resource sharing across societal application domains

- **Novel possibilities to help understand, support, and enhance the human condition and experience**
  - Including during development, healthy and otherwise

Human-centered Multimodal Machine Intelligence



Engineering methods and technologies to

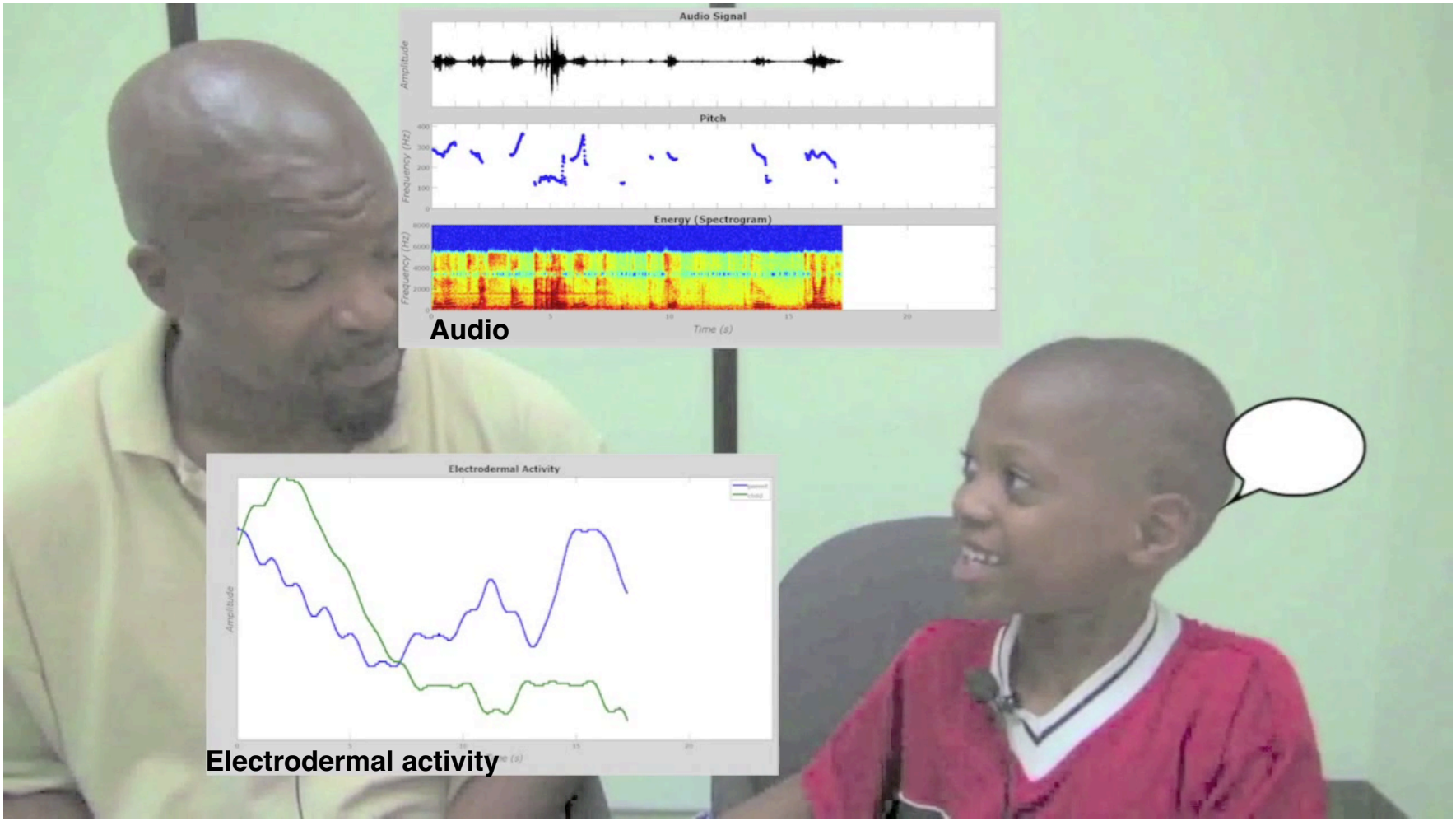
- *Understand human condition*
- *Support and enhance human experiences*

Human centered view: characterizing data/information ***about, from and for people***

- includes knowledge about how ***people*** perceive, process and use (human) data

Focus of this talk: *child-inclusive interactions*

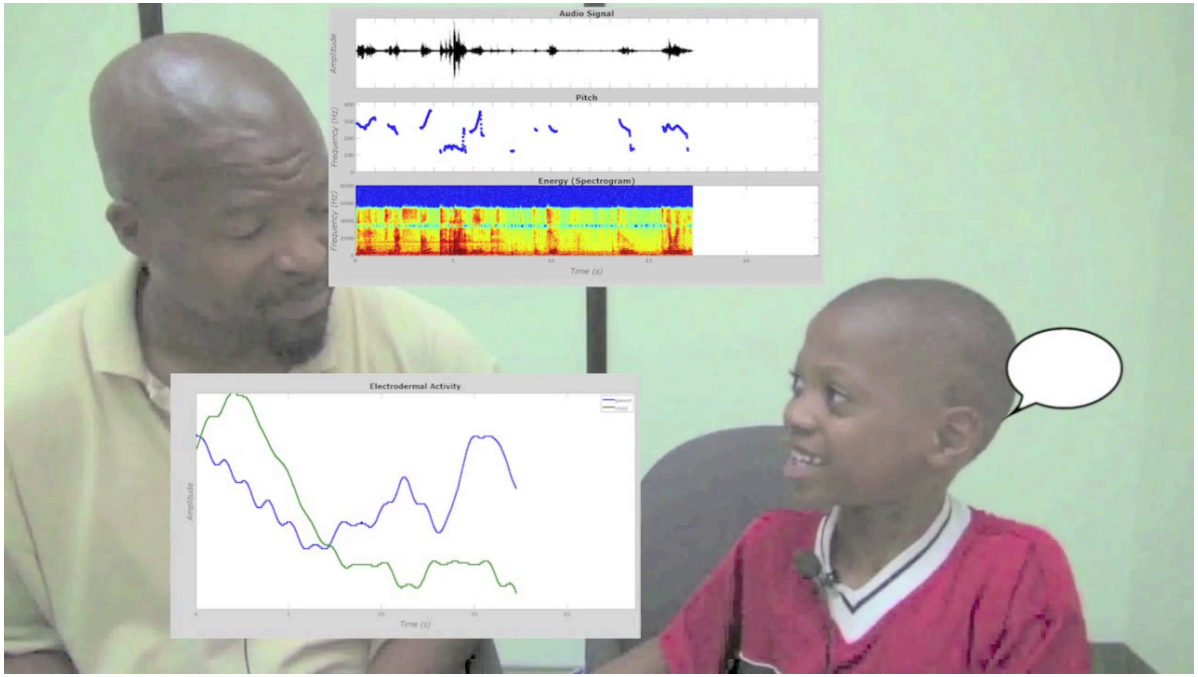
Rich Understanding of Multimodal Behavior and Interaction



Example: Parent and child creating a story together

Human-centered Machine Intelligence: Challenge and Opportunity

Rich Understanding of Multimodal Behavior and Interaction

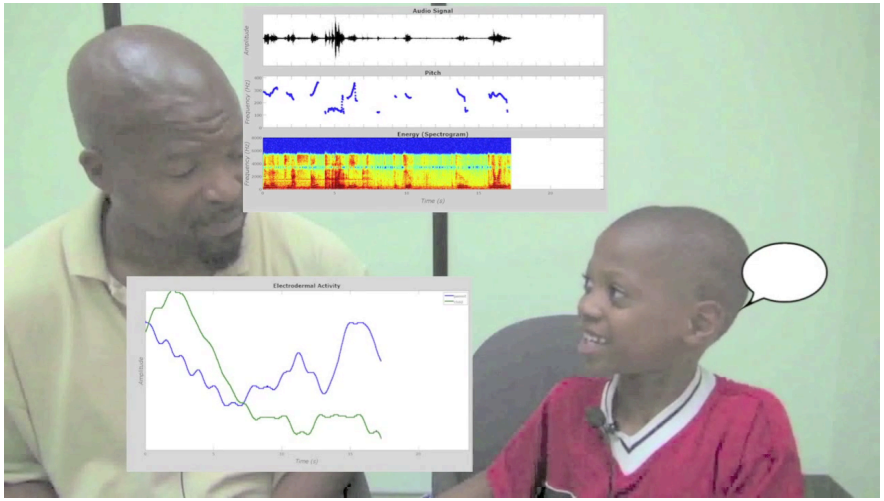


Example: Parent and child creating a story together

Verbal and nonverbal behavior encode and provide access to ***intent, emotions****, and a variety of information about* ***traits*** *(age, gender, appearance…),* ***physical/ psychological/health state,*** *and* ***interaction context.***
These attributes/constructs are often intricately related.

Human-centered Machine Intelligence: Challenge and Opportunity

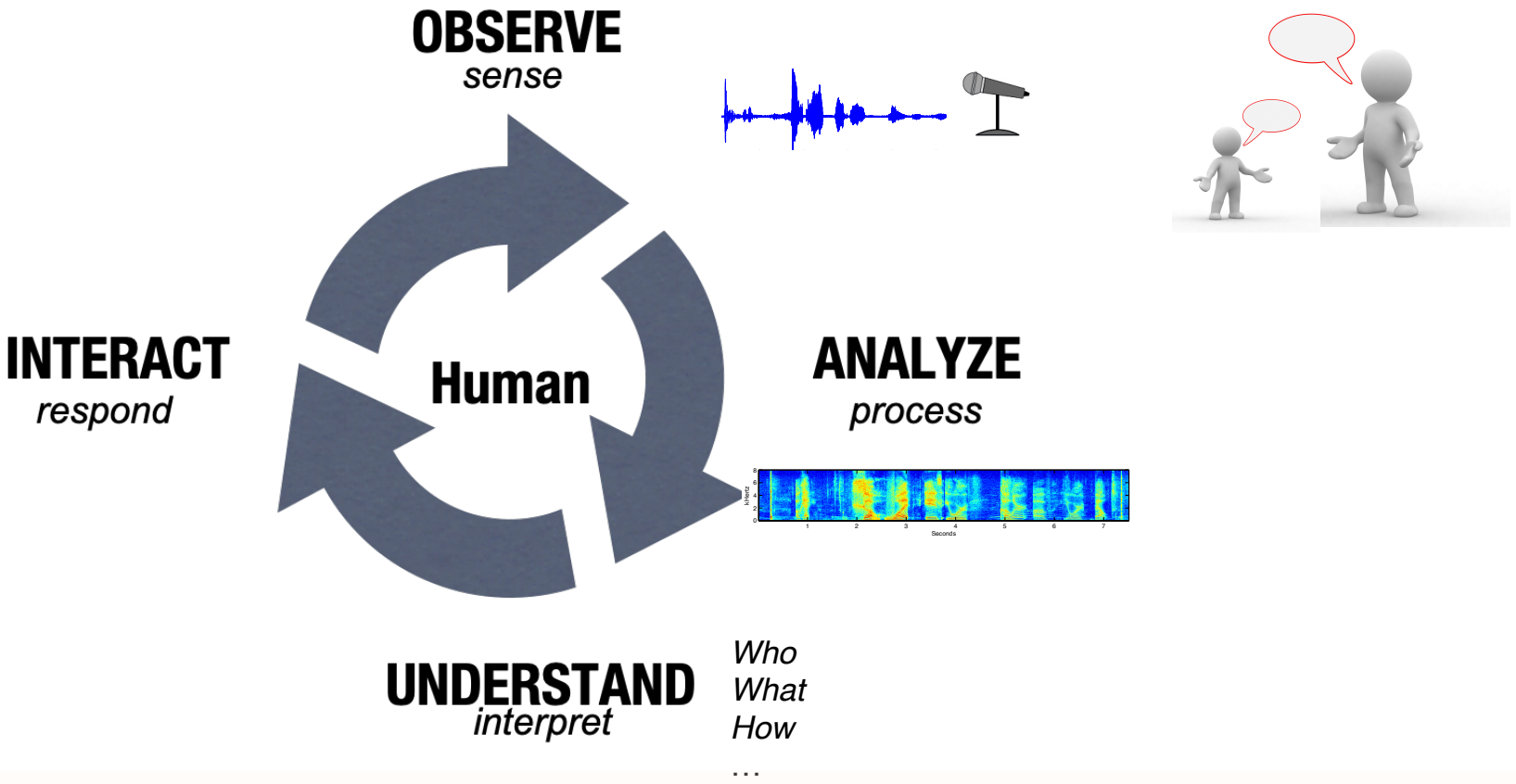
Rich Understanding of Multimodal Behavior and Interaction



- Technologies that work for <u>everyone</u> and in <u>all</u> contexts: understand and create experiences consistent with the **rich variety** in *who, what, where, how, when*,…

Twin goals: **Understanding <u>and</u> addressing variability within and across people and their contexts**
including in the presence of neuro-cognitive differences

Human-centered Machine Intelligence Ecosystem



- "Sounds, Words, Sight" offer a peek into and (hidden) human state and behavior
- Can be complemented by neural and physiological measures
- Screening, diagnostic, intervention support in learning, healthcare, entertainment,…

Child-inclusive Conversational Agents and Interfaces: A continuing journey

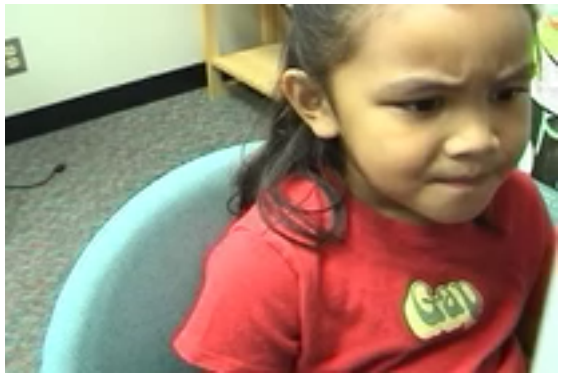
AT&T Bell Labs, 1996

FRUSTRATION

8-14 YEARS

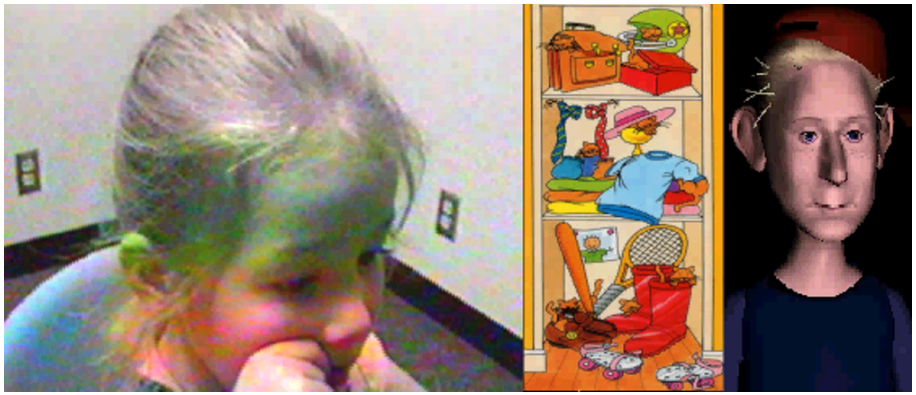
POLITENESS

CONFIDENT VS. UNCERTAIN



USC 2002

PRE-K



- S. YILDIRIM, S. NARAYANAN AND A. POTAMIANOS. DETECTING EMOTIONAL STATE OF A CHILD IN A CONVERSATIONAL COMPUTER GAME. COMPUTER, SPEECH, AND LANGUAGE. SPECIAL ISSUE ON AFFECTIVE SPEECH, 2010.
- MATTHEW BLACK, JEANNETTE CHANG AND SHRIKANTH NARAYANAN. AN EMPIRICAL ANALYSIS OF USER UNCERTAINTY IN PROBLEM-SOLVING CHILD-MACHINE INTERACTIONS. PROCEEDINGS OF THE WORKSHOP ON CHILD, COMPUTER AND INTERACTION, CHANIA, GREECE, OCTOBER 2008
- SHRIKANTH NARAYANAN AND ALEXANDROS POTAMIANOS. CREATING CONVERSATIONAL INTERFACES FOR CHILDREN. IEEE TRANS. SPEECH AND AUDIO PROCESSING, 10(2):65-78, 2002.

PREVALENCE OF SELECT HEALTH CONDITIONS (IN THE US)

Condition	Ages	Prevalence*
Autism spectrum disorder	Children (typically diagnosed as children, but persist over lifetime)	1.5% (lifetime)
Posttraumatic stress disorder	Adults	3.5% (one year)
Mood disorders (e.g., depression)	Adults	9.5% (one year)
Alcohol addiction/abuse	All	6.6% (one year)
Illicit drug use (nonmarijuana)	All	2.5% (one year)
Parkinson's disease	> 60 years old	1.9% (lifetime)
Dementia (e.g., Alzheimer's disease)	> 65 years old	6.5% (lifetime)

***Sources listed in:**
Daniel Bone, Chi-Chun Lee, Theodora Chaspari, James Gibson, and Shrikanth Narayanan. Signal Processing and Machine Learning for Mental Health Research and Clinical Applications. IEEE Signal Processing Magazine. 34(5): 189-196, September 2017

Autism Spectrum Disorder

Technologies for Rich Understanding of Expressive Behavior and Interaction?

Computational Targets

Joint Attention

Turn-taking

Shared enjoyment

Behavioral Synchrony

CREDIT: WPS/ADOS TRAINING VIDEO

- **1 in 44 US children diagnosed with ASD (CDC, 2021)**
- **ASD characterized by difficulties in social communication, reciprocity; repetitive or stereotyped behaviors and interests**
- **Economic Annual Cost of ASD in the US: $11.5 billion – $60.9 billion (2011 Dollars)**
  CDC https://www.cdc.gov/ncbddd/autism/data.html

Operationalizing…
Behavioral Machine Intelligence

- ***nuts and bolts***: foundational multimodal signal processing of data
  - *from people*: audio/speech, video, text, biosignals (ECG, EEG),..
  - *from their environment***:** location, temperature, light, sound, humidity, air quality,..
- ***construct prediction***: machine learning based methods for automated behavioral coding and characterization
- ***computational modeling***: of interaction processes & mechanisms
- ***translational applications notably in health***: screening, diagnostics, intervention support
  - just in time implementation, tracking response to treatment,..

Shrikanth Narayanan and Panayiotis Georgiou. Behavioral Signal Processing: Deriving Human Behavioral Informatics from Speech and Language. Proceedings of IEEE. 101(5): 1203-1233, May 2013
Daniel Bone, Chi-Chun Lee, Theodora Chaspari, James Gibson, and Shrikanth Narayanan. Signal Processing and Machine Learning for Mental Health Research and Clinical Applications. IEEE Signal Processing Magazine. 34(5): 189-196, September 2017

How is technology helping already?

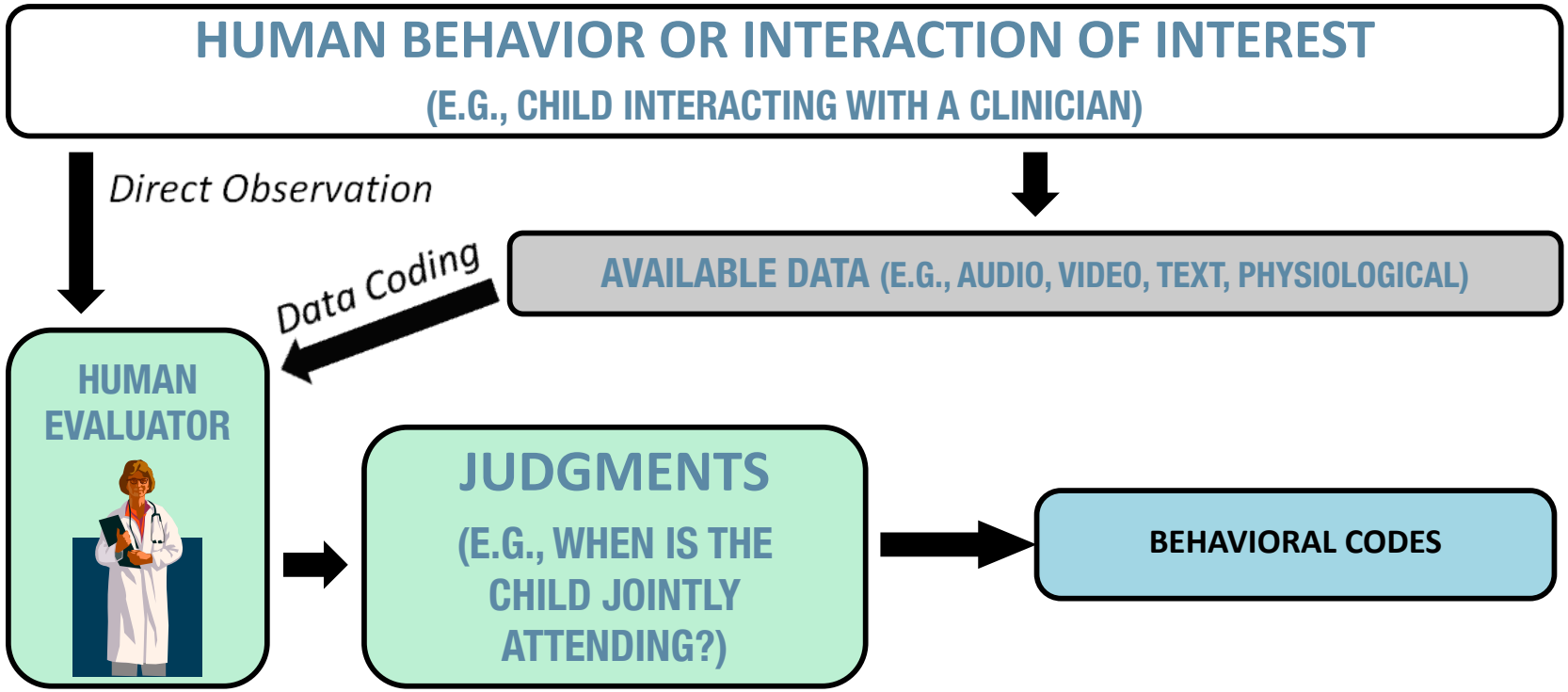
Sensing, signal processing and machine learning are key enablers.

- **wearable and environment/context sensing**
- **foundational technologies for behavior modeling: to 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, ..
- **multimodal affective computing**

SHIFT TO MODELING MORE ABSTRACT, DOMAIN-RELEVANT HUMAN BEHAVIORS
.....NEEDS NEW MULTIMODAL COMPUTATIONAL APPROACHES

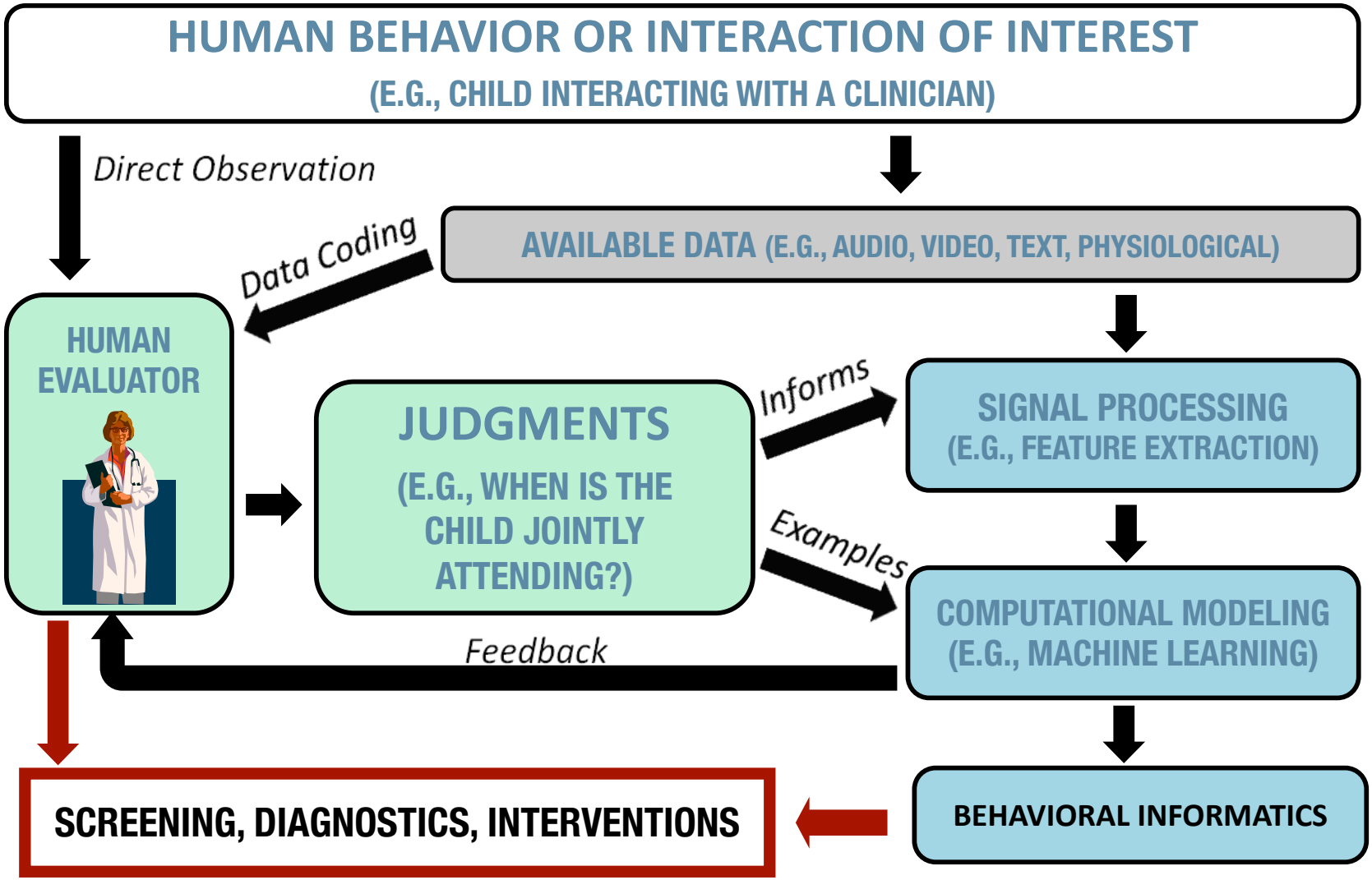
Behavior Coding: Humans in the loop

- Human assessments/judgments on human behavior



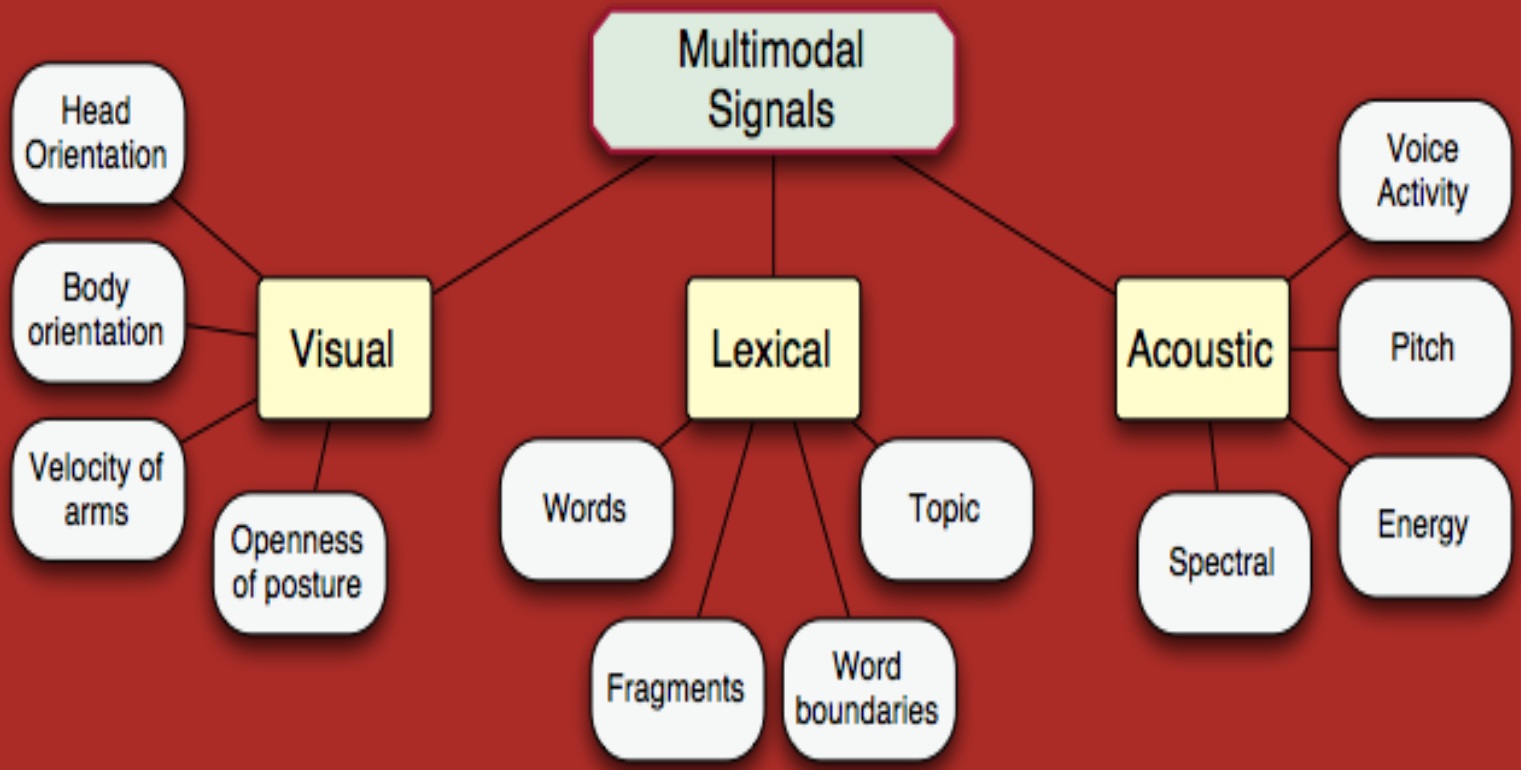
Behavior Modeling: Humans in/on the loop

- Support—than supplant—human (expert) analyses

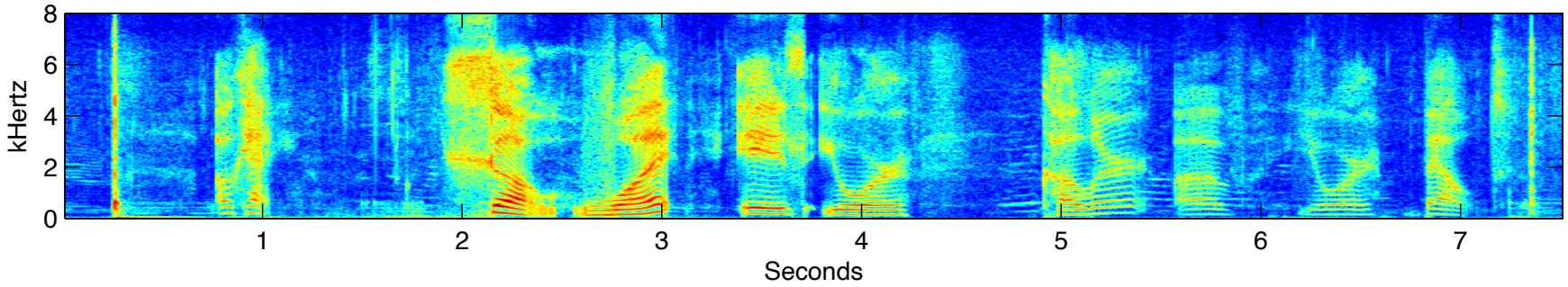


Collaborative integration of human and machine intelligence 15

Processing Human Communication Signals



Who spoke when, for how long, and about what?



Voice Activity Detection

S S S S S S

Detected Speech Regions

Speaker Diarization

1 1 2 1 3 3 2 2

Speakers in the conversation

Audio Event Diarization

Audio event e.g. banging

Audio event

Transcription (Speech Recognition)

Hello!

How are you?

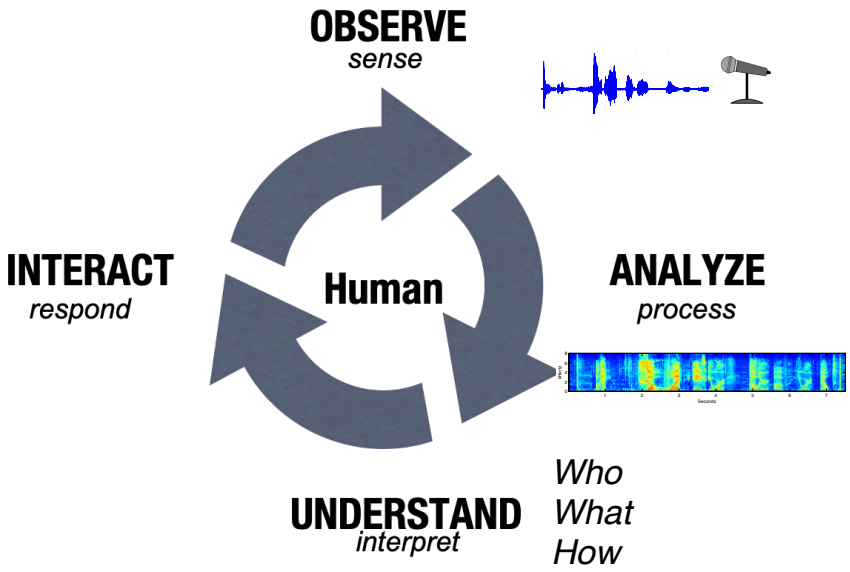
I am good.

We should go.

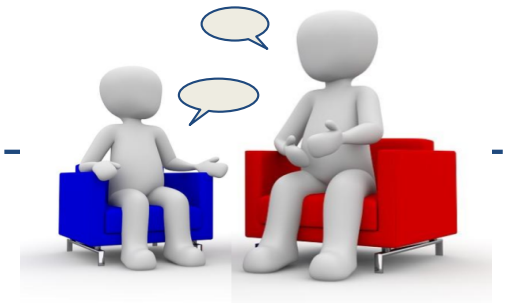
Come on!

Ok!

An Automated Speech Processing Pipeline



Eg., Dyädic child — adult interaction

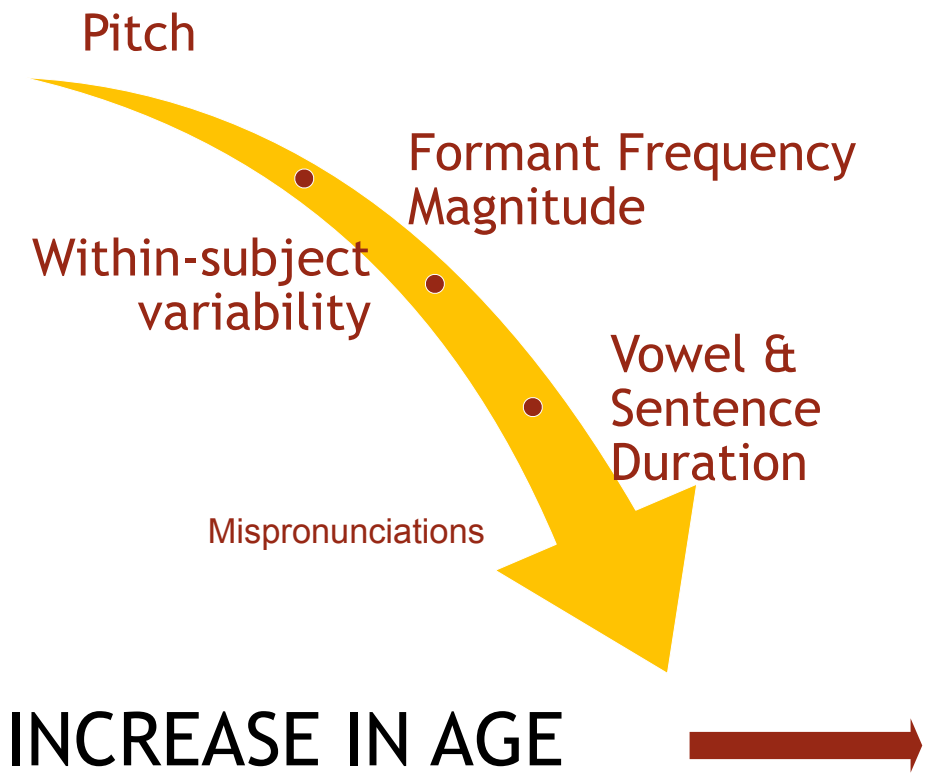


A spotlight on **Processing Children's Vocalization & Speech**

- **What is special about it?**
  - **Review acoustic properties**
- **Robust speech recognition techniques**

Developmental changes revealed in speech signal

Reduction in speech parameter values as a function of age



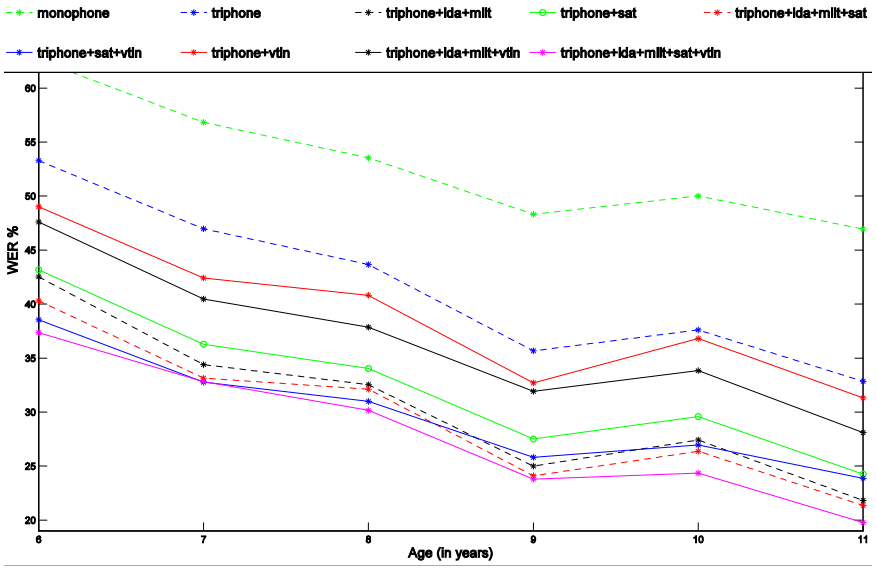
Sungbok Lee, Alexandros Potamianos, and Shrikanth Narayanan. Acoustics of children's speech: Developmental changes of temporal and spectral parameters. J. Acoust. Soc. Am., 105:1455-1468, Mar. 1999 (Selected Research Article)
Sungbok Lee, Alexandros Potamianos, and Shrikanth S. Narayanan. Developmental acoustic study of American English diphthongs. J. Acoust. Soc. Am., 136(4):1880–1894, oct 2014.

Children speech vs Adult speech

Sungbok Lee, Alexandros Potamianos, and Shrikanth Narayanan. Acoustics of children's speech: Developmental changes of temporal and spectral parameters. J. Acoust. Soc. Am., 105:1455-1468, Mar. 1999 (Selected Research Article)
Sungbok Lee, Alexandros Potamianos, and Shrikanth S. Narayanan. Developmental acoustic study of American English diphthongs. J. Acoust. Soc. Am., 136(4):1880–1894, oct 2014.

Early work in ASR of Children Speech

- **Performance varies with age: 2-5 times error than adult speech**
  - 50% relative error reduction due to frequency warping and model adaptation, larger for speakers under 12 years
  - Despite improvement relative error rate is at least 30% higher for 6-9 year olds
  - Age-dependent models provide an additional 10% relative error rate reduction
  - Front end vocal tract normalization (especially when training-testing age mismatch), speaker normalization, other spectral adaptation techniques



WER decreases almost linearly with increase in age

- Shrikanth Narayanan and Alexandros Potamianos. Creating conversational interfaces for children. IEEE Trans. Speech and Audio Processing, 10(2):65-78, 2002.
- Alexandros Potamianos and Shrikanth Narayanan. Robust recognition of children's speech. IEEE Trans. Speech and Audio Processing, 11:603-616, Nov. 2003.
- Prashanth Gurunath Shivakumar, Alexandros Potamianos, Sungbok Lee and Shrikanth Narayanan. Improving Speech Recognition for Children using Acoustic Adaptation and Pronunciation Modeling. In Proceedings of Workshop on Child Computer Interaction (WOCCI 2014), Singapore, September, 2014

Improving Speaker Detection & Diarization In Interactions Involving Children

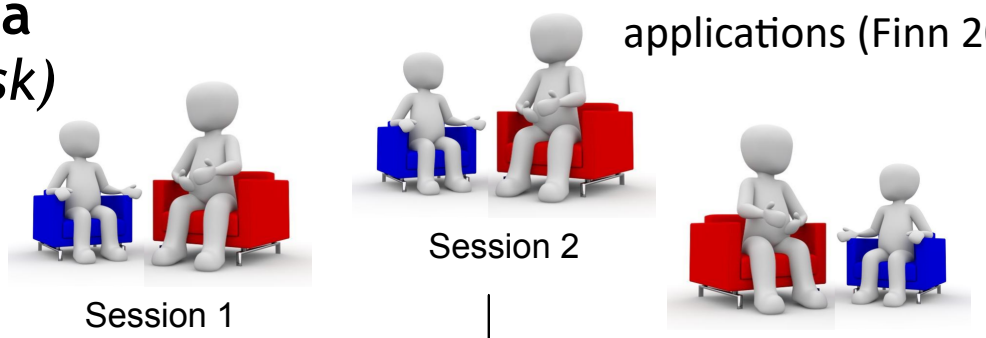
- **Large within-class variability** especially for child from age, gender, clinical symptom severity (Lee 1999; 2014, Gerosa 2009)
- **Lack of sufficient & balanced** training data covering different factors/conditions

- Rimita Lahiri, Manoj Kumar, Somer Bishop, and Shrikanth Narayanan. Learning domain invariant representations for child-adult classification from speech. In Proceedings of ICASSP, May 2020.
- Nithin Rao, Manoj Kumar, So Hyun Kim, Catherine Lord, and Shrikanth Narayanan. Meta-learning for robust child-adult classification from speech. In Proceedings of ICASSP, May 2020.
- Manoj Kumar, So Hyun Kim, Catherine Lord, and Shrikanth Narayanan. Speaker Diarization for Naturalistic Child-Adult Conversational Interactions using Contextual Information.. J. Acoust. Soc. Am., 147(2):EL196–EL200, February 2020.
- Monisankha Pal, Manoj Kumar, Raghuveer Peri, Tae Jin Park, So Hyun Kim, Catherine Lord, Somer Bishop, and Shrikanth Narayanan. Meta-learning with Latent Space Clustering in Generative Adversarial Network for Speaker Diarization. IEEE/ACM Transactions on Audio, Speech and Language Processing. 29: 1204-1219, 2021

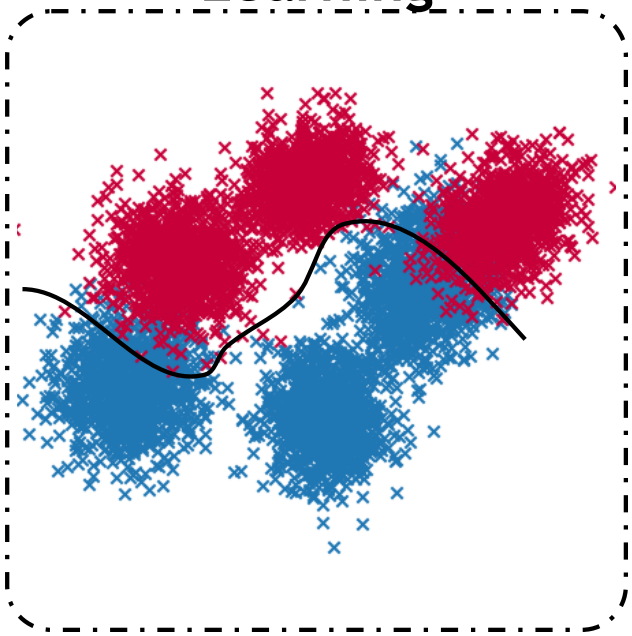
Meta-Learning for Child/Adult Classification

Meta Learning: (Learning to learn) Paradigm of supervised learning developed for low-resource applications (Finn 2017, Ravi 2016)

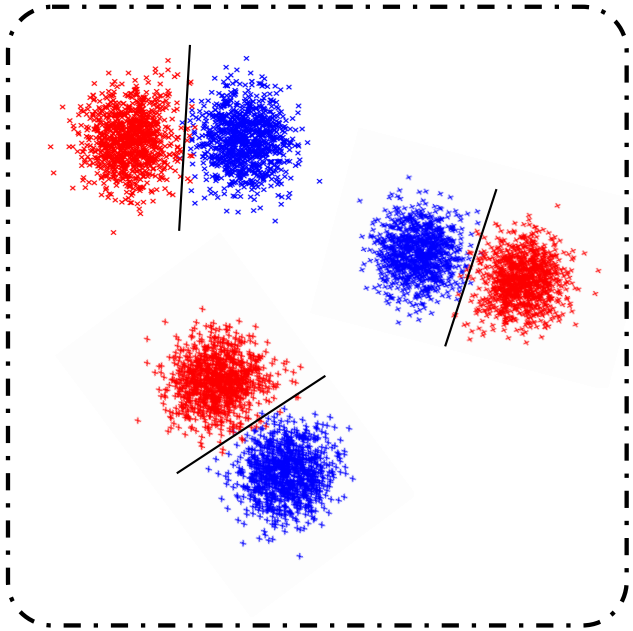
Training Data
(Session ≡ Task)



Conventional Learning



Meta Learning



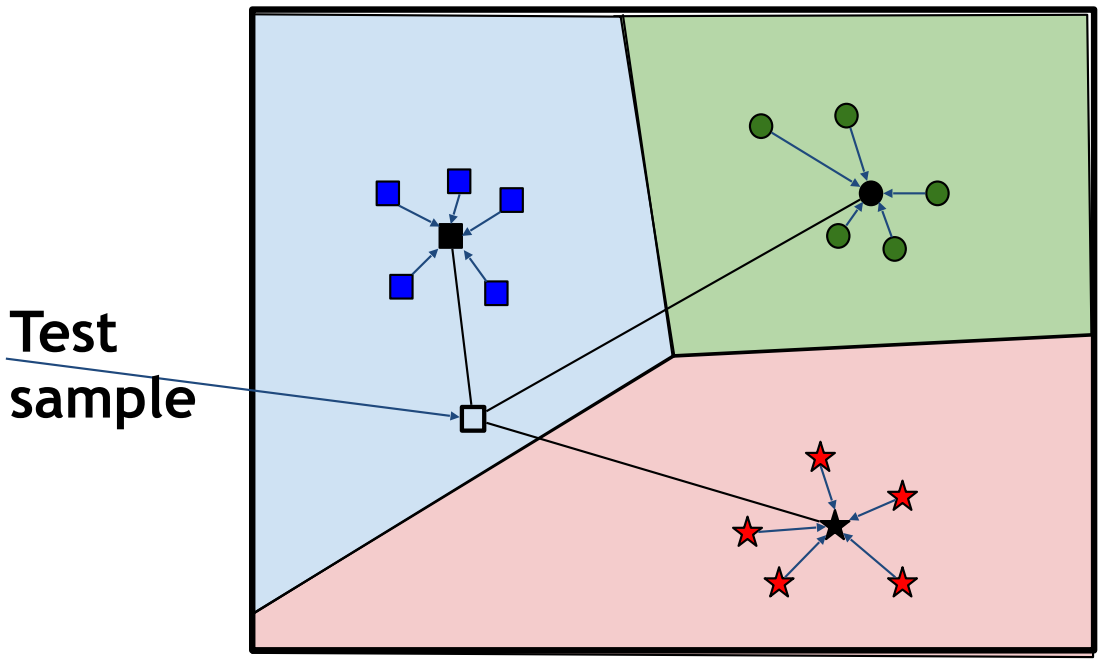
Key insight

+ Modeling child/adult classification across sessions as multiple, related tasks ➔ Learn task-invariant representations using meta-learning

- Learnable information constrained by input (e.g., x-vectors)

- At each training step, class labels constrained to *child* or *adult*

Meta learning



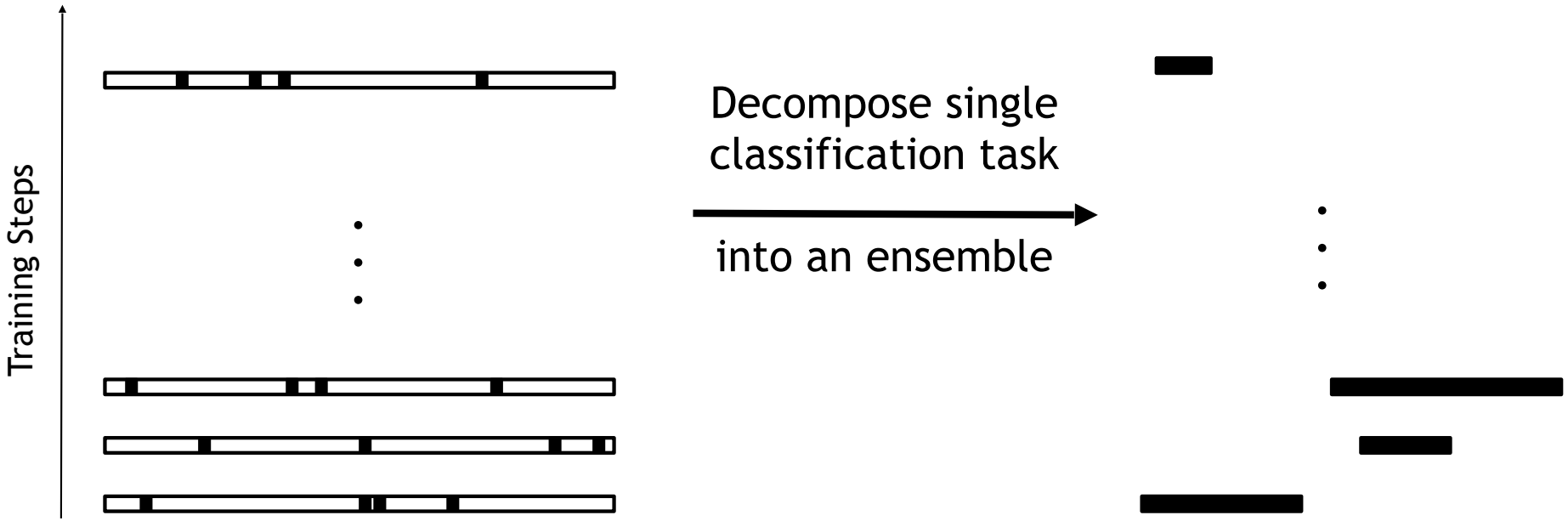
Using Prototypical Networks: Represent each class using centroid (prototype)

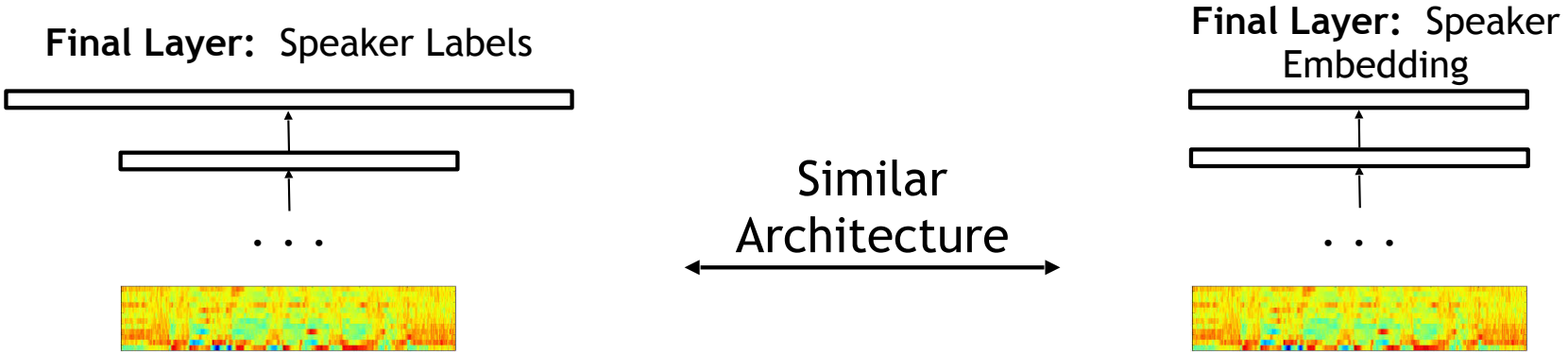
Extension to a generic speaker embedding training?

Converting to Ensemble of Tasks

Conventional Classification

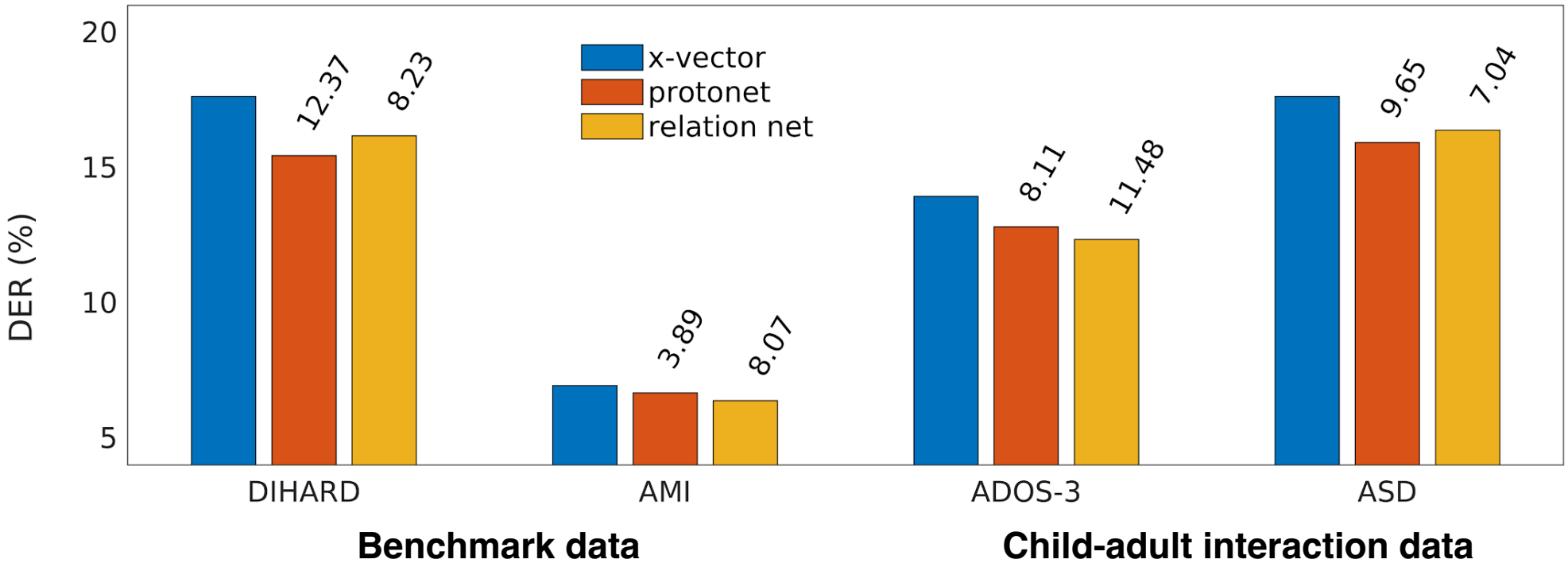
Proposed (Meta Learning)





Results: Speaker Diarization

Diarization Error Rate (%) for various public and internal corpora. Improvements with protonets and relation networks are indicated on top of the respective bars



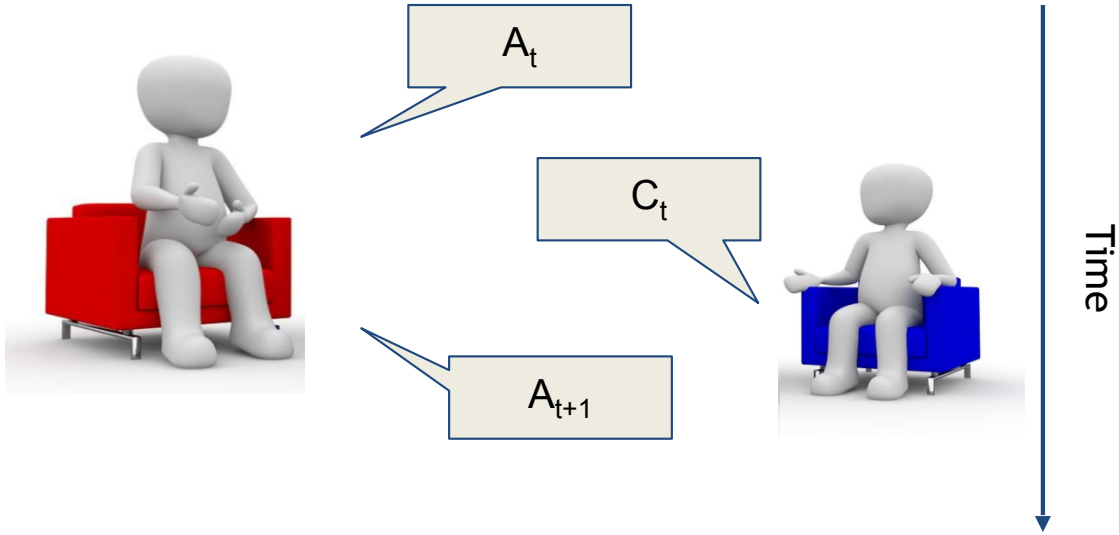
Improving Speech Recognition For Children

- Manoj Kumar, So Hyun Kim, Catherine Lord, Thomas Lyon, and Shrikanth Narayanan. Leveraging Linguistic Context in Dyadic Interactions to Improve Automatic Speech Recognition for Children. Computer, Speech and Language, 63, 2020.
- Prashanth Gurunath Shivakumar, Shrikanth Narayanan. End-to-End Neural Systems for Automatic Children Speech Recognition: An Empirical Study. Computer Speech & Language. 72:101289, 2022

Linguistic Context Adaptation

Goal:

Improve ASR for C_t using information from A_t ,A_{t-1} (backward context) and/or A_{t+1},A_{t+2} (forward context)



N-gram interpolation:

$$P(w|L_{adapt}) = \begin{cases} \lambda P(w|L_{base}) + (1-\lambda)P(w|L_{context}) & w \in L_{base} \cap L_{context} \\ \lambda P(w|L_{base}) & w \notin L_{context} \\ (1-\lambda)P(w|L_{context}) & w \notin L_{base} \end{cases}$$

where w represents n-gram

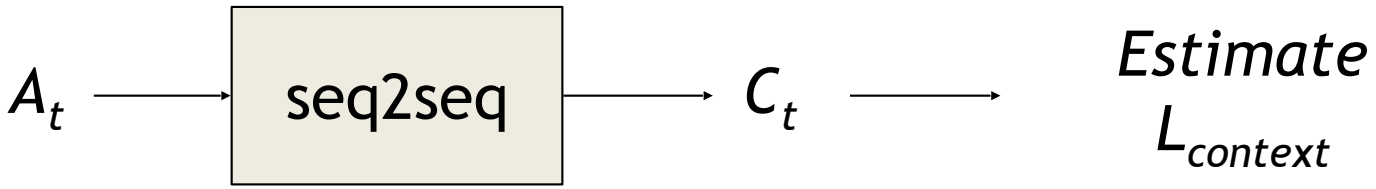
Lexical and Semantic Context

Lexical Context:

- Capture matched tokens
- Use the ASR hypothesis for P_t to estimate $L_{context}$

Semantic Context:

- Goal: Capture "Concept/Meaning" of the utterance
- Realized using a sequence-to-sequence neural network
  - Inspired by similar applications in conversational agents (Sutskever 2014, Vinyals 2015, Li 2016)
  - Decode context utterances (A_t) to obtain hypothesis for C_t



Experiments

TEST DATA
FI: Child Forensic Interview
ASD: Child-clinician interaction in Autism context

Baseline ASR Models

- Hybrid DNN-HMM model w/ TDNN + BiLSTM layers
  - Child: Trained with CSLU, CUKids, CHIMP, CMUKids[1] (ASD: 76.23, FI: 62.53)
  - Adult: Off-the-shelf model from the ASPIRE recipe[2] (ASD: 33.15, FI: 26.18)

Domain Adaptation

- Acoustic model: Final DNN layer trained for a single epoch
- Language model: Linearly interpolated, λ estimated using CV

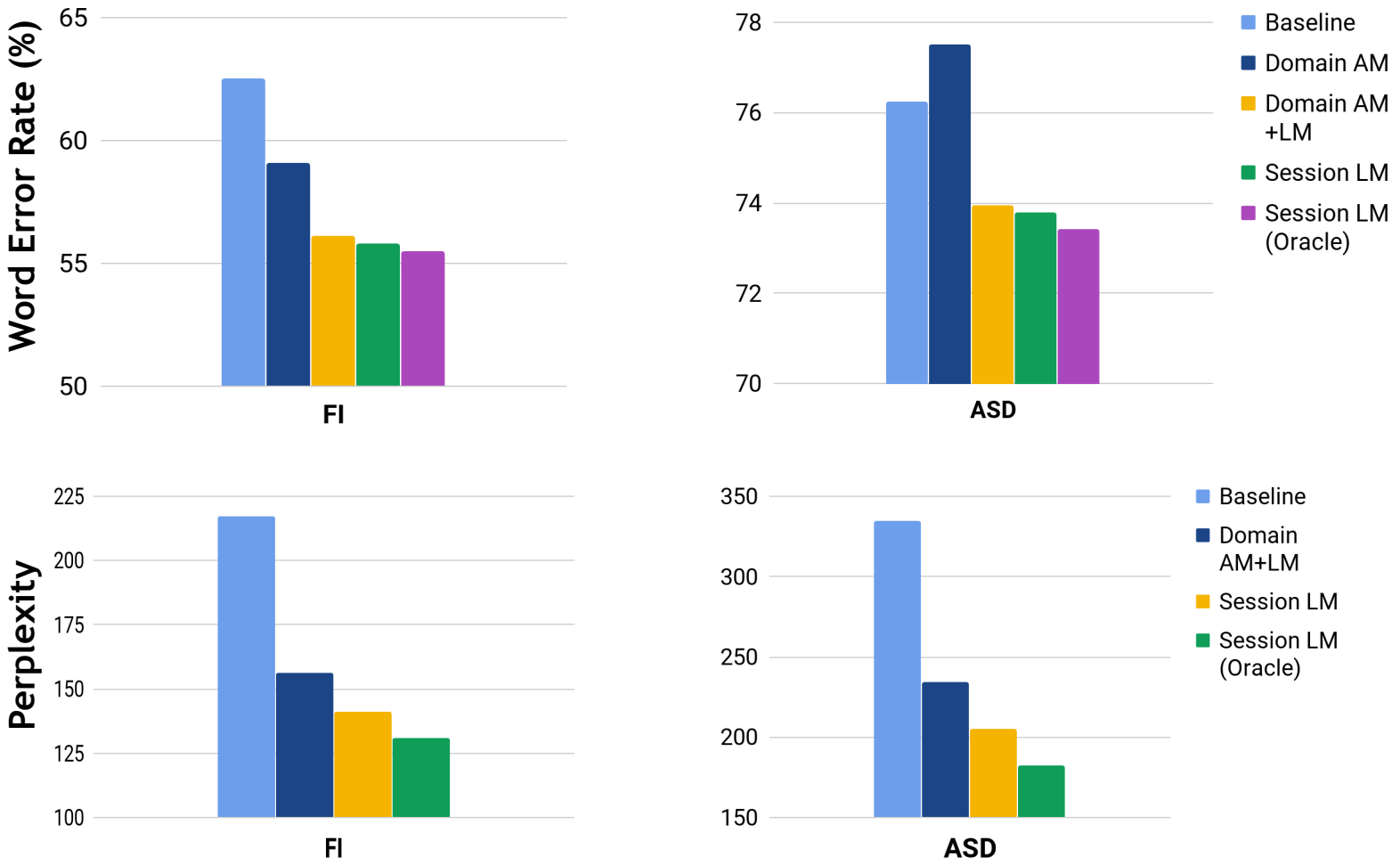
Session Adaptation

- Global *vs* Local adaptation; Language model only
- Analyze direction and strength of context within local adaptation
- Global: λ estimated using CV; Local: λ = 0.5

1. Kumar M, Kim SH, Lord C, Lyon TD, Narayanan S. *Leveraging Linguistic Context in Dyadic Interactions to Improve Automatic Speech Recognition for Children*. Computer Speech and Language 2020
2. https://kaldi-asr.org/models/m1

Results

- local interlocutor context adaptation using ASR hypotheses useful



Results for domain adaptation and global-level session adaptation in terms of word error rate (%) and language perplexity for both domains

32

End-to-End Speech Recognition for Children

- Will End-to-End ASR's ability to exploit large amounts of speech data impute for anomalies/challenges found in children speech?
- How do the end-to-end systems perform for children of different ages?
- Which neural network architectures are most effective for children speech recognition?

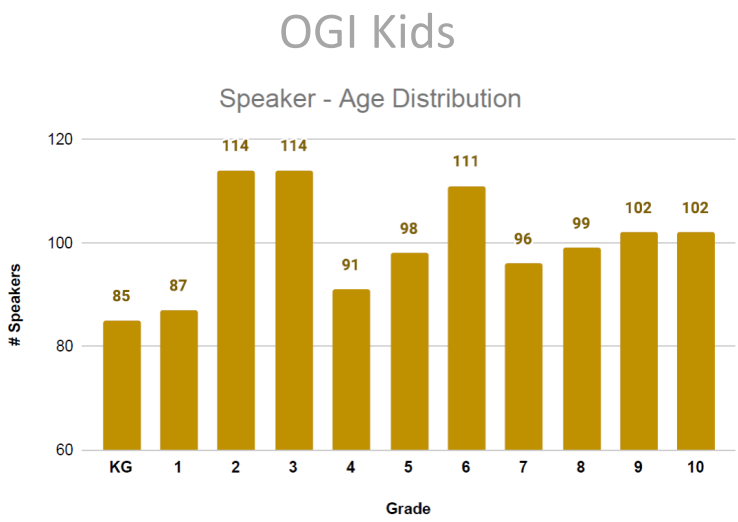
- **Assess State-of-the-art End-to-End Architectures for children speech**
  - ResNet (Residual Convolutional Neural Networks)
  - TDS (Time Depth Separable Convolutional Neural Networks)
  - Transformers
- **Training Objectives:**
  - CTC (Connectionist Temporal Classification)
  - Sequence-to-Sequence

Prashanth Gurunath Shivakumar, Shrikanth Narayanan. End-to-End Neural Systems for Automatic Children Speech Recognition: An Empirical Study. Computer Speech & Language. 72:101289, 2022

Datasets

- Adult Speech Corpora
  - LibriSpeech (960 hours)
  - LibriVox (53,800 hours)
- Children Speech Corpora

Corpus	Train	Development	Test
MyST	88318 Utterances 197.72 hours 678 speakers	5000 Utterances 12.23 hours 25 speakers	5000 Utterances 13.28 hours 34 speakers
OGI Kids			1099 Utterances 30.5 hours 1099 speakers



Results: Adult Acoustic Models

	AM	LM	LIB test-clean		MyST test		OGI Kids	
			LER	WER	LER	WER	LER	WER
KALDI	TDNN-F DNN-HMM	4-gram	2.22	5.94	26.98	47.90	36.04	53.55
WAV2LETTE R++	ResNet + CTC	GCNN	1.45	3.28	20.49	32.48	34.27	**49.06**
	ResNet + S2S	GCNN-wp	1.85	3.79	64.98	86.09	83.13	94.36
	Transformer + CTC	GCNN	1.12	2.58	17.43	**26.23**	52.58	55.92
	Transformer + S2S	GCNN-wp	**0.90**	**2.40**	32.73	45.96	72.27	88.84

- SOTA on Adult Speech Recognition
- E2E Systems significant improvement over DNN-HMM based systems
- 11x worse WER on Children Speech Recognition (MyST Corpus)
- 20x worse WER on OGI Kids Corpus

Results: Adult Acoustic Models: SSL data augmentation

	AM	LM	LIB test-clean		MyST test		OGI Kids	
			LER	WER	LER	WER	LER	WER
KALDI	TDNN-F DNN-HMM	4-gram	2.22	5.94	26.98	47.90	36.04	53.55
WAV2LETTE R++	ResNet + CTC	GCNN	1.45	3.28	20.49	32.48	34.27	**49.06**
	ResNet + S2S	GCNN-wp	1.85	3.79	64.98	86.09	83.13	94.36
	Transformer + CTC	GCNN	1.12	2.58	17.43	**26.23**	52.58	55.92
	Transformer + S2S	GCNN-wp	**0.90**	**2.40**	32.73	45.96	72.27	88.84
WAV2LETTE R++ (55k hours)	ResNet + CTC	GCNN	1.09	2.45	18.33	26.00	31.59	37.78
	ResNet + S2S	GCNN-wp	1.10	2.65	27.43	37.77	86.36	90.64
	Transformer + CTC	GCNN	1.03	2.41	16.79	**24.26**	52.15	55.67
	Transformer + S2S	GCNN-wp	0.80	**2.17**	26.89	40.51	70.77	85.15

- **Addition of 53.8k hours of Adult speech data -> Significant Improvements for Children**

Results: Adult Acoustic Models Adaptation on Children Speech

	AM	LM	LIB test-clean		MyST test		OGI Kids	
			LER	WER	LER	WER	LER	WER
KALDI	TDNN-F DNN-HMM	4-gram	2.22	5.94	26.98	47.90	36.04	53.55
WAV2LETTE R++	ResNet + CTC	GCNN	1.45	3.28	20.49	32.48	34.27	**49.06**
	ResNet + S2S	GCNN-wp	1.85	3.79	64.98	86.09	83.13	94.36
	Transformer + CTC	GCNN	1.12	2.58	17.43	**26.23**	52.58	55.92
	Transformer + S2S	GCNN-wp	**0.90**	**2.40**	32.73	45.96	72.27	88.84
WAV2LETTE R++	ResNet + CTC	GCNN	1.09	2.45	18.33	26.00	31.59	37.78
	ResNet + S2S	GCNN-wp	1.10	2.65	27.43	37.77	86.36	90.64
	Transformer + CTC	GCNN	1.03	2.41	16.79	**24.26**	52.15	55.67
	Transformer + S2S	GCNN-wp	0.80	**2.17**	26.89	40.51	70.77	85.15
(55K hours) WAV2LETTE R++ (Adapted on MyST Corpus 197 hours)	ResNet + CTC	4-gram			12.48	18.23	23.84	34.73
	ResNet + S2S	4-gram			20.30	27.22	62.27	73.64
	Transformer + S2S	4-gram			11.68	16.82	63.26	71.39
	Transformer + S2S	6-gram-wp			11.78	16.81	63.26	71.11

- **Adaptation on Children Speech Data -> Significant improvements**

End to End Child-centric ASR Summary

- *Benefits established with end-to-end ASR for adult speech **do not** translate completely to children speech*
- *ASR for children is 10 – 19 times worse than Adults and 6 times worse despite adaptation on children speech*
- *End-to-end systems provide near constant improvements over all age categories after adaptation on child speech (DNN-HMM is more sensitive to age categories of adaptation data).*
- *Absolute WER with the end-to-end systems better than DNN-HMM*
  - *Gap in performance between adult and children wider for end-to-end systems compared to DNN-HMM ASR*
- *Addition of large amounts of adult speech is found to be beneficial (more benefits for ASR for younger children)*
- *Transformer network architectures are the best performing models when the train–test mismatch is low, however they do not generalize well*
- *CTC loss based models are robust to children speech recognition; Sequence-to-sequence models can breakdown during high mismatch conditions*
- *Better performance with greedy decoding without language model*

Prashanth Gurunath Shivakumar, Shrikanth Narayanan. End-to-End Neural Systems for Automatic Children Speech Recognition: An Empirical Study. Computer Speech & Language. 72:101289, 2022

Some Case Studies

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

» READING ASSESSMENT

✓HELP HANDLE NEW DATA, CREATE NEW MODELS TO OFFER NEW INSIGHTS

✓CREATE TOOLS FOR SCIENTIFIC DISCOVERY

» CAUSAL INDICATORS OF TRUTHFULNESS IN FORENSIC INTERVIEWS

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

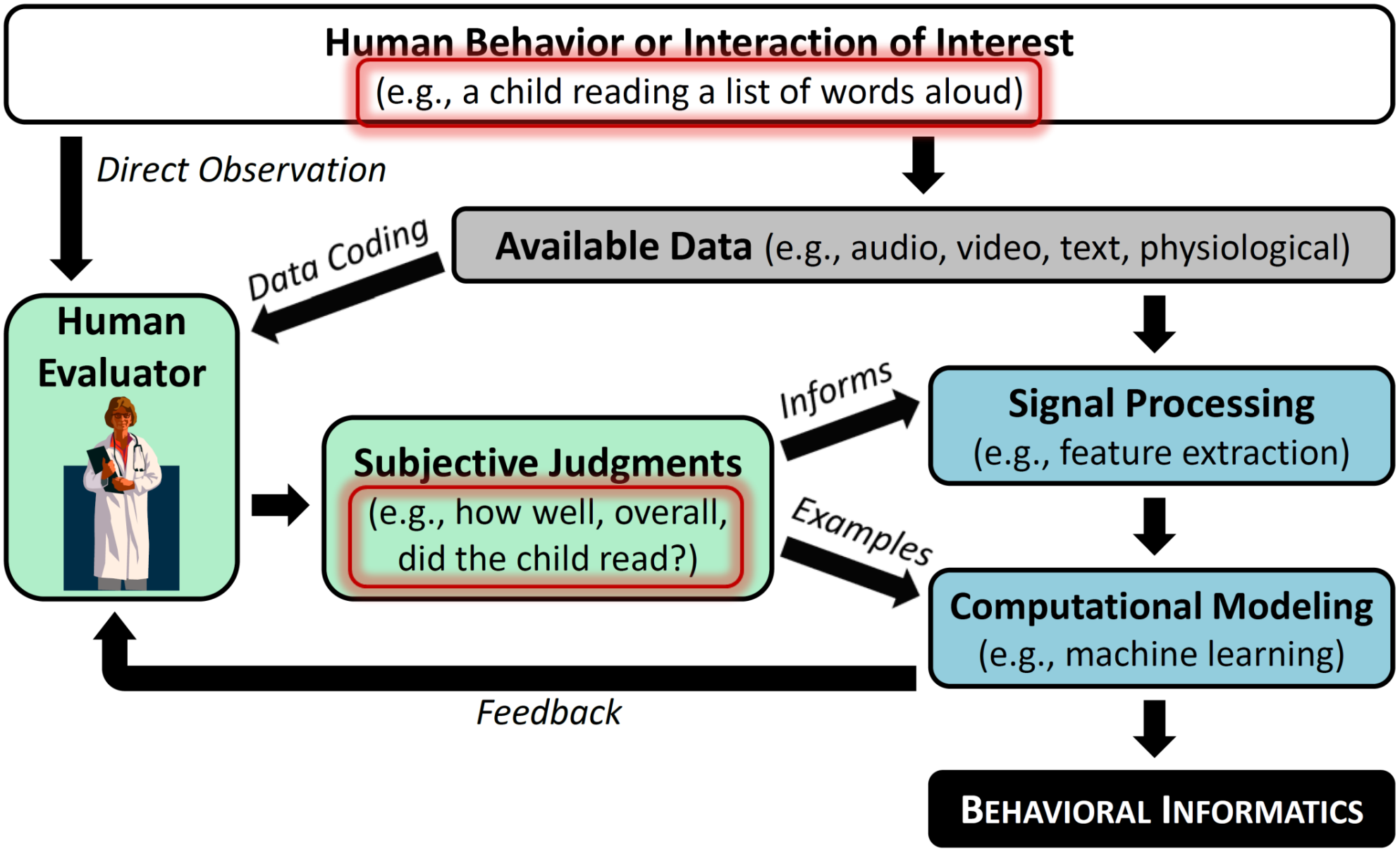
» SCREENING AND DIAGNOSIS IN AUTISM SPECTRUM DISORDER

Application:

Reading assessment

MATTHEW BLACK, JOSEPH TEPPERMAN AND SHRIKANTH NARAYANAN. AUTOMATIC PREDICTION OF CHILDREN'S READING ABILITY FOR HIGH-LEVEL LITERACY ASSESSMENT. IEEE TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING. 19(4): 1015 - 1028, 2011.

Automatic Literacy Assessment



MATTHEW BLACK, JOSEPH TEPPERMAN AND SHRIKANTH NARAYANAN. AUTOMATIC PREDICTION OF CHILDREN'S READING ABILITY FOR HIGH-LEVEL LITERACY ASSESSMENT. IEEE TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING. 19(4): 1015 - 1028, 2011.

Application:

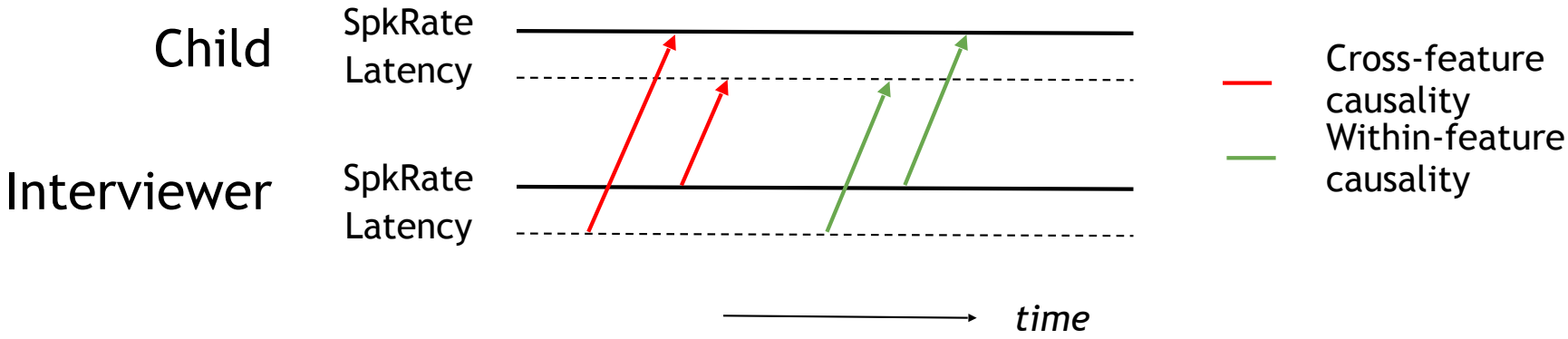
Tools for assisting in Child Forensic Interviews

ZANE DURANTE, VICTOR ARDULOV, MANOJ KUMAR, JENNIFER GONGOLA, THOMAS LYON, SHRIKANTH NARAYANAN. CAUSAL INDICATORS FOR ASSESSING THE TRUTHFULNESS OF CHILD SPEECH IN FORENSIC INTERVIEWS. COMPUTER SPEECH & LANGUAGE. 71:101263, 2022

Methods

Descriptors: Speaking Rate, Latency, Overall Silence Fraction

- Session-level functionals: Mean & Standard Deviation
- F-statistic from Granger causality test



Ensemble Models:

AdaBoost & Random Forest (Base model: Decision tree)

Baselines:

Human prediction performance (Gongola 2017)

Random guessing

Insights

Which features assist classification?

	Truth-Telling	Disclosure
Feature-Functionals	Recall: Interviewer Speaking Rate ⬇ Child Speaking Rate ⬆ Child Silence Fraction ⬇	Recall: Interviewer Speaking Rate ⬇ Child Speaking Rate ⬇ Interviewer Silence Fraction ⬇
		Rapport: Child Speaking Rate ⬇ Interviewer Latency ⬇
Granger Causality		Recall: Latency -> Speaking Rate ⬇
	Rapport: Speaking Rate -> Speaking Rate	Rapport:

Overall, children speak faster while being truthful, but slow down when forced to ignore the confederate's instructions

Insights

Which features assist classification?

	Truth-Telling	Disclosure
Feature-Functionals	Recall: Interviewer Speaking Rate ⬇ Child Speaking Rate ⬆ Child Silence Fraction ⬇ Rapport: Interviewer Latency Interviewer Silence F	Recall: Interviewer Speaking Rate ⬇ Child Speaking Rate ⬇ Interviewer Silence Fraction ⬇ Rapport:
Granger Causality	Recall: Speaking Rate -> Late Rapport: Speaking Rate -> Speaking Rate	Rapport:

Child pause for smaller duration while truthfully recalling the transgression

Takeaway:
Understanding interaction mechanisms

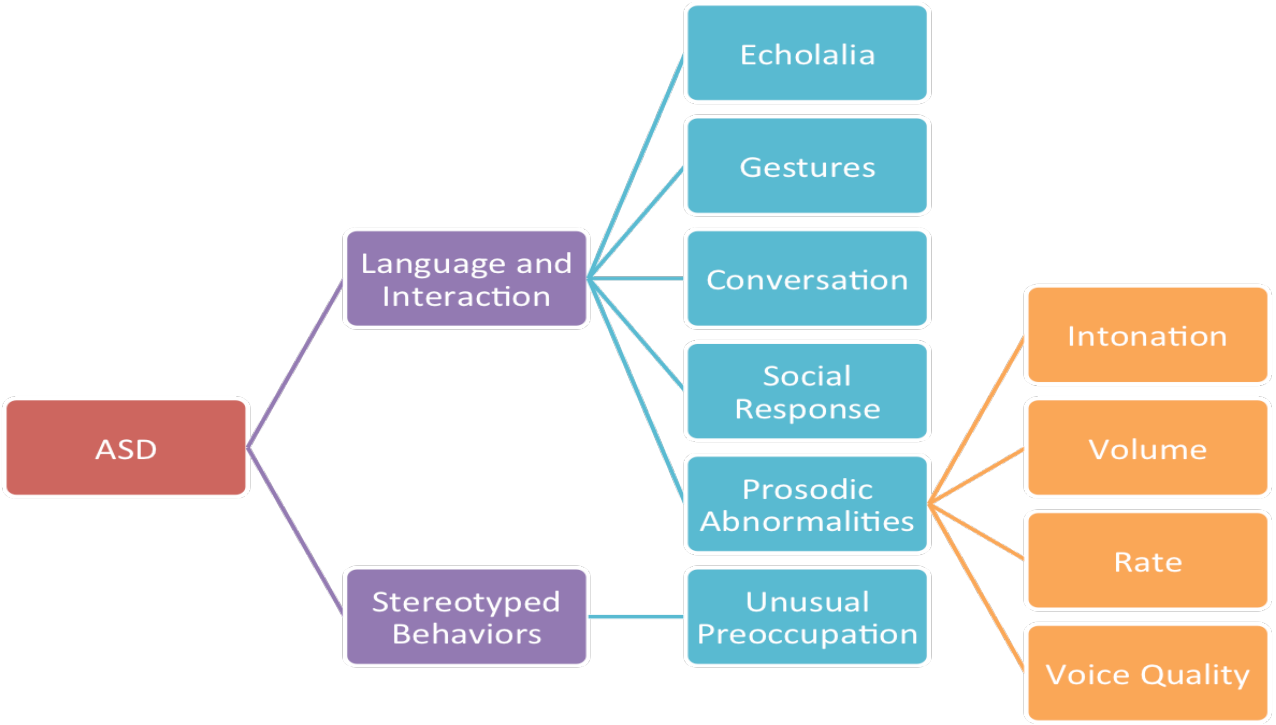
- Better-than-chance classification, but lots to improve!
- Group differences between truthful/deceit manifest in descriptors
- Turn-level truthfulness identification?

Application:

Autism Spectrum Disorder

Autism Spectrum Disorder (ASD)

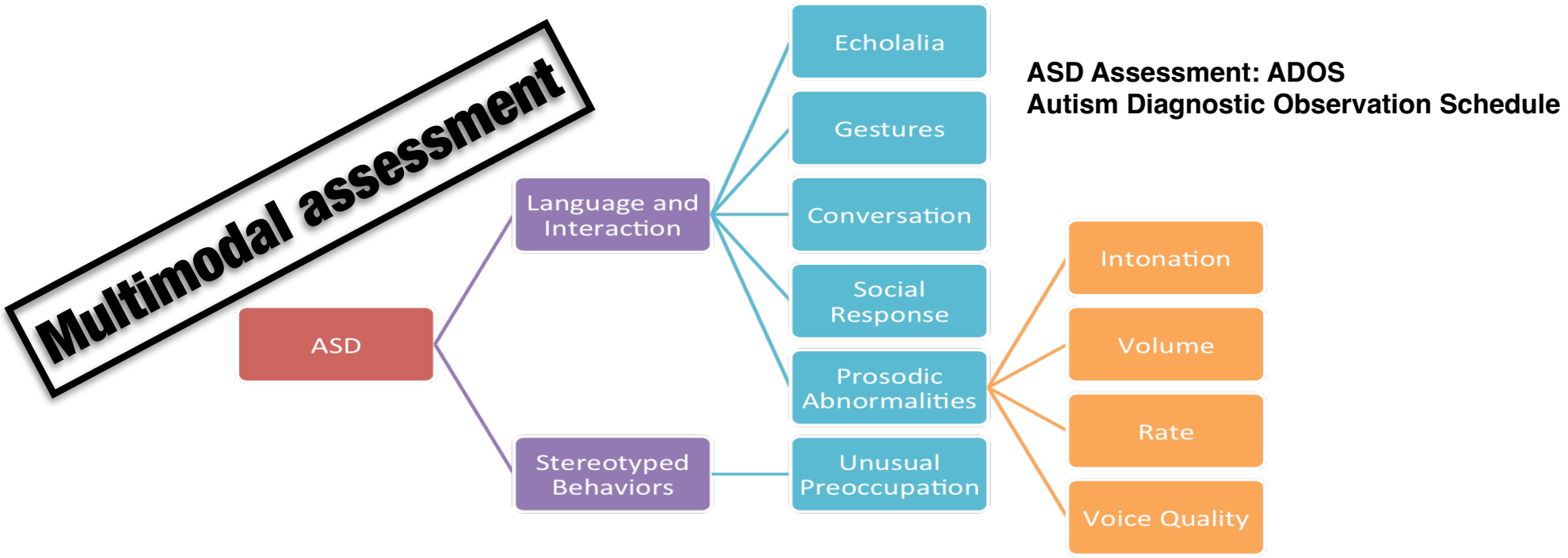
- **1 in 44 US children diagnosed with ASD (CDC[1], 2021)**
  - 1% prevalence in Asia, Europe, North America, 2.6% in S. Korea
- **ASD characterized by**
  - Difficulties in social communication, reciprocity
  - Repetitive or stereotyped behaviors and interests



[1]*CDC:* https://www.cdc.gov/ncbddd/autism/data.html

Opportunities for rich multimodal approaches in Autism Spectrum Disorder (ASD)

- Better understand communication and social patterns of children
- Stratify behavioral phenotyping with quantifiable and adaptable metrics
- Track, quantify children's progress during interventions



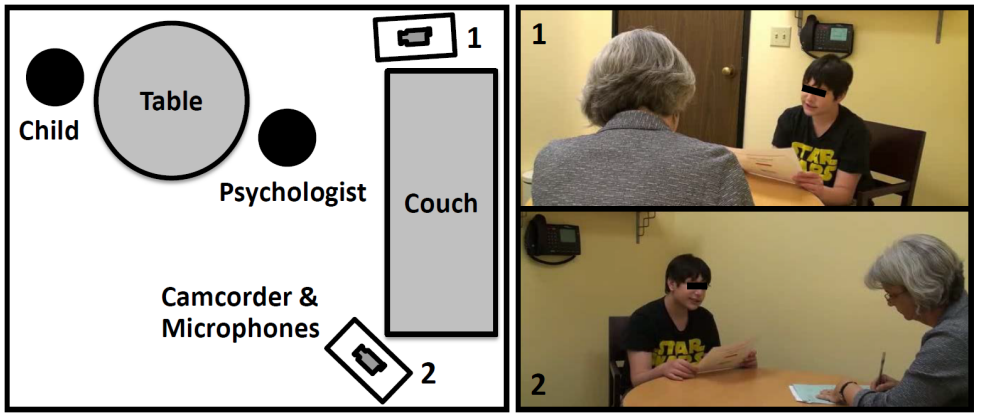
D. Bone, M. Goodwin, M. Black, C-C.Lee, K. Audhkhasi, and S. Narayanan. Applying Machine Learning to Facilitate Autism Diagnostics: Pitfalls and promises. Journal of Autism and Developmental Disorders. 45(5), 1121-1136, 2015

Daniel Bone, Somer Bishop, Matthew P. Black, Matthew S. Goodwin, Catherine Lord, Shrikanth S. Narayanan. Use of Machine Learning to Improve Autism Screening and Diagnostic Instruments: Effectiveness, Efficiency, and Multi-Instrument Fusion. Journal of Child Psychology and Psychiatry. 57(8): 927-937, August 2016

Our Case study Setup: ADOS Interactions

Approach

- **Automatic measures from spontaneous speech**
  - Create generally applicable tools for discovery
- **Data**
- **N=28 children.**
- **ADOS module 3 Interviews**
  - USC CARE Corpus

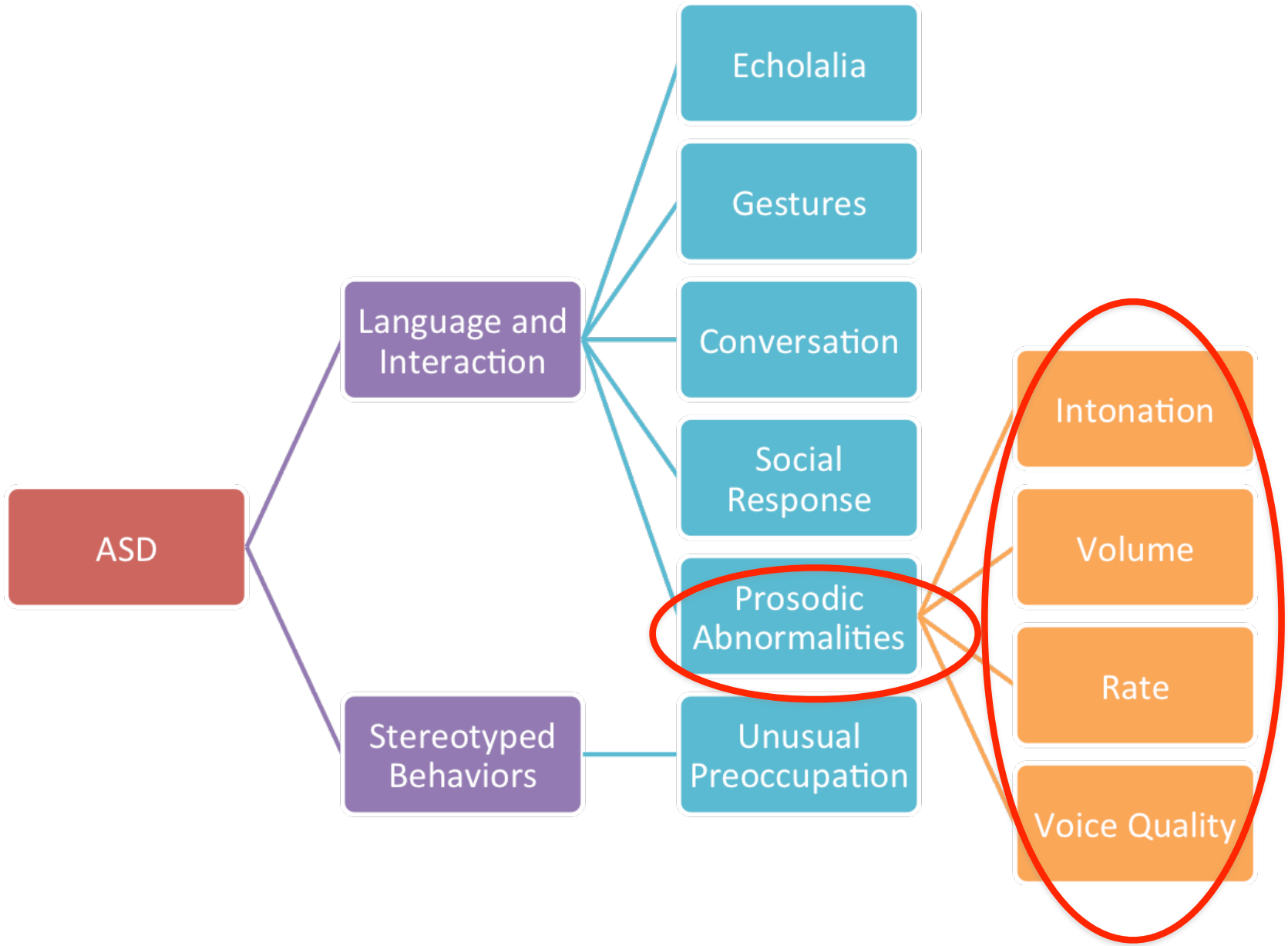


Hypotheses

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)**

MATTHEW P. BLACK, DANIEL BONE, MARIAN E. WILLIAMS, PHILLIP GORRINDO, PAT LEVITT, AND SHRIKANTH NARAYANAN. THE USC CARE CORPUS: CHILD-PSYCHOLOGIST INTERACTIONS OF CHILDREN WITH AUTISM SPECTRUM DISORDERS. PROCEEDINGS OF INTERSPEECH, 2011.

ASD Assessment: focus on atypical prosody



51

Case study: Quantifying Atypical Prosody

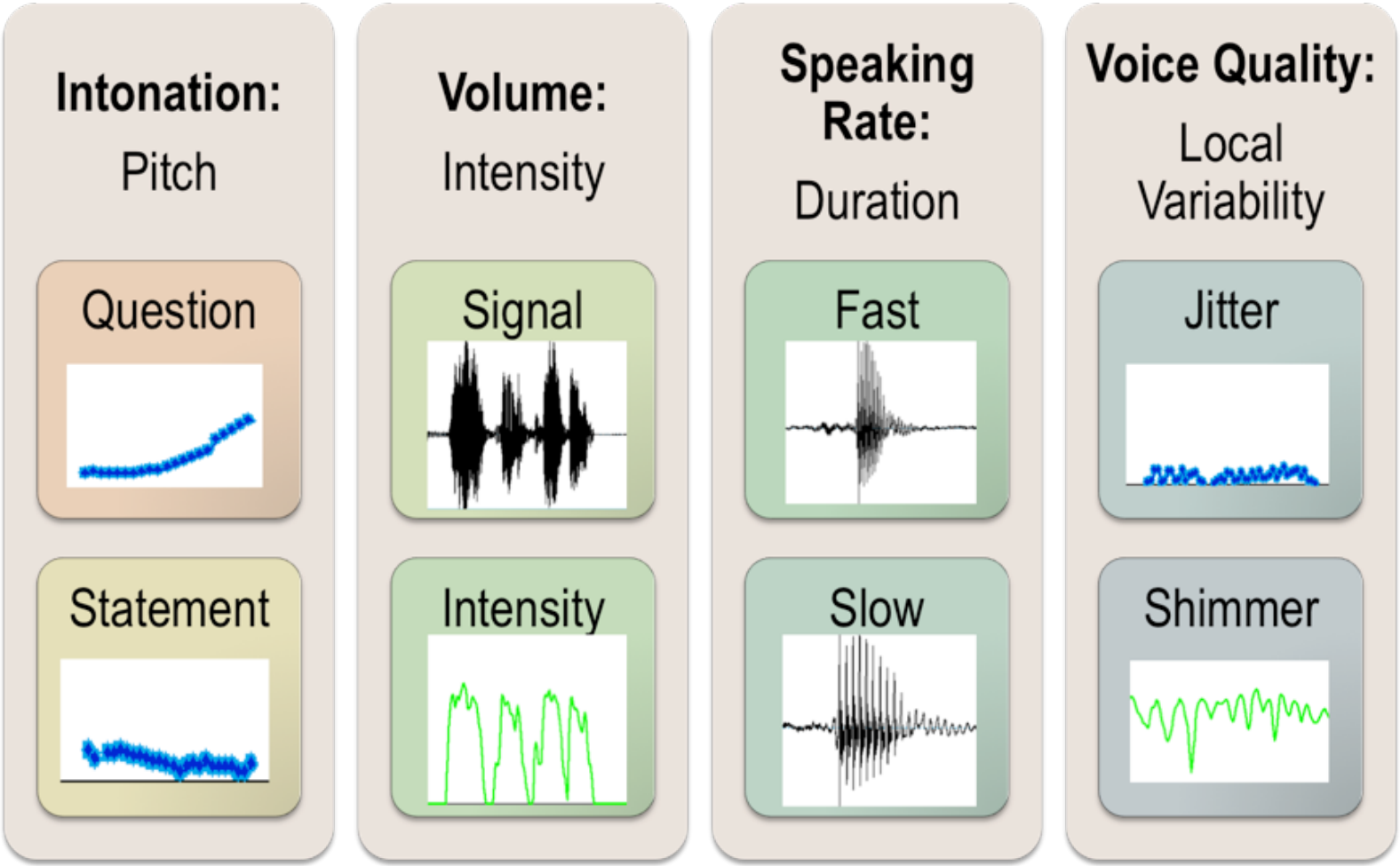
Qualitative descriptions are general and contrasting

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***

Results : Prosodic Features

ADOS MOD 3 Sessions*
N=28

	Trend with ASD Severity	*Psych* Feature	Sp. ρ	Trend with ASD Severity	*Child* Feature	Sp. ρ
Intonation Variability	***More variable***	***Pitch curv.***	***0.33***	***More variable***	***Pitch curv.***	***0.45***
	More variable	Intensity curv.	0.51	More variable	Intensity curv.	0.43
Intonation Dynamics	***More Positive***	***Pitch slope***	***0.32***	***More negative***	***Pitch curv.***	***-0.56***
	More Positive	Intensity curv.	0.31			
Voice Qual.	***Decreased***	***HNR***	***-0.47***			

* Sp. ρ = Spearman's correlation coef.

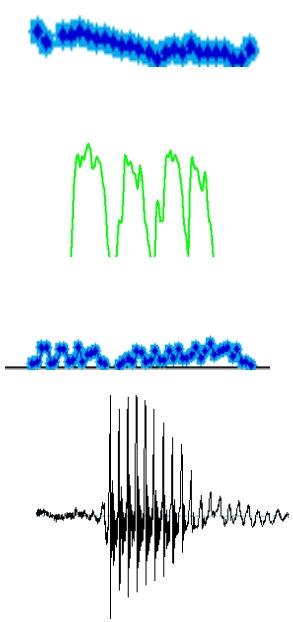
- Psychologist and child speak with more intonational variability
  - ➢Psychologist may be exaggerating intonation in order to engage
  - ➢Child may have less control of intonation dynamics or higher arousal
- Child displays negative pitch curvature -> statement-like?
- Lower periodicity for psychologist → `breathy/harsh' quality?
  - ➢ Relation to other studies: [Boucher et al., 2011] & [Bone et al., 2012]

*MATTHEW P. BLACK, DANIEL BONE, MARIAN E. WILLIAMS, PHILLIP GORRINDO, PAT LEVITT, AND SHRIKANTH NARAYANAN. THE USC CARE CORPUS: CHILD-PSYCHOLOGIST INTERACTIONS OF CHILDREN WITH AUTISM SPECTRUM DISORDERS. PROCEEDINGS OF INTERSPEECH, 2011.

ASD: Quantifying Atypical Prosody & Interaction

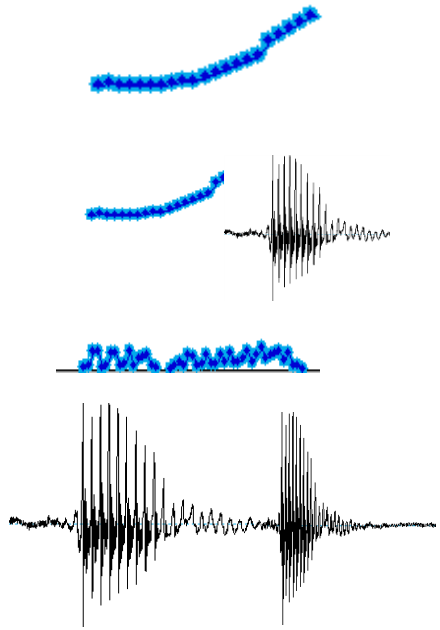
Child's Prosody

- "Monotone"
  p<0.01
- "Abnormal volume"
  p<0.05
- "Breathy/Rough"
  p<0.01
- Slower speaking rate
  p<0.05



Psychologist's Prosody

- Questions/affect
  p<0.05
- Variable prosody
  p<0.01
- also higher jitter
  p<0.01
- slower/then faster
  p<0.01



Spearman's Correlation between rated severity and prosodic cues (ADOS 3, N=28)

Prosody of	Child	Psych.	Child+Psych.
	0.5**	0.71**	0.5**

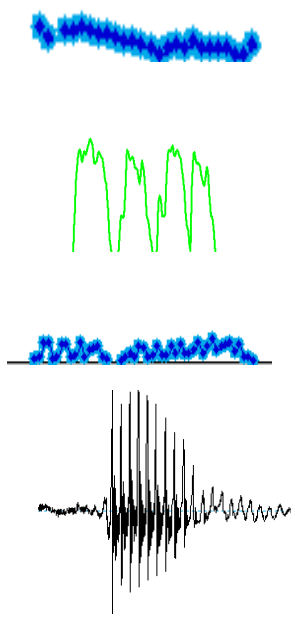
The psychologists may be varying their engagement strategies

DANIEL BONE, CHI-CHUN LEE, MATTHEW P. BLACK, MARIAN E. WILLIAMS, SUNGBOK LEE, PAT LEVITT, AND SHRIKANTH NARAYANAN, "THE PSYCHOLOGIST AS AN INTERLOCUTOR IN AUTISM SPECTRUM DISORDER ASSESSMENT: INSIGHTS FROM A STUDY OF SPONTANEOUS PROSODY", JOURNAL OF SPEECH, LANGUAGE, AND HEARING RESEARCH, 57:1162–1177, AUGUST 2014.

ASD: Quantifying Atypical Prosody & Interaction

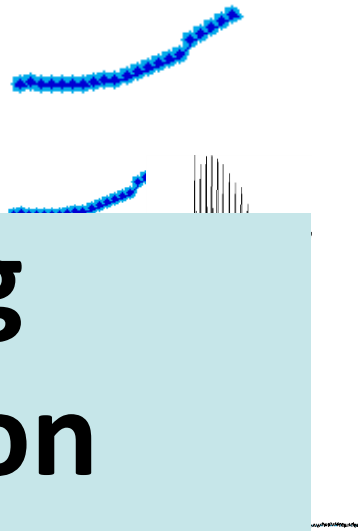
Child's Prosody

- "Monotone"
  p<0.01
- "Abnormal volume"
  p<0.05
- "Breathy/Rough"
  p<0.01
- Slower speaking rate
  p<0.05



Psychologist's Prosody

- Questions/affect
  p<0.05
- Variable prosody
  p<0.01
- also higher jitter
- s



Modeling Interaction Dynamics is Critical

Spearman's Correlation between rated seve =28)

Prosody of	Child	Psyc	
	0.5**	0.71**	0.5**

The psychologists may be varying their engagement strategies

DANIEL BONE, CHI-CHUN LEE, MATTHEW P. BLACK, MARIAN E. WILLIAMS, SUNGBOK LEE, PAT LEVITT, AND SHRIKANTH NARAYANAN, "THE PSYCHOLOGIST AS AN INTERLOCUTOR IN AUTISM SPECTRUM DISORDER ASSESSMENT: INSIGHTS FROM A STUDY OF SPONTANEOUS PROSODY", JOURNAL OF SPEECH, LANGUAGE, AND HEARING RESEARCH, 57:1162–1177, AUGUST 2014.

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).

DANIEL BONE, CHI-CHUN LEE, THEODORA CHASPARI, MATTHEW P. BLACK, MARIAN E. WILLIAMS, SUNGBOK LEE, PAT LEVITT AND SHRIKANTH NARAYANAN, ACOUSTIC-PROSODIC, TURN-TAKING, AND LANGUAGE CUES IN CHILD-PSYCHOLOGIST INTERACTIONS FOR VARYING SOCIAL DEMAND, INTERSPEECH, 2013.

Results: Language & Turn taking

	Trend w/ Sev.	Psych Feature	Sp. ρ	Trend w/ Sev.	Child Feature	Sp. ρ
Speech Amount	***Increased***	***Speech %***	***0.54***	***Decreased***	***Speech %***	***-0.36***
				Decreased	WPS (MLU)	-0.42
Turn-taking	***Increased***	***Articulatory R***	***0.38***	***Decreased***	***Articulatory R***	***-0.34***
	Increased	Intra-turn sil.	0.32	Increased	Latency	0.34
Pronouns	***Increased***	***Personal Pron.***	***0.38***	***Decreased***	***Personal Pron.***	***-0.40***
Language Use	Decreased	Assent Lang.	-0.48	Decreased	Affect Lang.	-0.40
	Decreased	***Non-fluencies***	***-0.48***	***Decreased***	***Fillers***	***-0.41***

- **Interaction patterns influence one another'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
- Overall, the conversational quality degrades
  - ➢ 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

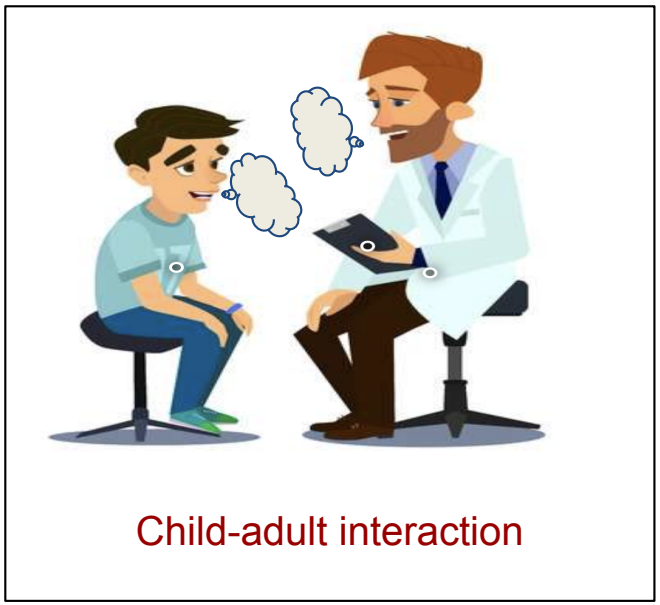
Future Work

- Investigate interplay between these varied features
- Larger datasets that include TD and non-ASD DD
- Unsupervised behavioral signals e.g., arousal dynamics, entrainment

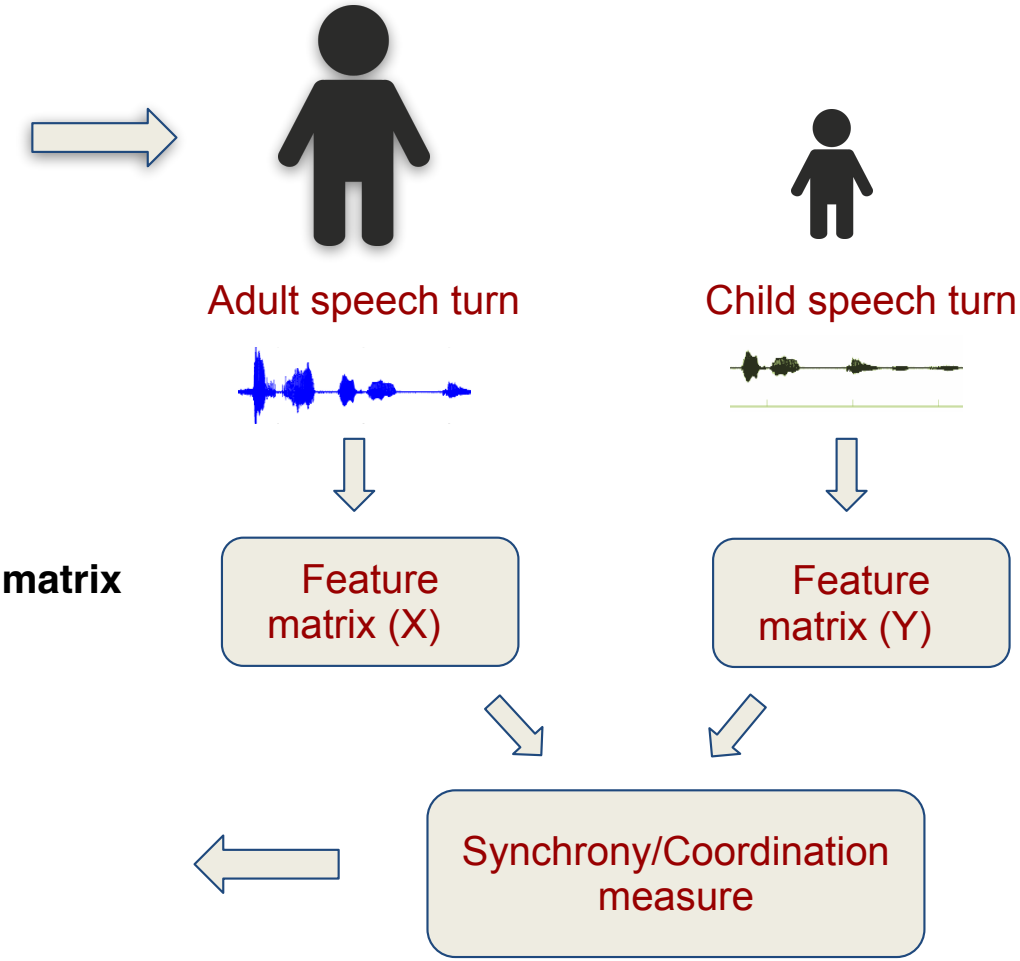
DANIEL BONE, CHI-CHUN LEE, THEODORA CHASPARI, MATTHEW P. BLACK, MARIAN E. WILLIAMS, SUNGBOK LEE, PAT LEVITT AND SHRIKANTH NARAYANAN, ACOUSTIC-PROSODIC, TURN-TAKING, AND LANGUAGE CUES IN CHILD-PSYCHOLOGIST INTERACTIONS FOR VARYING SOCIAL DEMAND, INTERSPEECH, 2013.

YOUNG KYUNG KIM, RIMITA LAHIRI, MD NASIR, SO HYUN KIM, SOMER BISHOP, CATHERINE LORD AND SHRIKANTH NARAYANAN. ANALYZING SHORT TERM DYNAMIC SPEECH FEATURES FOR UNDERSTANDING BEHAVIORAL TRAITS OF CHILDREN WITH AUTISM SPECTRUM DISORDER. PROCEEDINGS OF INTERSPEECH, BRNO, CZECH REPUBLIC, 2021

Quantifying synchrony



For every consecutive turn pair, coordination measures are computed, and they are averaged across all such turn pairs to get a session level measure



Acoustic: **Prosodic/Spectral Feature matrix**
***Lexical*: BERT embeddings matrix**

Acoustic: **Squared cosine distance of complexity measure;**
Dynamic Time Warping distance
***Lexical*: Word Movers Distance**

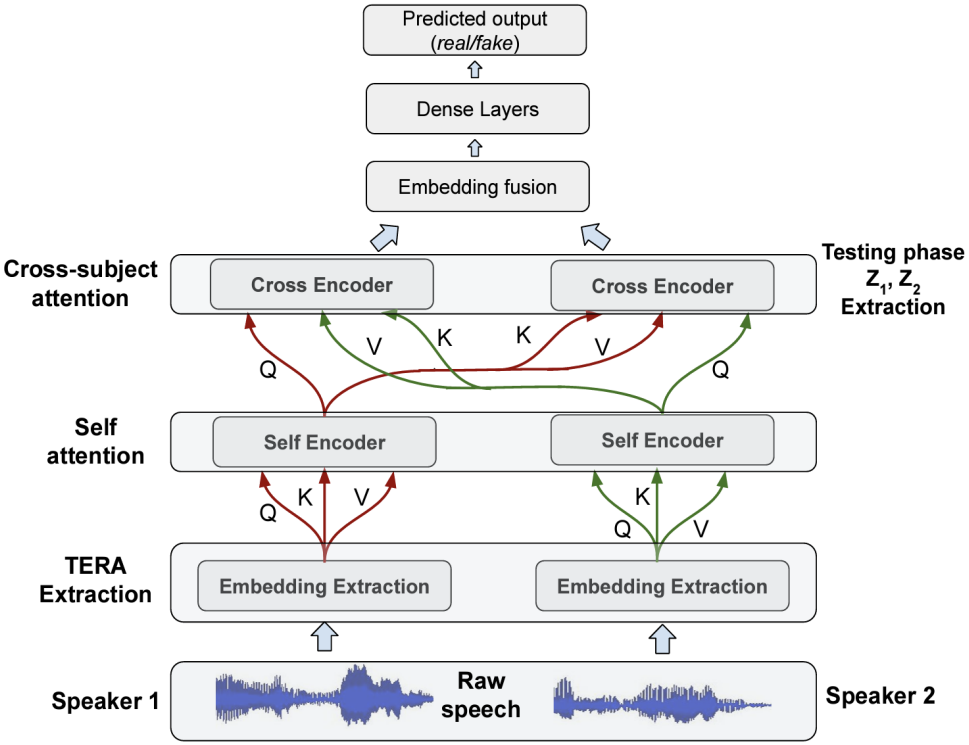
Quantifying synchrony

Child adult interaction

- Children with ASD diagnosis show less synchrony in both social and emotional subtasks in terms of the absolute values of the introduced measures
- Measures are ***complementary***: improved distinction between ASD and non-ASD groups when vocal and lexical synchrony measures are fused: 40% relative improvement in F1 score
- An improvement in the fused F1 classification score is observed with increasing age in female children:
  - supports further investigation into why female children are more likely to go undiagnosed than males until an older age

Rimita Lahiri, Md Nasir, Manoj Kumar, SoHyun Kim, Somer Bishop, Cathy Lord, Shrikanth Narayanan. Interpersonal synchrony across vocal and lexical modalities in interactions involving children with Autism Spectrum Disorder. J. Acoust. Soc. Am. Express Letters 9(2): 095202, 2022

Improved Methods Quantifying synchrony: Contextual Entrainment Distance



Model Architecture for CED extraction

CED is directional:
PC (psychologist->child) & CP (child->psychologist)

Clinical scores	Pearson's correlation			
	CED-PC		CED-CP	
	ρ	p-value	ρ	p-value
VINELAND ABC	-0.061	0.237	0.012	0.827
VINELAND Social	-0.021	0.345	0.071	0.073
VINELAND Communication	**-0.158**	**0.003**	0.043	0.428
CSS	**0.222**	**0.004**	0.023	0.672
CSS-SA	**0.231**	**0.012**	0.03	0.472
CSS-RRB	0.158	0.055	0.091	0.262

Correlation between CED and clinical scores relevant to ASD

Rimita Lahiri, Md Nasir, Cathy Lord, SoHyun Kim, Shrikanth Narayanan. A Context-Aware Computational Approach for Measuring Vocal Entrainment in Dyadic Conversations. Proc. ICASSP, 2023 (to appear)

ASD: Understanding the expression of social cues

Production of Affective Facial Expressions (During Smile Imitation Task)

Computational Targets

Quantify atypicality of smile

Region-based activation

Synchrony & symmetry

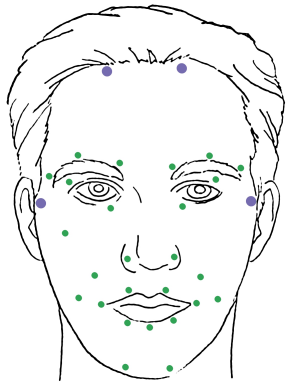
Reduced complexity in dynamic facial behavior primarily in the eye region

- Complexity measured in terms of *Multiscale Sample Entropy (MSE)* [Costa et al. 2011]
- MoCAP data from 20 HFA, 19 TD children, 8 - 12 years of age, no group difference in IQ, age or gender

Tanaya Guha, Zhaojun Yang, Ruth Grossman and Shrikanth Narayanan. A Computational Study of Expressive Facial Dynamics in Children with Autism. IEEE Transactions on Affective Computing. 9(1): 14-20, January 2018

Emily Zane, Zhaojun Yang, Lucia Pozzan, Tanaya Guha, Shrikanth Narayanan, Ruth Grossman. Motion-Capture Patterns of Voluntarily Mimicked Dynamic Facial Expressions in Children and Adolescents With and Without ASD. Journal of Autism and Developmental Disorders. 49(3): 1062-1079, March 2019

Social communication difficulties in autism involve deficits in cross-modal coordination



Objective

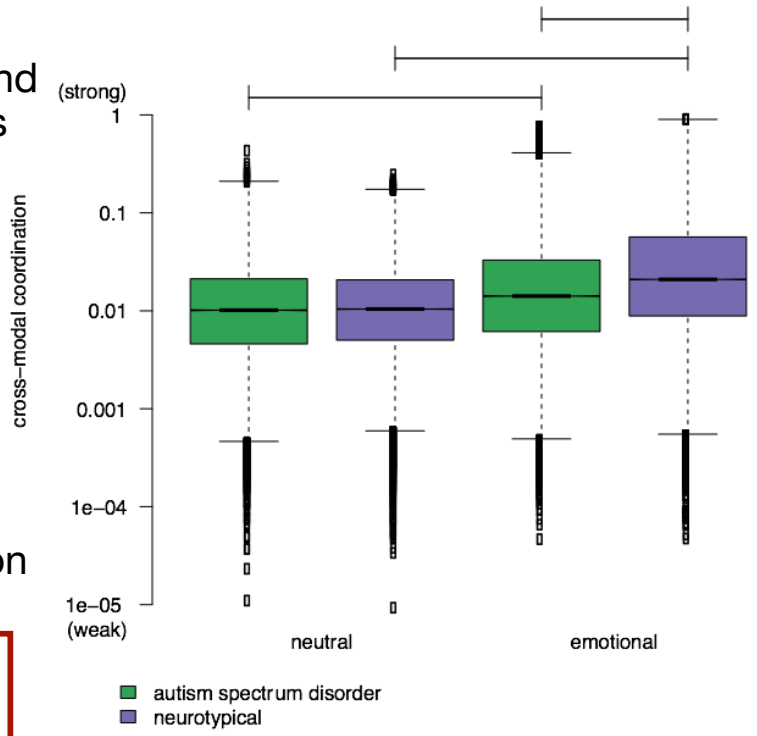
- Dynamic relation between *speech production and facial expression* in children with autism?
- How face-directed gaze modulates this cross-modal coordination?

Method

- Mimicry task in which participants watched and repeated neutral and emotional spoken sentences with accompanying facial expressions
- Cross-modal coordination measure: Granger causality analysis of dependence between audio and motion capture signals

Results

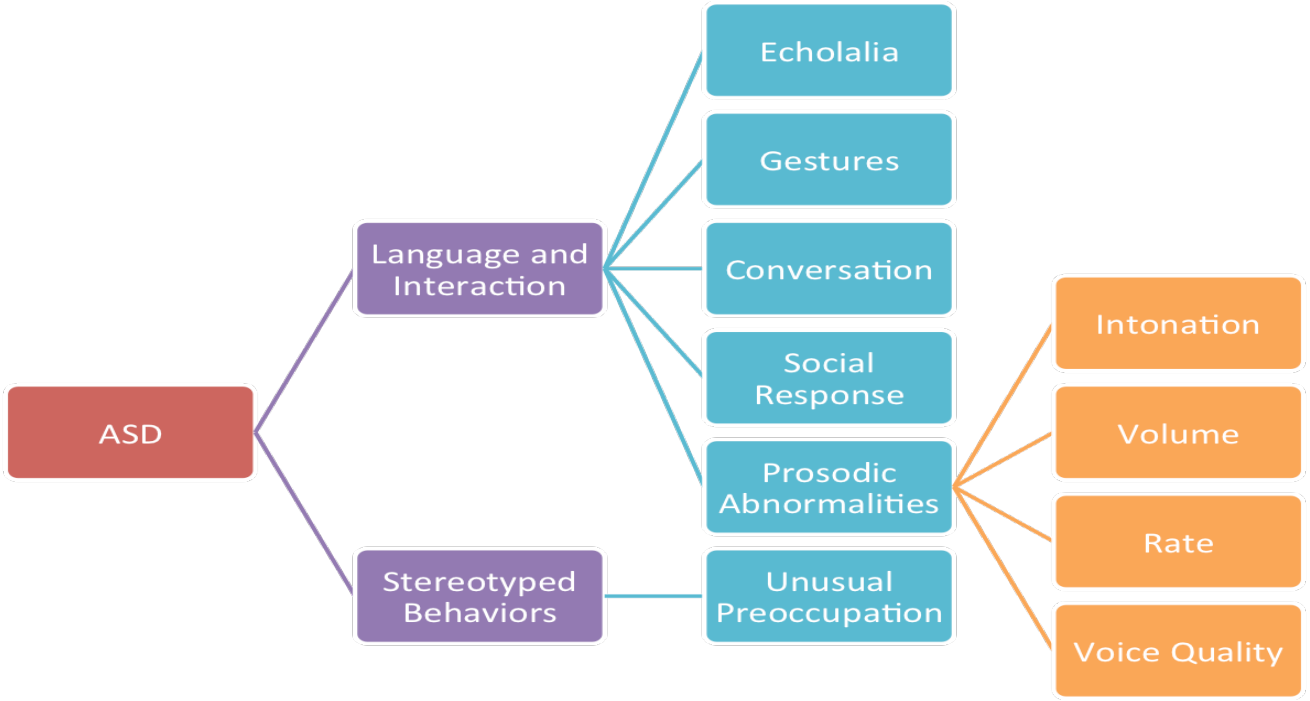
- Neurotypical children produced emotional sentences with strong cross-modal coordination and produced neutral sentences with weak cross-modal coordination (*differential expressions*)
- Autistic children produced similar levels of cross-modal coordination for both neutral and emotional sentences (*no differentiation*)
- *Cross-modal coordination was greater when the non-ASD child spent more time looking at the face, but weaker when the autistic child spent more time looking at the face*



Tanner Sorensen, Emily Zane, Tiantian Feng, Shrikanth Narayanan, and Ruth Grossman. Cross-Modal Coordination of Face-Directed Gaze and Emotional Speech Production in School-Aged Children and Adolescents with ASD. Scientific Reports (Nature Press). 9, 18301, 2019

Opportunities for rich multimodal learning approaches

- Better understand communication and social patterns of children
- Stratify behavioral phenotyping with quantifiable and adaptable metrics
- Track, quantify children's progress during interventions

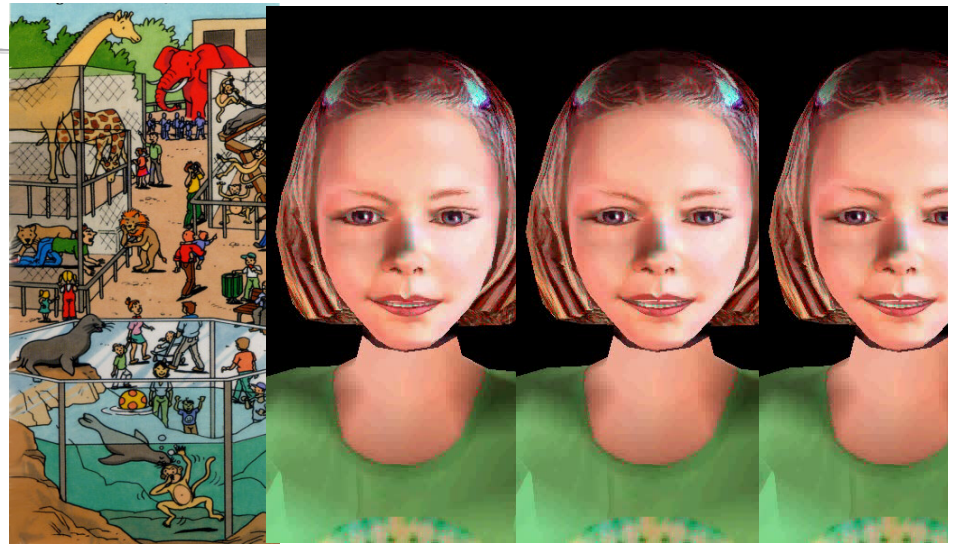


D. Bone, M. Goodwin, M. Black, C-C.Lee, K. Audhkhasi, and S. Narayanan. Applying Machine Learning to Facilitate Autism Diagnostics: Pitfalls and promises. Journal of Autism and Developmental Disorders. 45(5), 1121-1136, 2015

D. Bone, S. Bishop, M. Black, M. Goodwin, C. Lord, S. Narayanan. Use of Machine Learning to Improve Autism Screening and Diagnostic Instruments: Effectiveness, Efficiency, and Multi-Instrument Fusion. Journal of Child Psychology and Psychiatry. 57(8): 927-937, August 2016

Children with Autism interacting with Conversational Agents

from ~ 10 years ago



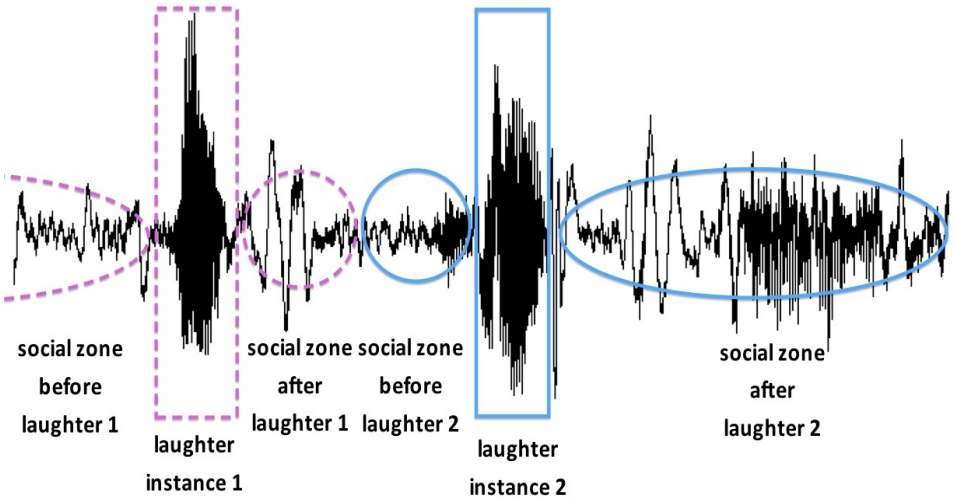
- **Automatic data logging**
  - Agent behavior
  - Wizard flag

- **Recorded data: Clinic**
  - Three Sony Handycams
  - Two shotgun microphones

EMILY MOWER, CHI-CHUN LEE, JAMES GIBSON, THEODORA CHASPARI, MARIAN WILLIAMS, SHRIKANTH NARAYANAN. ANALYZING THE NATURE OF ECA INTERACTIONS IN CHILDREN WITH AUTISM. IN PROCEEDINGS OF INTERSPEECH, 2011.
EMILY MOWER, MATTHEW BLACK, ELISA FLORES, MARIAN WILLIAMS AND SHRIKANTH NARAYANAN. DESIGN OF AN EMOTIONALLY TARGETED INTERACTIVE AGENT FOR CHILDREN WITH AUTISM. IN PROCEEDINGS OF IEEE INTERNATIONAL CONFERENCE ON MULTIMEDIA & EXPO (ICME), 2011

Modeling "Shared Enjoyment"

- **Rates social interaction in autism**
- **Laughter: expression of shared enjoyment**
  - Voiced laughter is associated with positive affect
  - Unvoiced laughter acts as a social facilitator



Task 1: Classification of the type of laughter based on acoustic information extracted from the laughter region: 85-90% accuracy

Task 2: Prediction of the type of laughter from the acoustic features of the surrounding social zones: 90-95% accuracy

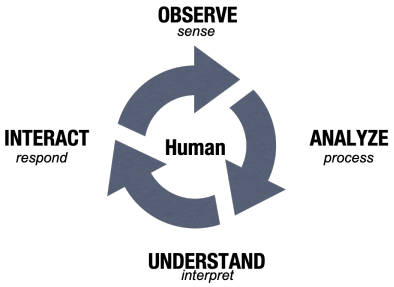
- Acoustic features of social zones discriminate type of shared enjoyment
- Humorous events (voiced instances) easier to recognize than social facilitators (unvoiced instances)

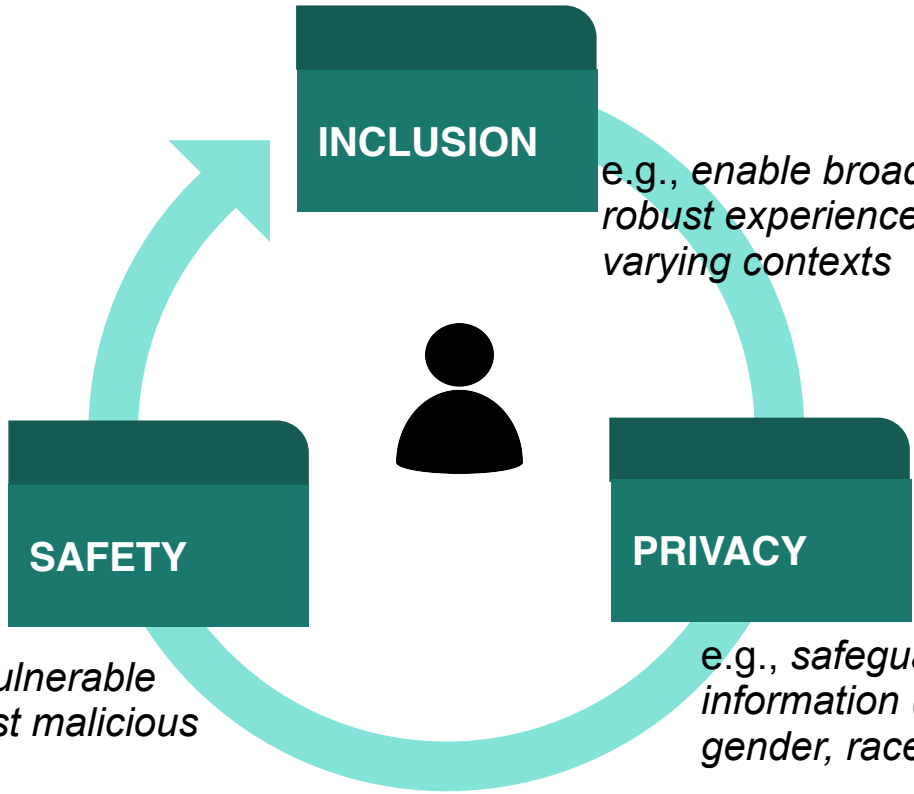
Theodora Chaspari, Emily Mower Provost, Athanasios Katsamanis and Shrikanth Narayanan, An Acoustic Analysis Of Shared Enjoyment In ECA Interactions Of Children With Autism, in: Proceedings of ICASSP, 2012

New possibilities for child-inclusive multimodal conversational interactions

Elements toward enabling

Trustworthy Human-centered Machine Intelligence





e.g., *enable broad access, robust experiences across varying contexts*

others
explainability
transparency
resilience
…

e.g., *protect vulnerable people; against malicious attacks*

e.g., *safeguard protected information (PII: age, gender, race, etc.)*

69

Twin goals: Understanding and addressing variability within and across people and their contexts

S. Narayanan and A. Madni. Inclusive Human centered Machine Intelligence. *The Bridge*. 50(S): 113-116. National Academy of Engineering, 2020

69

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,
SIMONS FOUNDATION, GUGGENHEIM, GOOGLE, APPLE, AMAZON, DISNEY, TOYOTA

Signal Analysis and Interpretation Laboratory

....technologies to understand the human condition
and to support and enhance human capabilities and experiences

creating inclusive technologies and technologies for inclusion

Signal Analysis and Interpretation Laboratory

....technologies to understand the human condition
and to support and enhance human capabilities and experiences

creating inclusive technologies and technologies for inclusion

http://sail.usc.edu

Shri Narayanan, Director