USC Viterbi
School of Engineering

Human Behavioral Informatics

TECHNICAL & SOCIETAL

Challenges and Opportunities for Multimedia R&D

Shrikanth (Shri) Narayanan

Signal Analysis and Interpretation Laboratory (SAIL)

http://sail.usc.edu/shri

University of Southern California

ICME JULY 2011

1

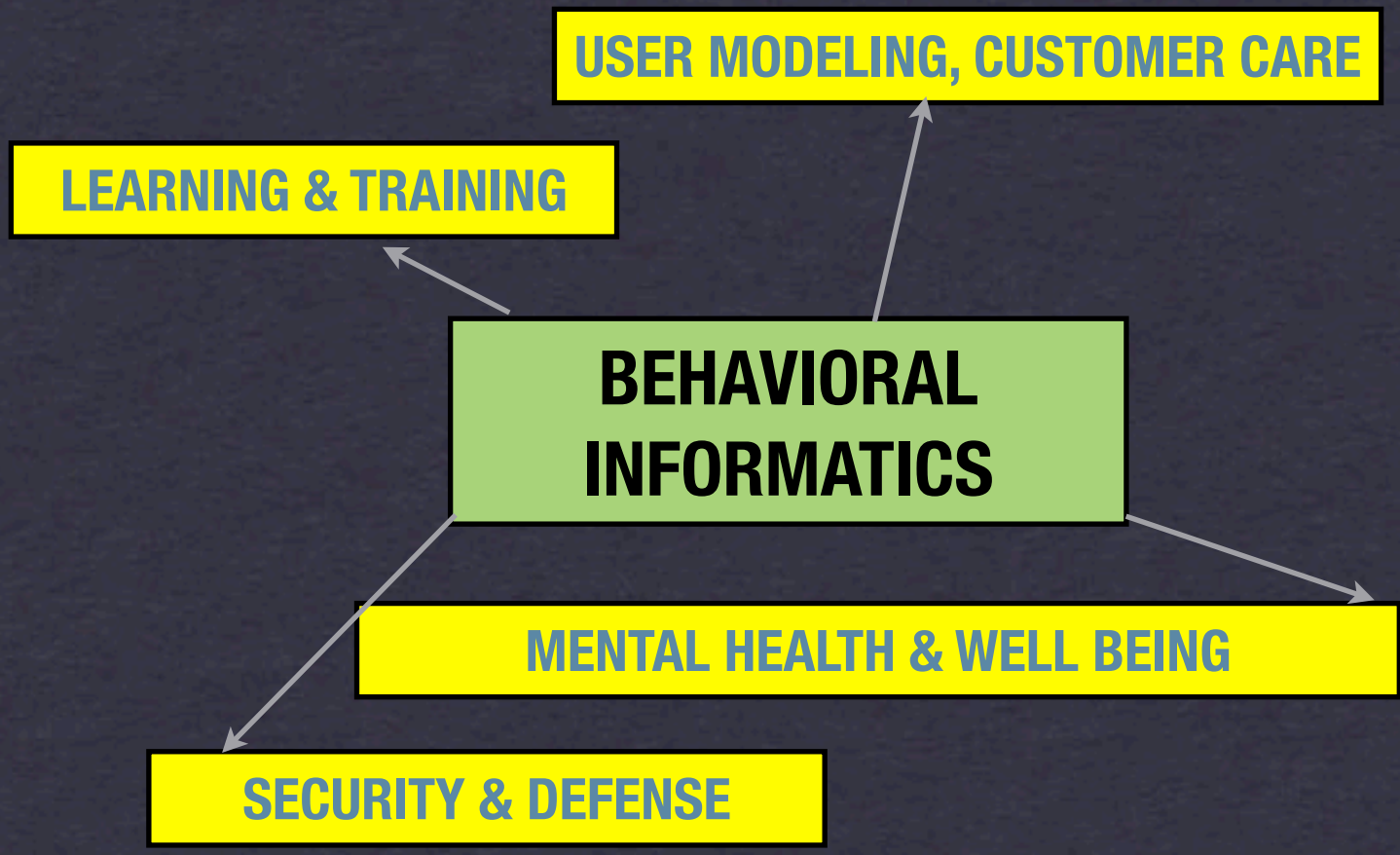
Human Behavior

Complex and multifaceted: involves

- Intricate brain-body interplay
- Effect of environment and interaction with others
- Communication, Social interaction, Emotions, Personality
- Typical, Atypical, and Disordered characterizations
- Generation and processing of multimodal cues

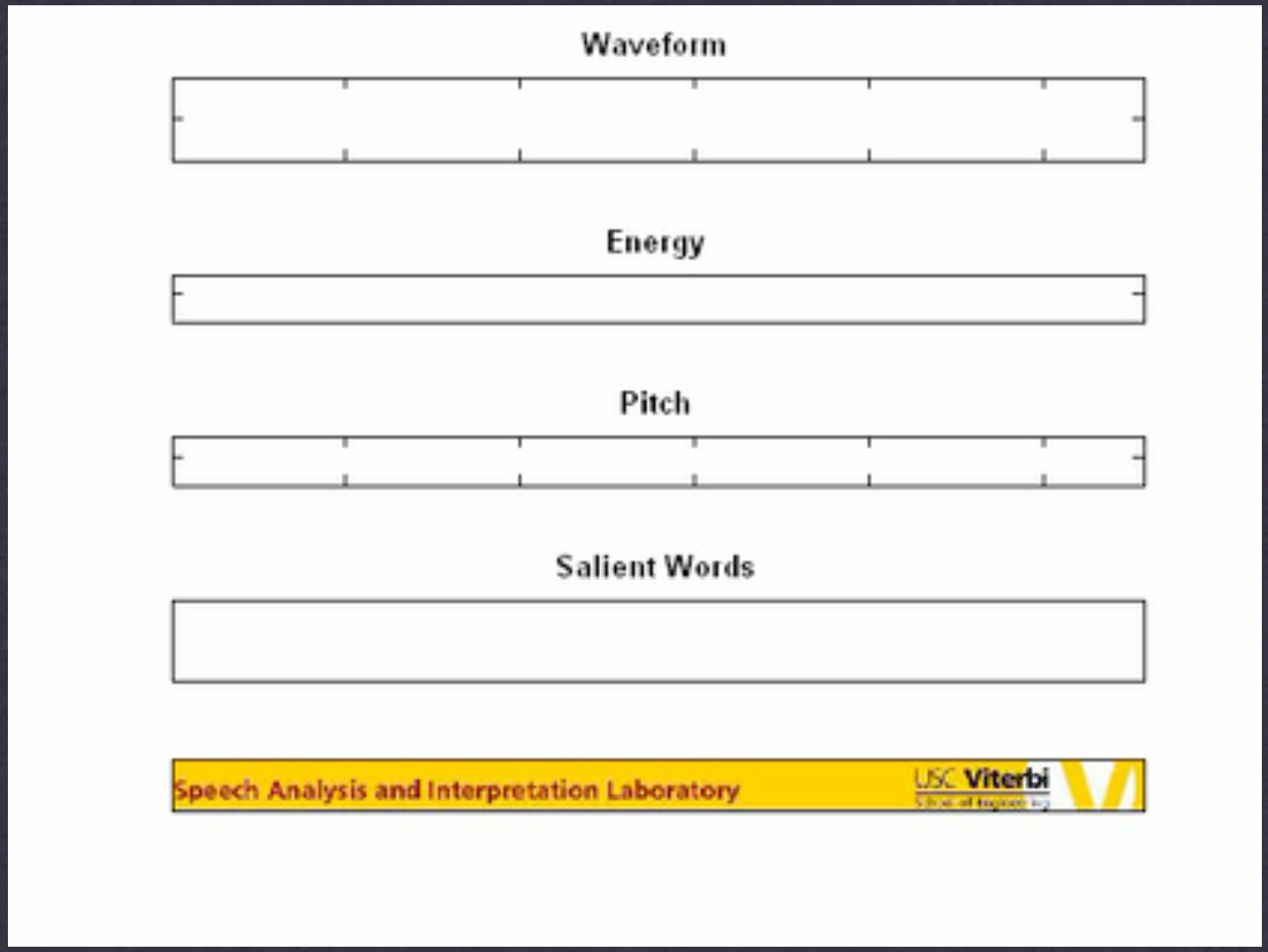
ENGINEERS, WHY BOTHER?

EXAMPLE APPLICATIONS



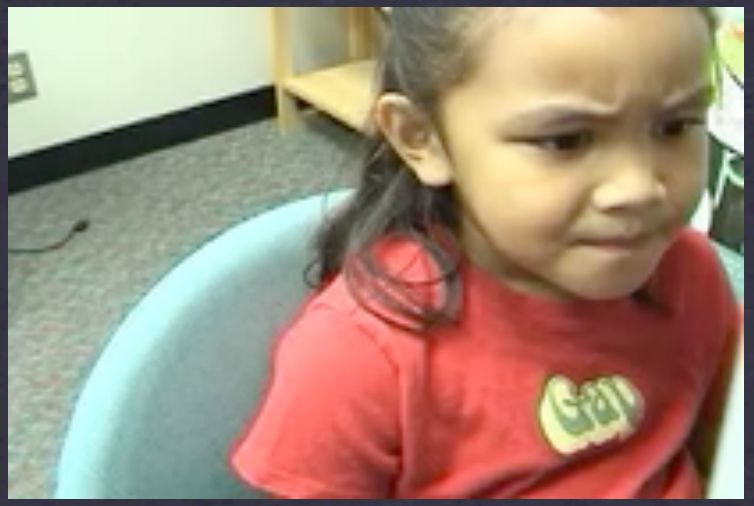
BEHAVIOR ANALYSIS IS CENTRAL
BUT LARGELY MANUAL!

Customer care
Is the customer on the telephone upset? (only customer side played)



Problem solving Game: "Cognitive state" Characterization

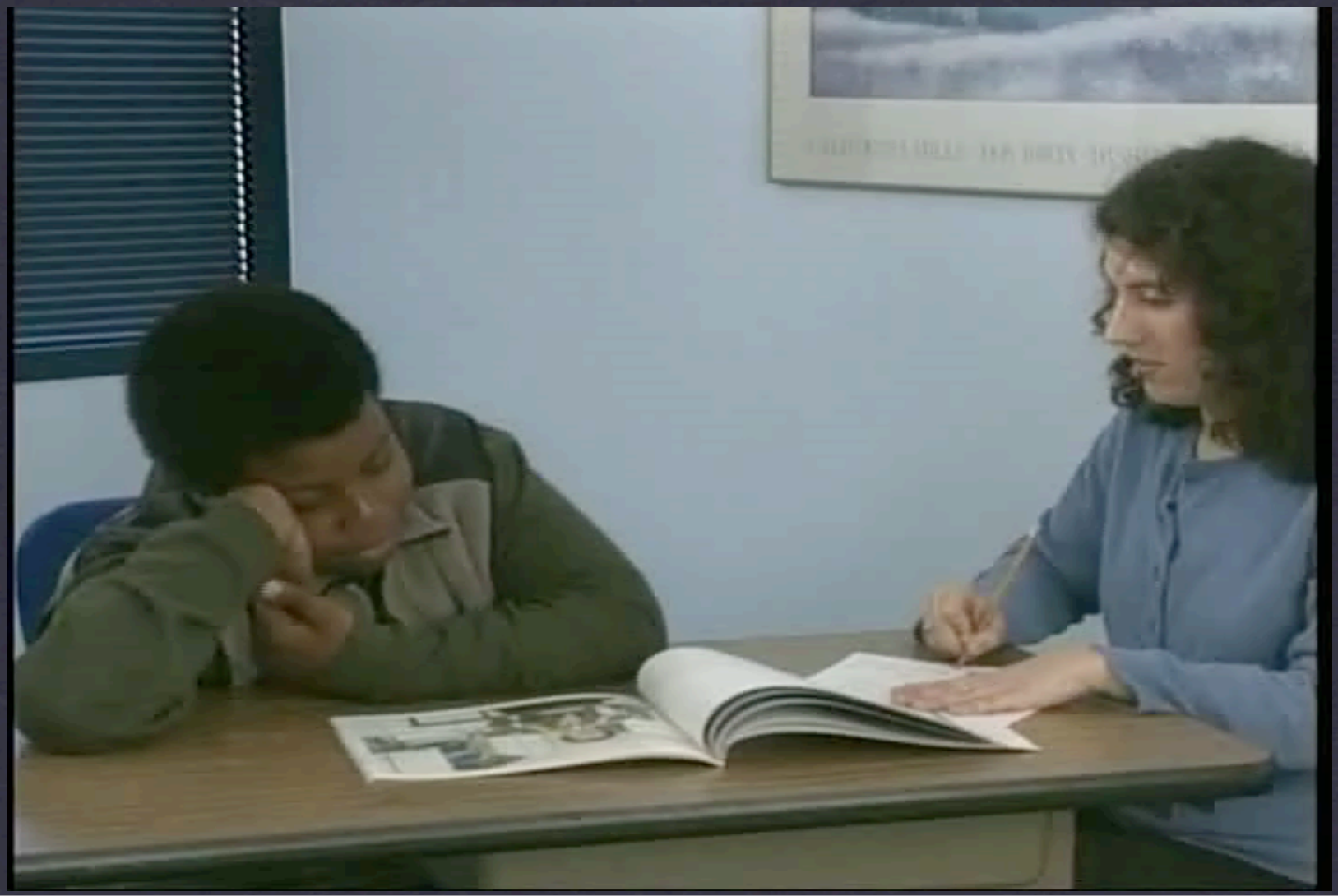
- **CONFIDENT** VS. **UNCERTAIN**



Uncertainty manifests itself through combination of vocal, language, and visual cues

Couple therapy
Characterizing affective dynamics, blame patterns

Autism Spectrum Disorders
Characterizing joint attention; quantifying socio-emotional discourse



Multimodal Behavior Signals

- Some overt and observable e.g., vocal and facial expressions, body posture
- Others covert e.g., heart rate, brain activity
  - ✓ Provide a window into higher level processes
  - ✓ Information at multiple time scales, through multiple cues
  - ✓ Implications for understanding/using “brain-body” relations

MEASURING & QUANTIFYING HUMAN BEHAVIOR:
A CHALLENGING ENGINEERING PROBLEM

Operationally defining Behavior Signal Processing (BSP)

COMPUTATIONAL METHODS THAT MODEL HUMAN BEHAVIOR SIGNALS

- MANIFESTED IN BOTH OVERT AND COVERT CUES
- PROCESSED AND USED BY HUMANS EXPLICITLY OR IMPLICITLY
- THAT SUPPORT HUMAN ANALYSIS AND DECISION MAKING

OUTCOME OF BSP: "BEHAVIORAL INFORMATICS"

How technology has helped already?

- **Significant advances in foundational aspects of behavior modeling**
  - Speech recognition: what was spoken
  - Audio (video) diarization: who spoke when; doing what,..
  - Activity recognition: head pose; face/hand gestures,...
  - Physiological Affective signal processing with EKG, GSR, ..

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

Ongoing Engineering Pursuits: Multifaceted

- **Rich speech and spoken language processing**
  - ***who* said *what* to *whom* and *how***
- **Affective computing & Emotion recognition**
  - Modeling affective behavior in acted and natural scenarios
- **Social signal processing**
  - Modeling social behavior: turn taking, non verbal cues such as smiles, laughters and sighs, rapport, ..
- **Sensing: From Smartrooms to Body area networks**

ALL THESE CAN BE VIEWED AS PART OF BSP UMBRELLA

–

Behavioral Signal Processing: Ingredients

- **Acquisition**
  - Behavior data sensing: audio, video, physiological, location,..
  - Measurements in controlled and free living settings
    - instrumented environs & instrumented people (mobile sensing)
    - use of virtual interaction paradigms and techniques
- **Analysis**
  - Deriving low level cues: who, what, when, how, where, why
- **Modeling**
  - High level descriptions desired by domain experts
  - Descriptive and predictive models using multimodal data

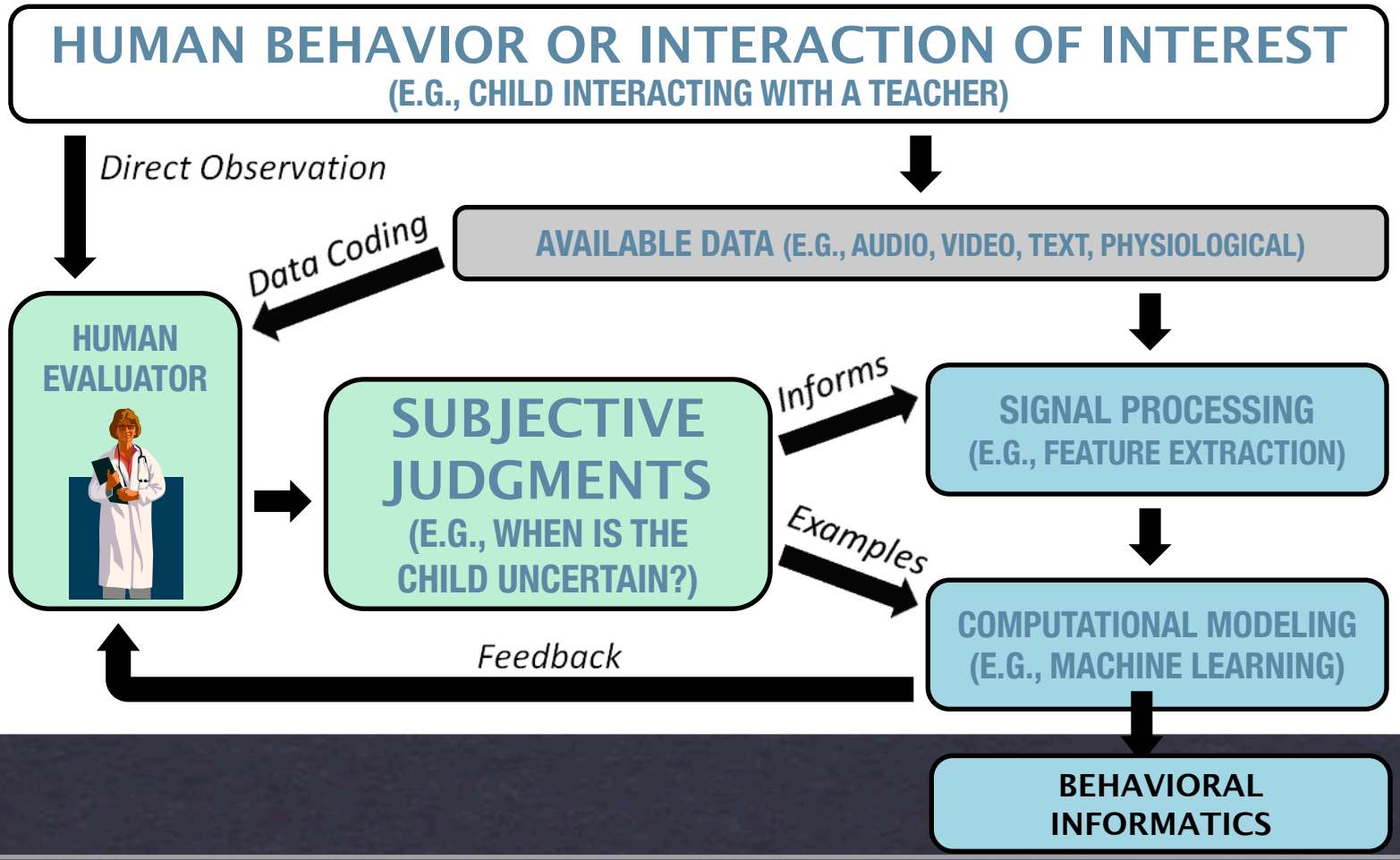
- **Handle varying types of abstraction in data and descriptions**
  - Uncertainty in observations (partial, noisy)
  - Subjectivity in descriptions (especially of higher level behavior)
  - Heterogeneity and variability in how data is generated and used

Some Methodology Themes

- **Many domain-specific features**
  - *Language assessment*: fluency of read speech
  - *Obesity*: physical activity
  - *Couples Therapy*: emotional dynamics
- **Many common features across domains**
  - e.g., spoken and body language use, interaction patterns
- **Overall approach is the same!**
  - Analyze the signals that the experts observe
  - Learn from, and augment their capabilities
  - Support but *not* supplant!

Behavior Coding: Humans in the loop

- Modeling subjective judgments on human behavior



REST OF THIS TALK

- Some BSP building blocks

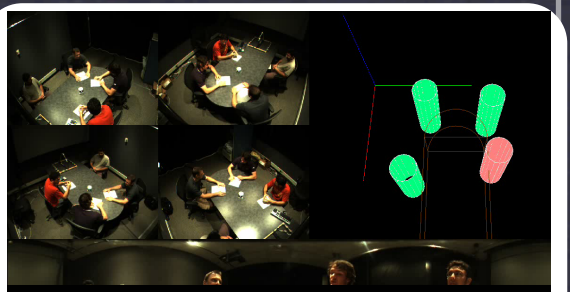
Example Behavioral Analysis Studies

- Family Studies: Marital couples
  - Blame patterns; positiveness/negativeness; humor/sarcasm
- Autism Spectrum Disorders
  - Characterizing and quantifying socio-emotional discourse
  - Technology interfaces for elicitation & personalized interventions
- Metabolic Health Monitoring
  - Characterizing physical behavior in context

BSP BUILDING BLOCKS

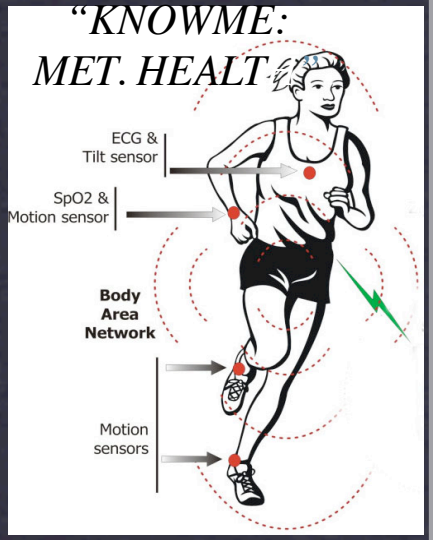
Multimodal Signal Acquisition & Processing

- **Instrumented Environments**: arrays of microphones, cameras, mocap
  - Sense the user
    - Identity/Location of speaker
    - Recognition of speech, visual activity
    - Emotion recognition
  - interaction
    - Back channels, head orientation, proxemics
    - Engagement ....
  - and the audio-visual environment:
    - Lab, classroom, home,playground,..



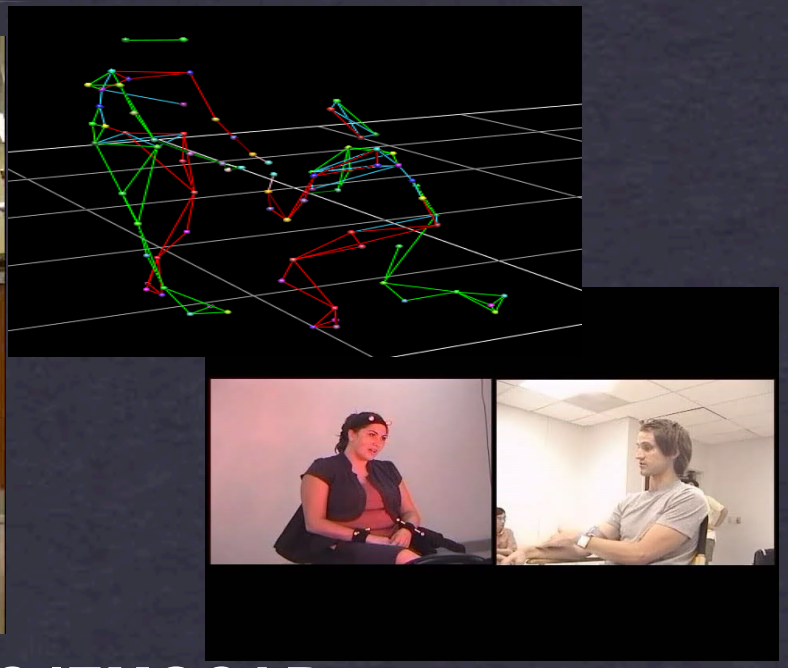
"SMARTROOM" 2002

"DYADIC MOCAP" 2005



- **Instrumented People**: Body sensing, mobile settings
  - Sense user state, activity, context
  - Wireless (cell phone) based: sensing & actuation,
  - Data from real life free living conditions

EXPRESSIVE BEHAVIOR: USING ACTORS



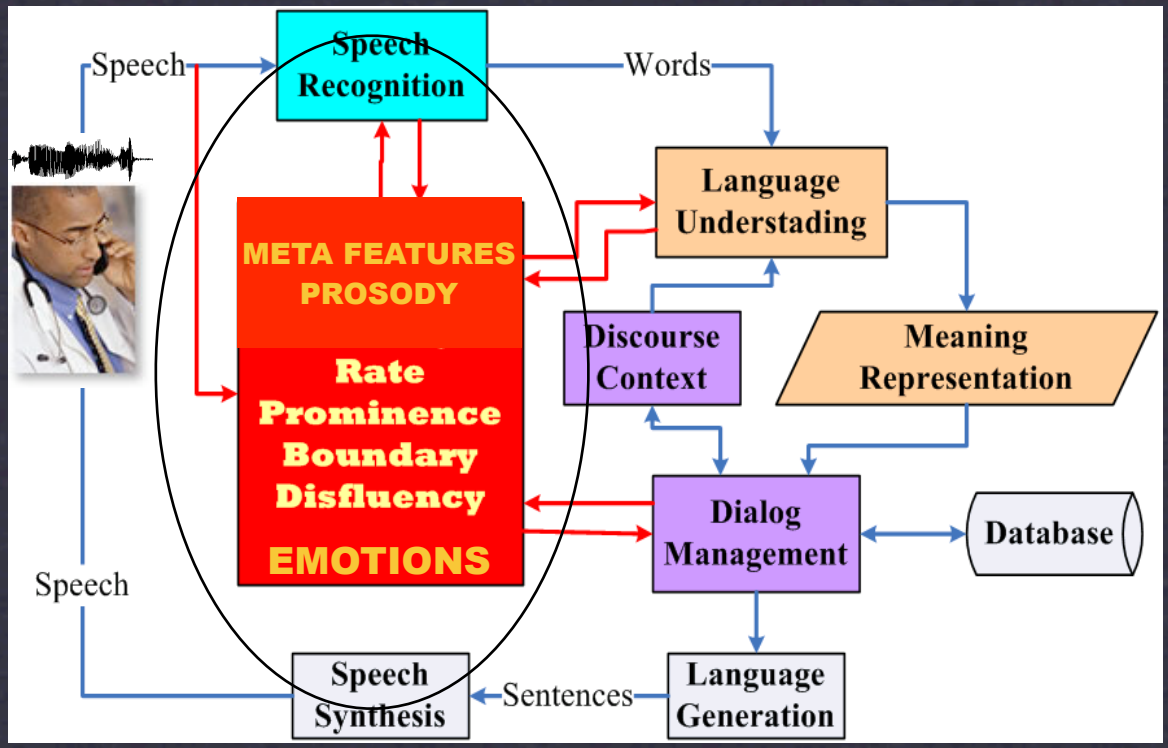
USC CreativeIT database

- **Multimodal/multidisciplinary** database
- USC Engineering and Theatre
- Dyadic Theatrical Improvisations
- Motion Capture, Video, Audio
- http://sail.usc.edu/improv/

USC IEMOCAP: Interactive and emotional motion capture database

- Dyadic interaction
- 5 sessions, 2 actors each
- Emotions elicited in context
- ~12 hours of data
- Markers on the face and hands
- http://sail.usc.edu/iemocap/

Integrated Spoken Language Processing
mapping speech to words, and beyond



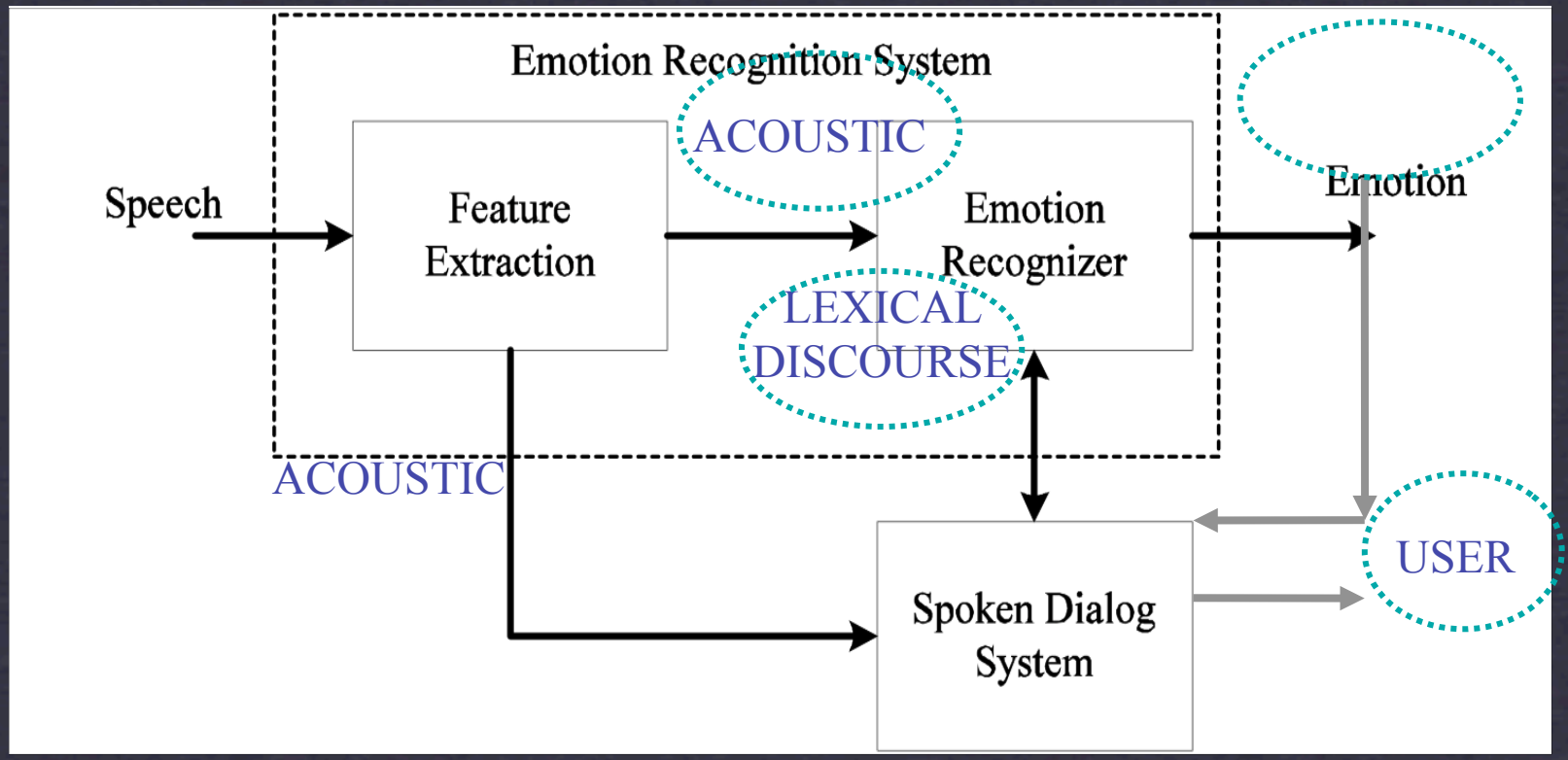
RECOGNIZE

- **WHAT**: SPOKEN LANGUAGE CONTENT
- **WHO**: SPEAKER IDENTITY,
- **HOW**: SPEAKING STYLE AND EMOTIONS

AUTOMATICALLY FROM SPOKEN LANGUAGE

Recognizing Emotions?

- **Expression versus experience**
- **Description**
  - Categorical (e.g., happy, sad..), Dimensional (arousal, valence)
  - Emotion Profiles
  - Dynamic descriptions



Lee & Narayanan, Toward recognizing emotions in spoken dialogs, IEEE Trans. Speech&Audio, 2005

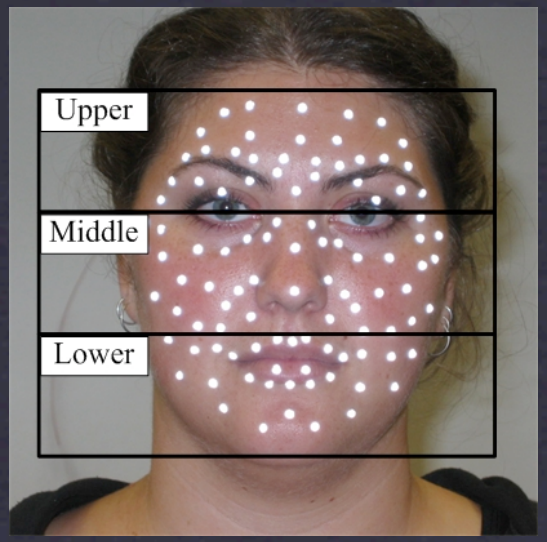
Enriching behavior descriptions further....
"Situated" Interactions & Conversational Computing

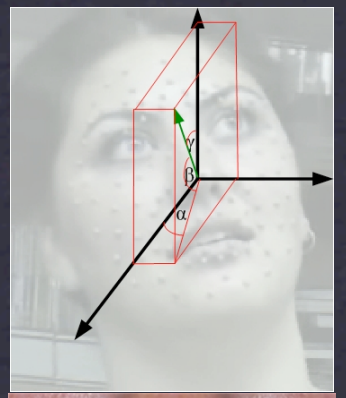
- Multimodality
- Interaction dynamics

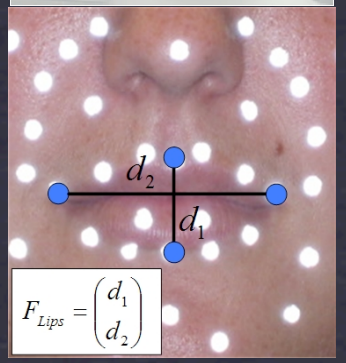
Modeling gestures/speech interrelation

VISUAL AND VOCAL FEATURES

- Speech
  - Prosodic features (source of the speech): Pitch, energy and they first and second derivatives
  - MFCC coefficients (vocal tract)
- Visual features
  - Head motion
  - Eyebrow
  - Lips
  - Each marker grouped in Upper, middle and lower face regions







C. Busso and S. Narayanan. Interrelation between Speech and Facial Gestures in Emotional Utterances: A single subject study. IEEE Transactions on Audio, Speech and Language Processing. 15(8): 2331 – 2347, November 2007.

Multimodal Emotion Recognition

- From speech
  - Average ~70%
  - Confusion sadness-neutral ( )
  - Confusion happiness-anger ( )
- From facial expression
  - Average ~85%
  - Confusion anger-sadness ( )
  - Confusion neutral-happiness ( )
  - Confusion sadness-neutral ( )
- Multimodal system (feature-level)
  - Average ~90%
  - Confusion neutral-sadness ( )
  - Other pairs are correctly separated

REDUNDANCY & COMPLEMENTARITY IN INFORMATION ENCODING

USING SVM

	Anger	Sadness	Happiness	Neutral
Anger	0.68	0.05	0.21	0.05
Sadness	0.07	0.64	0.06	0.22
Happiness	0.19	0.04	0.70	0.08
Neutral	0.04	0.14	0.01	0.81

	Anger	Sadness	Happiness	Neutral
Anger	0.79	0.18	0.00	0.03
Sadness	0.06	0.81	0.00	0.13
Happiness	0.00	0.00	1.00	0.00
Neutral	0.00	0.04	0.15	0.81

	Anger	Sadness	Happiness	Neutral
Anger	0.95	0.00	0.03	0.03
Sadness	0.00	0.79	0.03	0.18
Happiness	0.02	0.00	0.91	0.08
Neutral	0.01	0.05	0.02	0.92

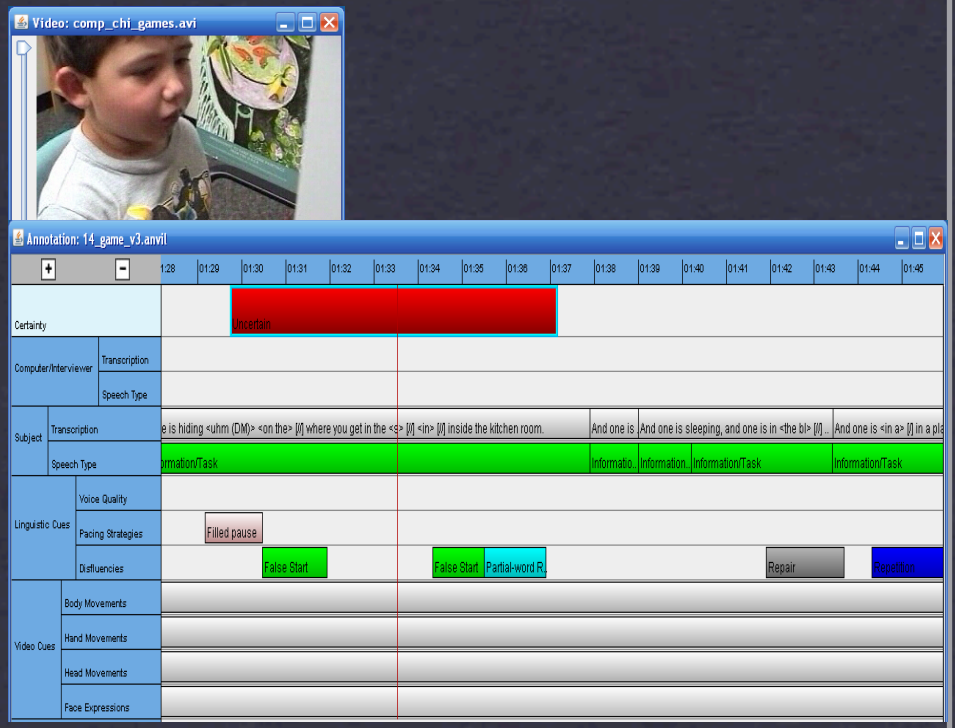
Busso et al, Analysis of emotion recognition using facial expressions, speech and multimodal information, ICMI, 2004

Is the child "certain"? (learning context)

TRAINED DECISION TREES (C4.5) USING "LEAVE-ONE-PERSON-OUT" TECHNIQUE

- USED DIFFERENT SUBSETS OF FEATURES: LEXICAL, ACOUSTIC, AND VISUAL
  - 6 LEXICAL: "I DON'T KNOW", FALSE STARTS, REPETITIONS, ETC.
  - 10 ACOUSTIC: PAUSES, ELONGATIONS, VOICE LOUDNESS, ETC.
  - 20 VISUAL: HEAD/FACE/HAND MOVEMENTS, FACIAL EXPRESSIONS

System	% Agreement *
Baseline (always guess certain)	72.32
Lexical	73.41
Visual	74.17
Acoustic	82.66
Lexical + Visual	74.17
Acoustic + Visual	81.94
Acoustic + Lexical	82.29
Acoustic + Lexical + Visual	82.66
Average Human Agreement	86.15



* ALL AGREEMENT STATISTICS ARE PAIRWISE AGREEMENT PERCENTAGES WITH THE GROUND TRUTH

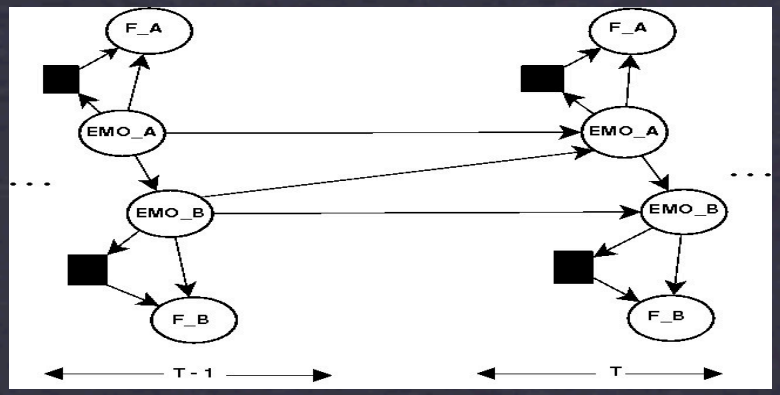
Emotion Tracking in Dyadic Spoken Interaction

- Problem
  - Emotion states tracking in dyadic spoken interaction
  - Incorporating "mutual influence" in the model
- Approach
  - Dynamic Bayesian Network: Joint modeling both speakers

•F0 Frequency
• Intensity/Energy
• Speech Rate

•Harmonic to Noise Ratio (HNR)
• 13 MFCC Coefficients
• 27 Mel Frequency Bank Filter Output

And functionals: Mean, Standard Deviation, Minimum, Maximum, 25% Quantile, 75% Quantile, Range, InterQuantile Range, Median, Kurtosis, Skewness



- Result
  - Emotion state tracking accuracy improves absolute 3.7 %

Chi-Chun Lee, C. Busso, S. Lee and S. Narayanan, Modeling mutual influence of interlocutor emotion states in dyadic spoken interactions, in Proceedings of InterSpeech, 2009

THIS TALK

- Some BSP building blocks

Example Behavioral Analysis Studies

- Family Studies: Marital couples
  - Blame patterns; positiveness/negativeness; humor/sarcasm
- Metabolic Health Monitoring
  - Characterizing physical behavior in context
- Autism Spectrum Disorders
  - Characterizing joint attention; quantifying socio-emotional discourse
  - Technology interfaces for elicitation and personalized interventions

An illustrative BSP application
Couples Marital Therapy: Behavior Coding

Human Behavior Observations

"YOU WORK TOO MUCH…"

"SO HARD TO TALK ABOUT…"

"..IS REALLY HOUSEHOLD CHORES STUFF"

"..TEMPER AND PATIENCE…"

CHRISTENSEN ET AL, JOURNAL OF CONSULTING AND CLINICAL PSYCHOLOGY, 2004

Human Behavior Coding

10-MINUTES LONG PROBLEM SOLVING INTERACTION

CODING IS PERFORMED AT THE SESSION-LEVEL

EXAMPLE CODING GOAL:

"IS THE HUSBAND SHOWING ACCEPTANCE?" (SCALE 1-9)

FROM THE MANUAL:

"INDICATES UNDERSTANDING AND ACCEPTANCE OF PARTNER'S VIEWS, FEELINGS, AND BEHAVIORS. LISTENS TO PARTNER WITH AN OPEN MIND AND POSITIVE ATTITUDE. . . . "

HUSBAND SPEAKING TURNS:

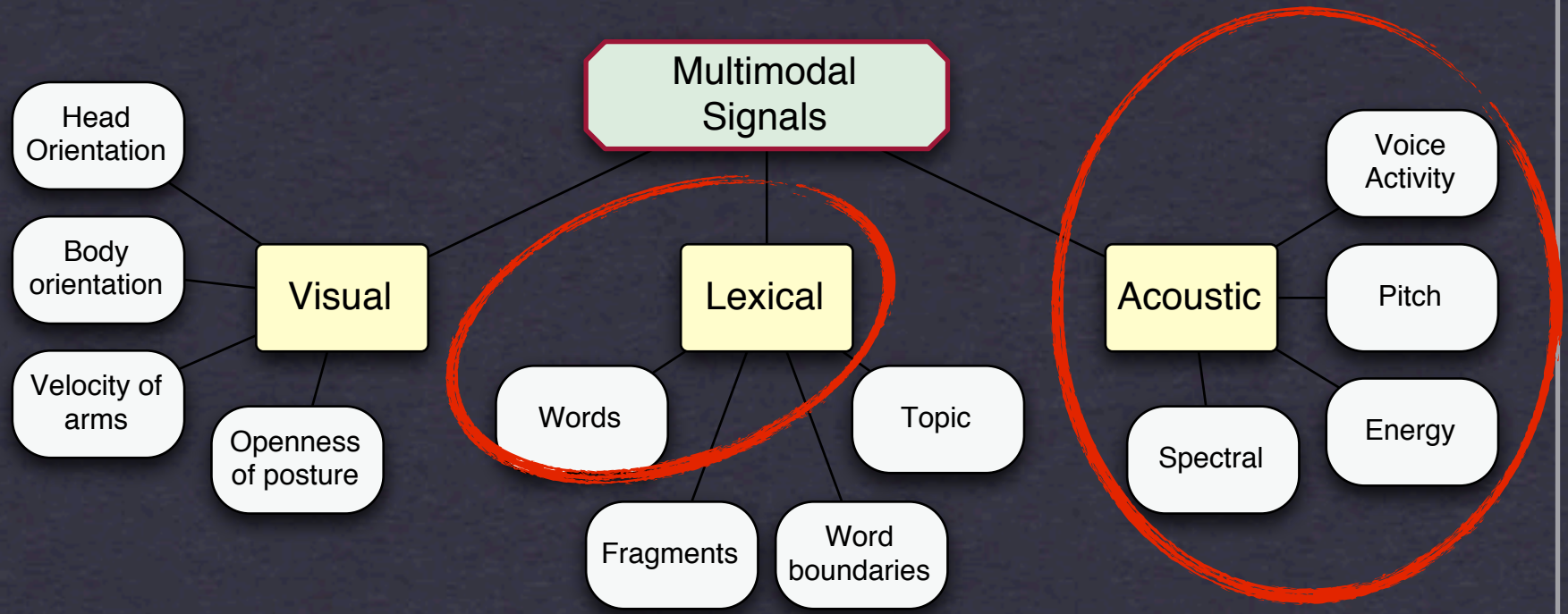
. . .

JONES AND CHRISTENSEN, SOCIAL SUPPORT INTERACTION RATING SYSTEM, UCLA, 1998

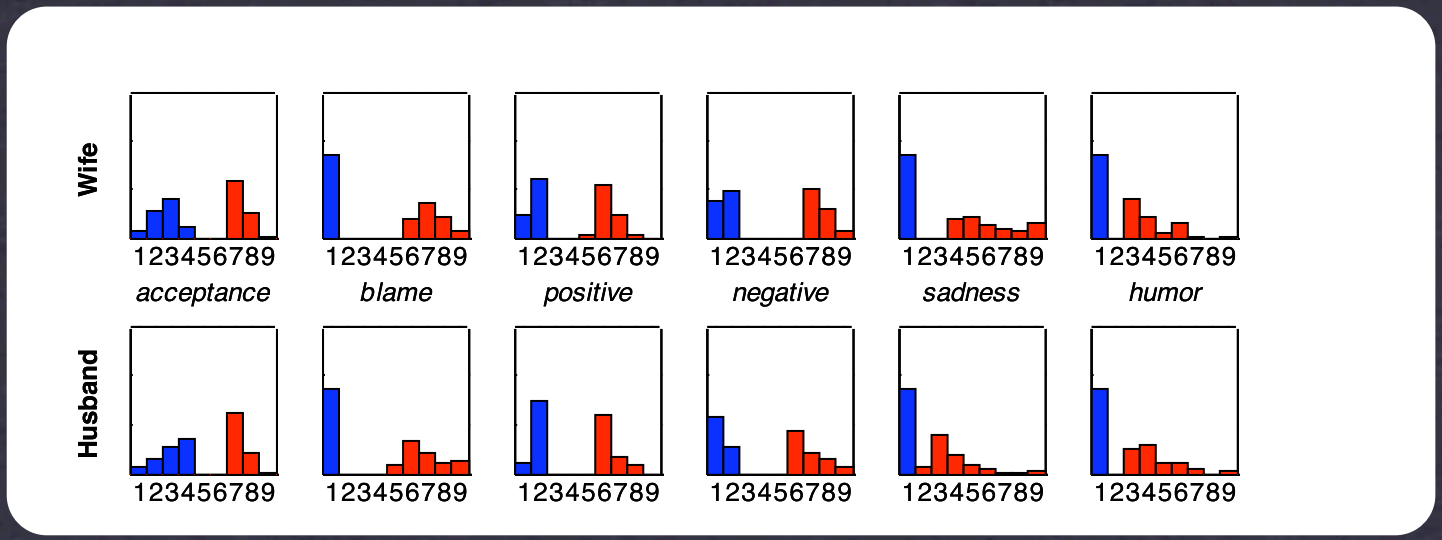
Corpus

- **Real couples in 10-minute problem-solving interactions**
  - Longitudinal study at UCLA and U. of Washington [Christensen et al. 2004]
  - 134 distressed couples received couples therapy for 1 year
- **574 sessions (96 hours)**
  - Split-screen video (704x480 pixels, 30 fps)
  - Single channel of far-field audio
- **Data originally only intended for manual coding**
  - Recording conditions not ideal
  - Video angle, microphone placement, and background noise varied

Estimate codes from audio information?



Separate extreme cases of session-level perceptual judgments using audio-derived information



- **So far we have been working on 6 codes:**
  - acceptance, blame, positive affect, negative affect, sadness, humor

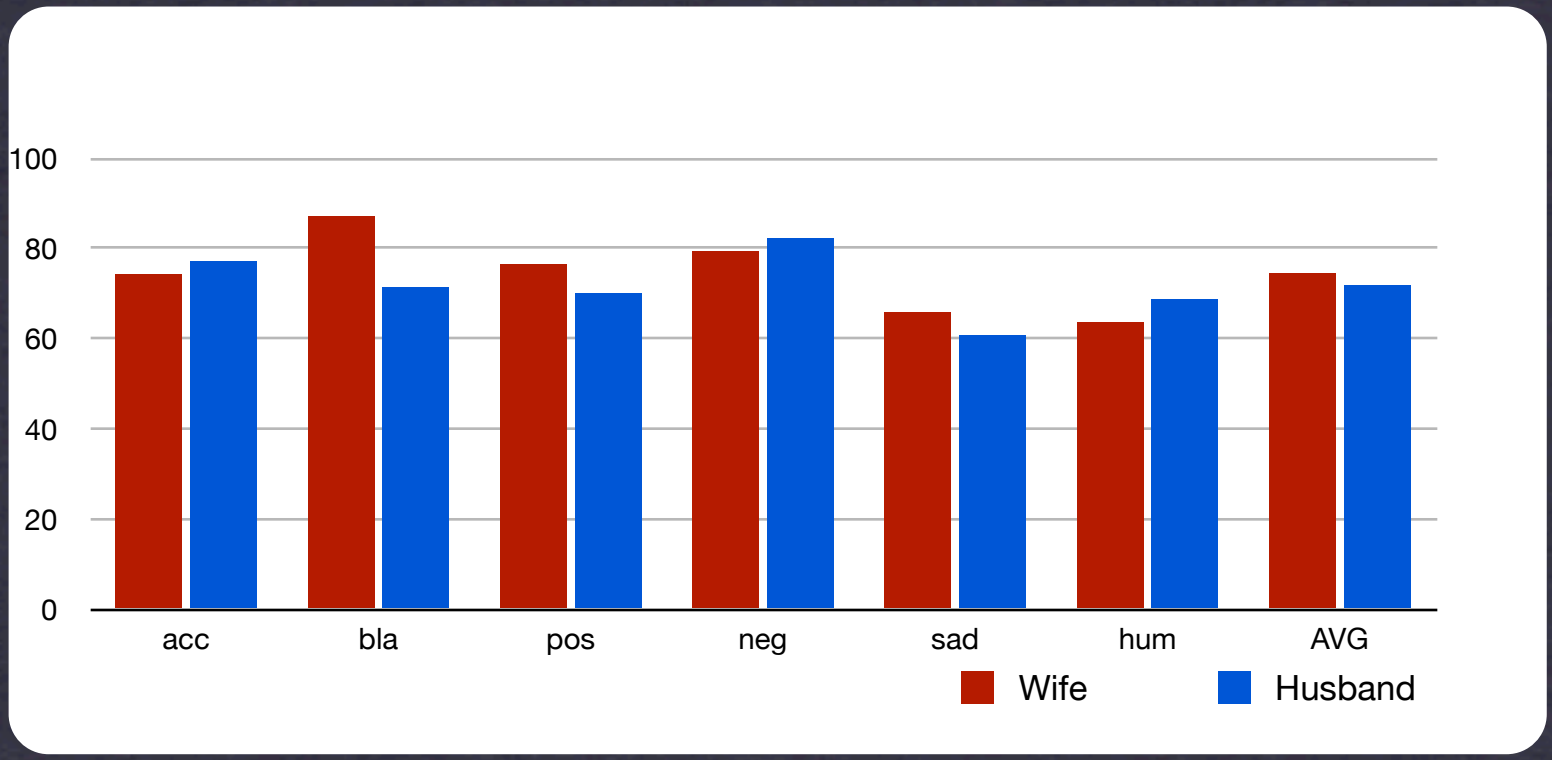
Acoustic Classifier

- **Capture relevant speech properties of spouses:**
  - **1)** Multiple low-level descriptors from speechevery 10 ms:
    - Prosody (pitch, energy), spectral (MFCCs), voice quality (jitter, shimmer)
  - **2)** Separate features for each spouse (wife, husband)
  - 3) Features at various temporal granularities (e.g., 5 second windows)
  - 4) Final features produced by computing functionals (e.g., mean, std. dev.) for each low-level descriptor, speaker, and temporal granularity

- **Apply binary classifier**
  - Finds mapping from acoustic feature space to blame class labels
  - Classifier: Support Vector Machine (SVM) with linear kernel
  - Confidence score: class probability estimates using LIBSVM [Chang and Lin 2001]

Acoustic-feature based Behavior Estimation

- **Use of acoustic low-level descriptors (LLDs)**
  - Binary classification task
  - Linear-SVM (best so far)
  - Global speaker-dependent cues capture evaluators' codes well



M. BLACK, ET AL "AUTOMATIC CLASSIFICATION OF MARRIED COUPLES' BEHAVIOR USING AUDIO FEATURES" INTERSPEECH 2010

Lexical-information based Behavior Estimation

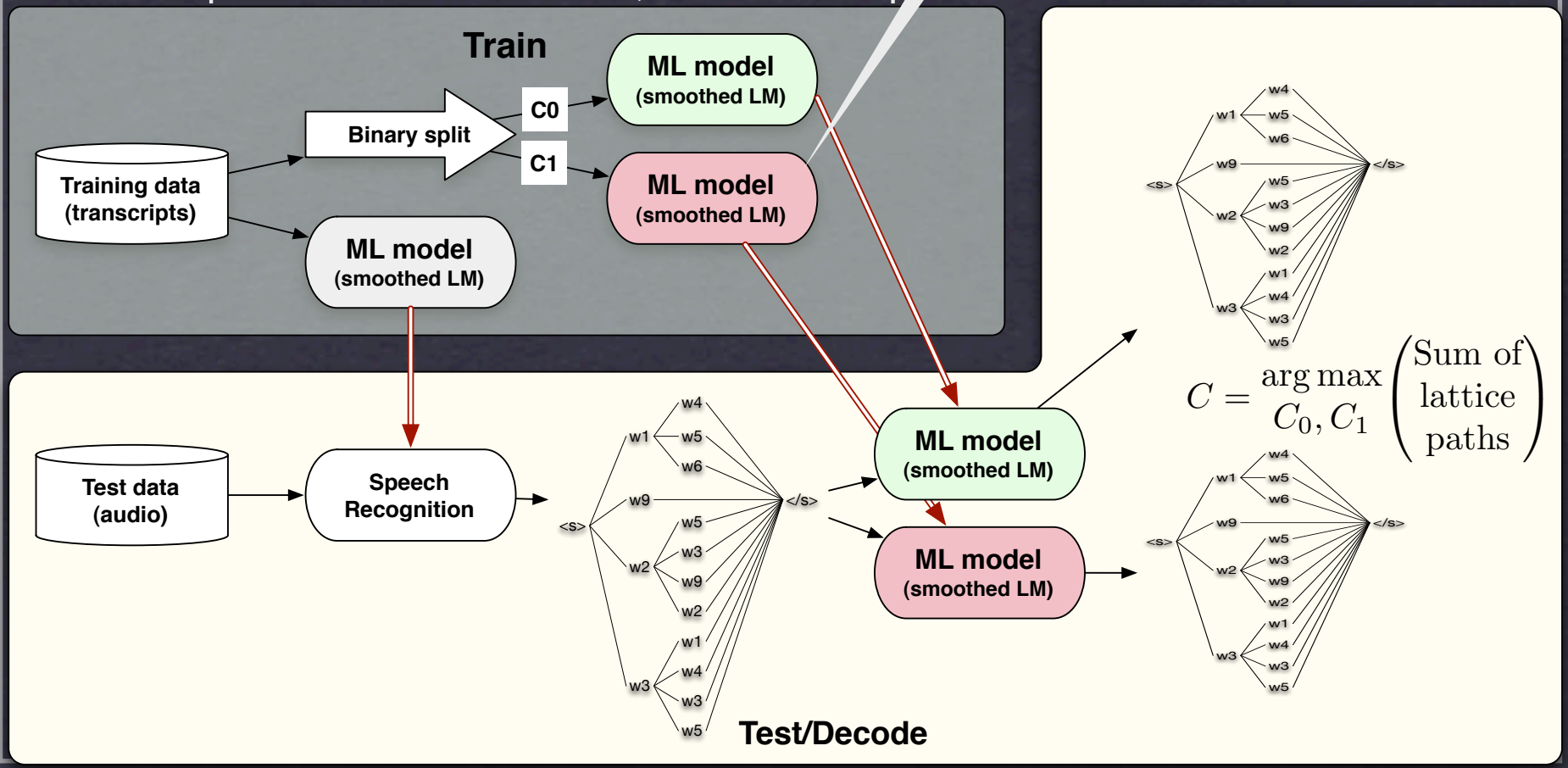
Partner	Transcript
H	WHAT DID I TELL YOU YOU CAN DO THAT AH AND EVERYTHING
W	BUT WHY DID YOU ASK THEN WHY DID TO ASK
H	AND DO IT MORE AND GET US INTO TROUBLE
W	YEAH WHY DID YOU ASK SEE MY QUESTION IS
H	MM HMMM
W	IF IF YOU TOLD ME THIS AND I AGREE I WOULD KEEP TRACK OF IT AND EVERYTHING
H	THAT'S THAT'S
W	**THAT'S AGGRAVATING VERY AGGRAVATING**
H	A BAD HABIT THAT
W	VERY AGGRAVATING
H	CAUSES YOU TO THINK THAT I DON'T TRUST YOU
W	THAT'S EXACTLY WHY THAT'S ABSOLUTELY THE WAY IT IS
H	AND IF I DON'T THE REASON FOR THAT IS AH
W	I DON'T CARE THE REASON YOU GET IT I GET IT TOO
H	THE REASON IS THE LONG TERM BAD PERFORMANCE
W	YEAH AND YOU KNOW WHY
H	MM HMMM
W	ALL YOU GET IS A NEGATIVE REACTION FROM ME

GEORGIOU, BLACK, LAMMERT, BAUCOM AND NARAYANAN. "THAT'S AGGRAVATING, VERY AGGRAVATING": IS IT POSSIBLE TO CLASSIFY BEHAVIORS IN COUPLE INTERACTIONS USING AUTOMATICALLY DERIVED LEXICAL FEATURES? PROCEEDINGS ACII, 2011

Lexical Behavior Estimation

- Logical way is to transcribe
- Speech recognition is inaccurate
- Use probabilistic decisions, i.e. lattices: probabilistic trees of words

Context: Data mining can enrich these



Informing experts

- Through the analysis we can inform experts
  - Example: Top and bottom words that contributed to (correct) classification of a partner as “blaming”

Most blaming words in terms of discriminative contribution				Least blaming words in terms of discriminative contribution			
Word	No Bl. log prob	Blame log prob	Δ	Word	No Bl. log prob	Blame log prob	Δ
YOU	-95.49	-85.88	-9.61	EXPECTS	-16.70	-17.84	1.14
YOUR	-51.24	-47.18	-4.06	CONSIDERATION	-16.11	-17.31	1.21
ME	-40.27	-37.74	-2.53	KNOW	-35.10	-36.62	1.53
TELL	-33.97	-32.46	-1.51	INABILITY	-16.76	-18.32	1.55
ACCEPT	-25.44	-23.99	-1.45	SESSION	-20.51	-22.07	1.56
CARING	-27.05	-25.91	-1.14	OF	-44.50	-46.26	1.76
KITCHEN	-21.22	-20.21	-1.02	ANTICIPATION	-22.22	-24.21	2.00
TOLD	-29.04	-28.19	-0.85	THINK	-35.70	-37.77	2.07
NOT	-40.32	-39.59	-0.73	WE	-29.39	-31.75	2.36
WHAT	-51.47	-50.77	-0.69	I	-99.92	-102.49	2.57
INTIMACY	-43.16	-42.53	-0.63	THAT	-91.30	-93.97	2.67
IT	-42.70	-42.18	-0.52	UM	-64.75	-70.76	6.01

Some challenges (and preliminary) approaches

- Any single feature stream offers partial, noisy code information

  ➡ **Multimodal approach**

- Not all portions of the feature stream are equally relevant in explaining an overall behavior description

  ➡ **Salient instances: Multiple instance learning**

- Behavior ratings are relative, often on an ordered scale

  ➡ **Ordinal regression**

- Not all human observers/evaluators are equally reliable, and reliability is data dependent

  ➡ **Realistic models of human observers/evaluators**

- Behavior is a part of an interaction: mutual dependency between interlocutors

  ➡ **Models of entrainment**

Fusion Results: Estimating "Blame"

- Exploit complementary information from language and speech
- Score-level fusion of classifiers using confidence scores

Classifier Type	Accuracy
Baseline Chance	50.0%
Language	75.4%
Acoustic	79.6%
Fusion	82.1%

- **REMARKS**
  - Lower performance of language classifier due to ASR issues
  - Fusion advantageously uses language and acoustic information
  - Feasible to model high-level behaviors with automatically derived speech and language information

Behavior Collection Space: Multichannel Multimodal

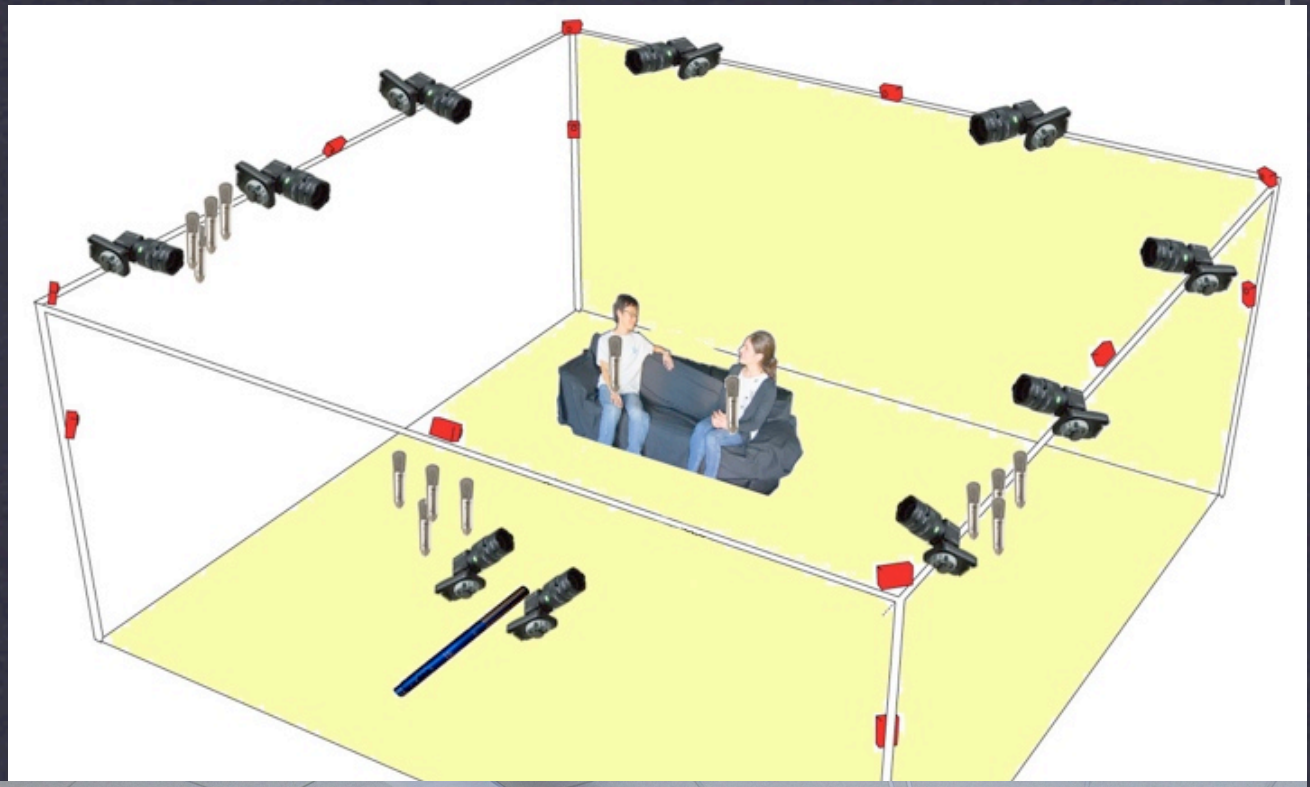
Audio:

- 3 4-mic T-arrays
- 2 lapel mics
- 1 shotgun mic

Video:

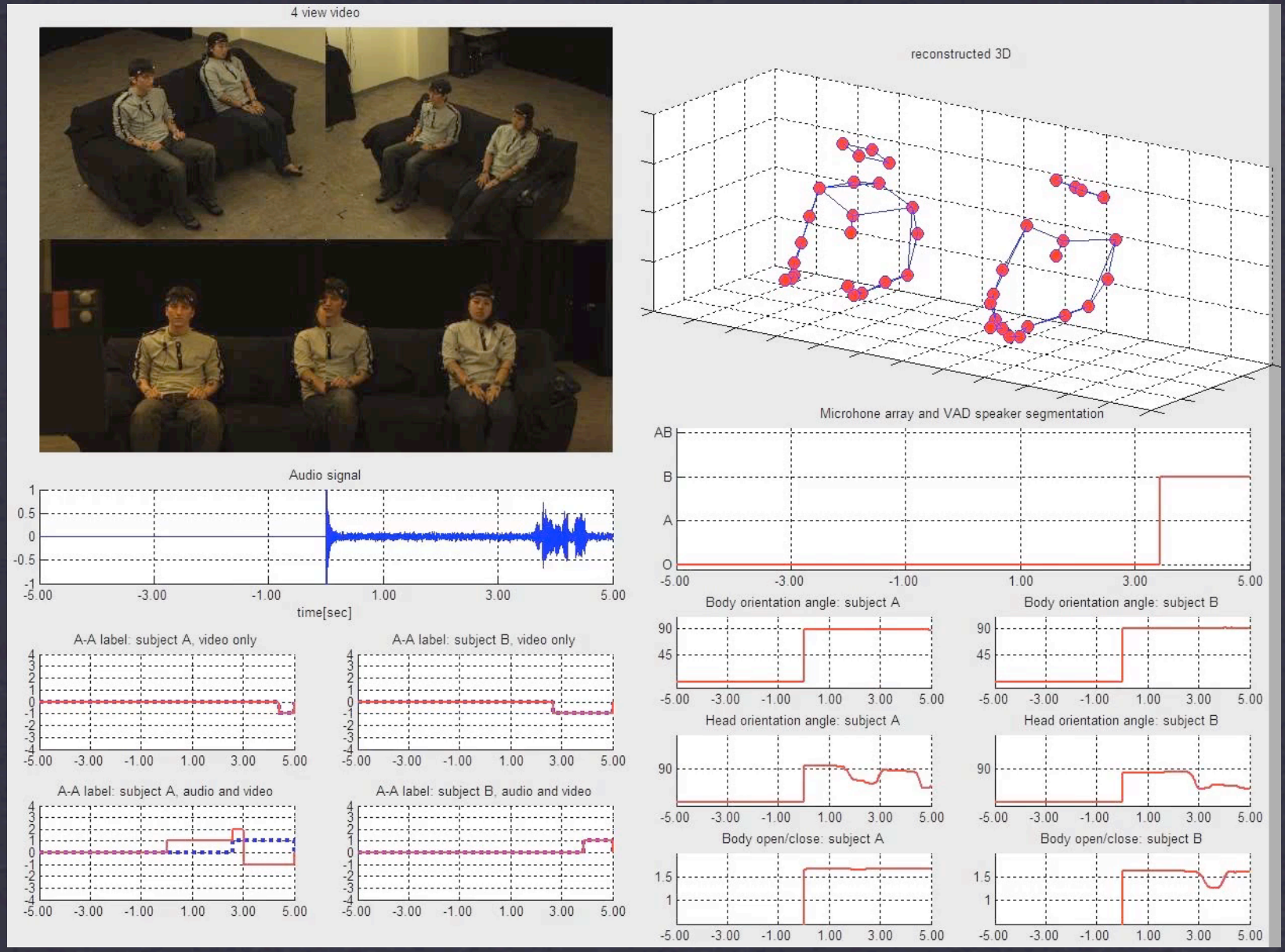
- 10 HD cameras (PointGrey Flea 2)
- Motion capture: 12 ViconQ Sensors

Accurate synchronization



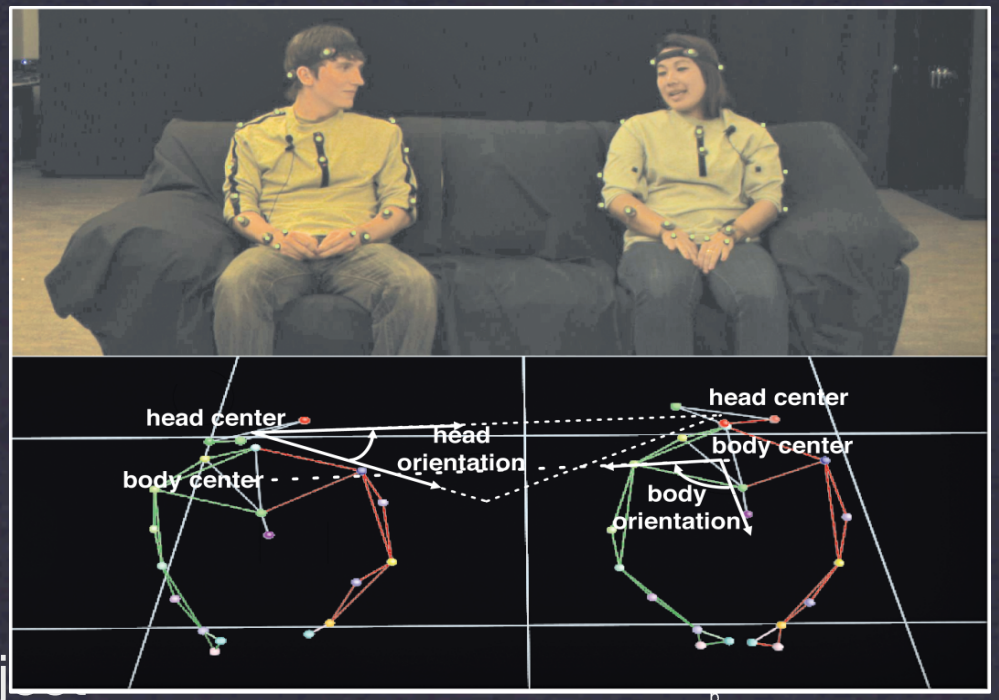
V. ROZGIĆ, B. XIAO, A. KATSAMANIS, B. BAUCOM, P. G. GEORGIOU, AND S. NARAYANAN, "A NEW MULTICHANNEL MULTIMODAL DYADIC INTERACTION DATABASE" INTERSPEECH 2010

Example Multimodal Data



Low Level Descriptor Features

- Audio features:
  - pitch
  - energy
  - MFCCs
  - speaker segmentation using VAD/Mic array
- Motion capture features:
  - Head/body orientation relative to the other subj
  - Arm velocity maximized over left and right hand
  - Body open/close in terms of average distance of left and right forearms to chest
- Functionals: mean, min, max, std of features on 3sec intervals
  - (3s windows with 1s overlap are motivated by the **3s coding rule**)



ORDINAL LOGISTIC REGRESSION FOR BEHAVIORAL CODE ESTIMATION

Case Study: Approach-Avoidance (A-A) Codes

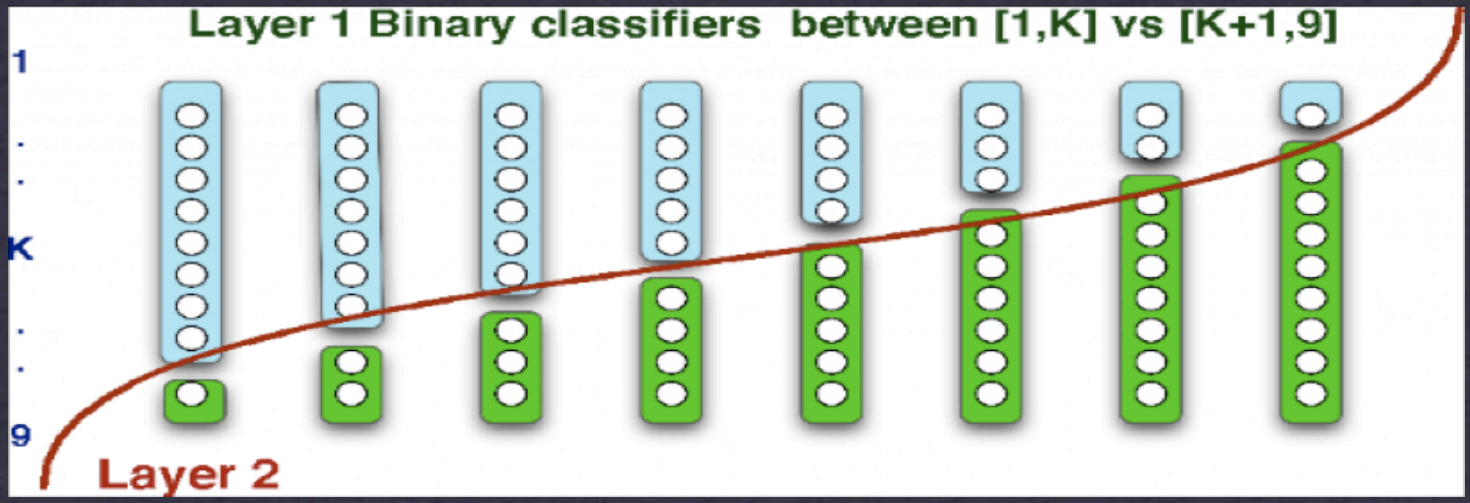
- Describes Moving Towards And Away From Objects, Situations Or Feelings
- Estimation: Via Multimodal Signal Processing And Statistical Modeling

Why Ordinal? A-A Labels Are Ordered Integers From 1 To 9

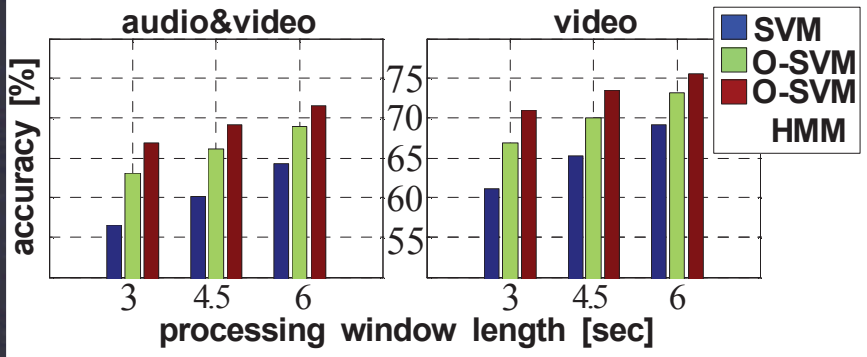
- Layer 1: A Series Of Independent Score Outputting Binary SVM Classifier
- Layer 2: Cumulative Logit Logistic Regression Model With Proportional Odds

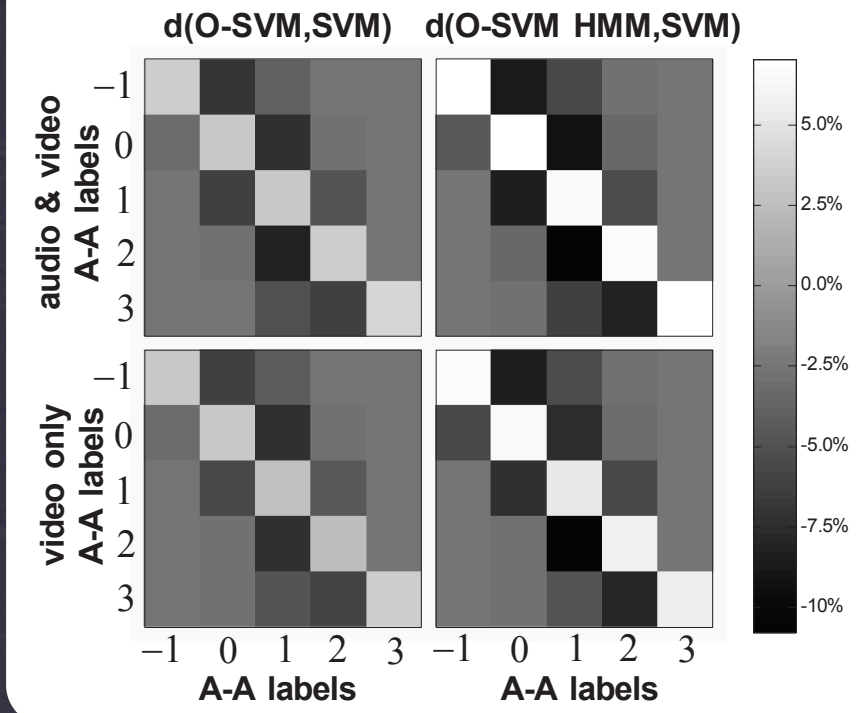
Multimodal Features

- Motion Capture Derived Measures, Acoustic Pitch And Energy Measures



APPROACH-AVOIDANCE (A-A) CODE ESTIMATION RESULTS





- Reference codes from experts:
  - USING AUDIO & VIDEO, OR ONLY VIDEO
- Three methods are compared:
  - PLAIN MULTI-CLASS SVM
  - ORDINAL SVM
  - ORDINAL SVM WITH HMM SMOOTHING

ON THE ESTIMATED LABEL SEQUENCE

- Ordinal SVM takes inherent order information of A-A codes into account
- Difference of confusion matrices show the advantage of ordinal regression method

Some challenges (and preliminary) approaches

- Any single feature stream offers partial, noisy code information

  ➡ **Multimodal approach**

- Behavior ratings are relative, often on an ordered scale

  ➡ **Ordinal regression**

✓ Not all portions of the feature stream are equally relevant in explaining an overall behavior description

➡ **Salient instances: Multiple instance learning**

- Not all human observers/evaluators are equally reliable, and reliability is data dependent

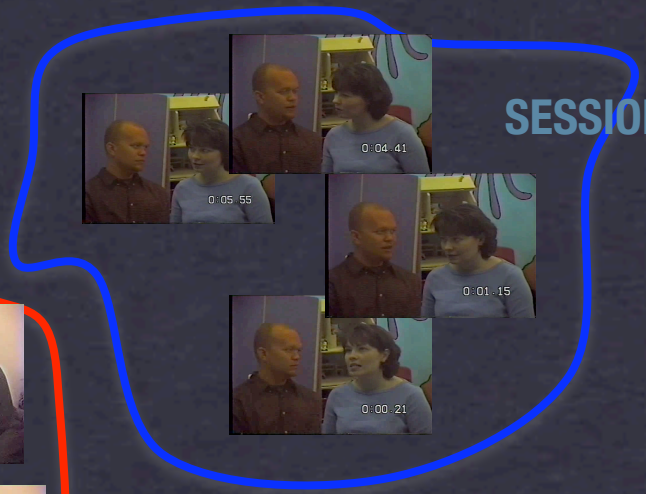
  ➡ **Realistic models of human observers/evaluators**

- Behavior is a part of an interaction: mutual dependency between interlocutors

  ➡ **Models of entrainment**

Multiple Instance Learning

EACH SPEAKER-TURN IS AN INSTANCE (OF BEHAVIOR)



THE PROBLEM:
CAN WE IDENTIFY THE SPEAKER TURNS (INSTANCES) THAT ARE SALIENT, GIVEN THAT WE ONLY HAVE THE SESSION-LEVEL CODES?

RED SESSIONS: NON-ACCEPTING SPOUSE
BLUE SESSIONS: ACCEPTING SPOUSE

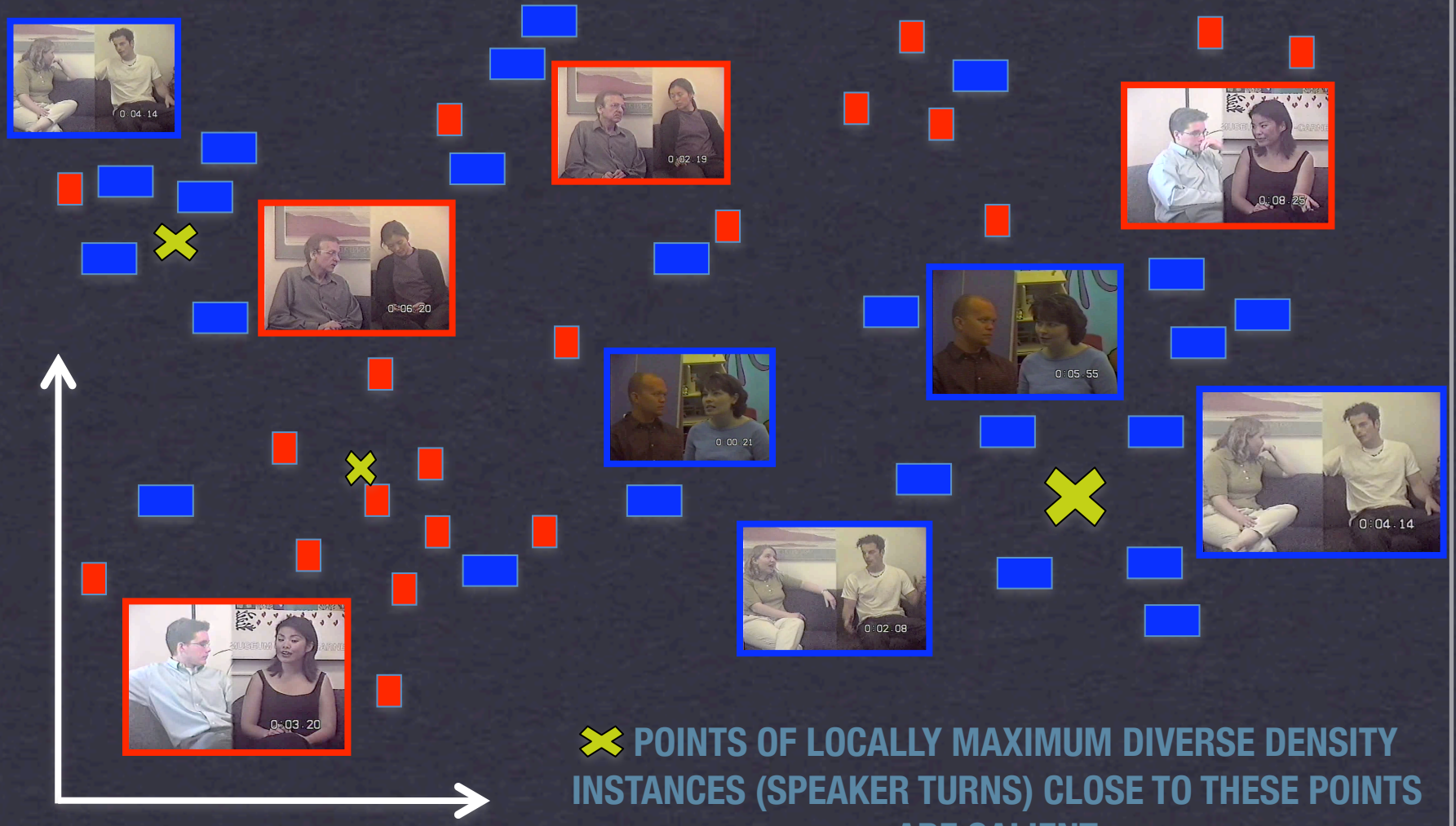
KATSAMANIS, GIBSON, BLACK, NARAYANAN, MULTIPLE INSTANCE LEARNING FOR CLASSIFICATION OF HUMAN BEHAVIOR OBSERVATIONS, ACII 2011

Diverse Density

TURNS FROM ACCEPTING SPOUSES

TURNS FROM NON-ACCEPTING SPOUSES

INSTANCES CLOSE TO POSITIVE BAGS AND FAR AWAY FROM NEGATIVE BAGS



× POINTS OF LOCALLY MAXIMUM DIVERSE DENSITY

INSTANCES (SPEAKER TURNS) CLOSE TO THESE POINTS ARE SALIENT

GIBSON, KATSAMANIS, BLACK, NARAYANAN, AUTOMATIC IDENTIFICATION OF SALIENT ACOUSTIC INSTANCES IN COUPLES' BEHAVIORAL INTERACTIONS USING DIVERSE DENSITY SUPPORT VECTOR MACHINES, INTERSPEECH 2011

Instance (speaker turn) representation

- **BAG-OF-WORDS REPRESENTATION:**
- **FREQUENCY HISTOGRAM OF A SET OF WORDS AND INTONATION PATTERNS**

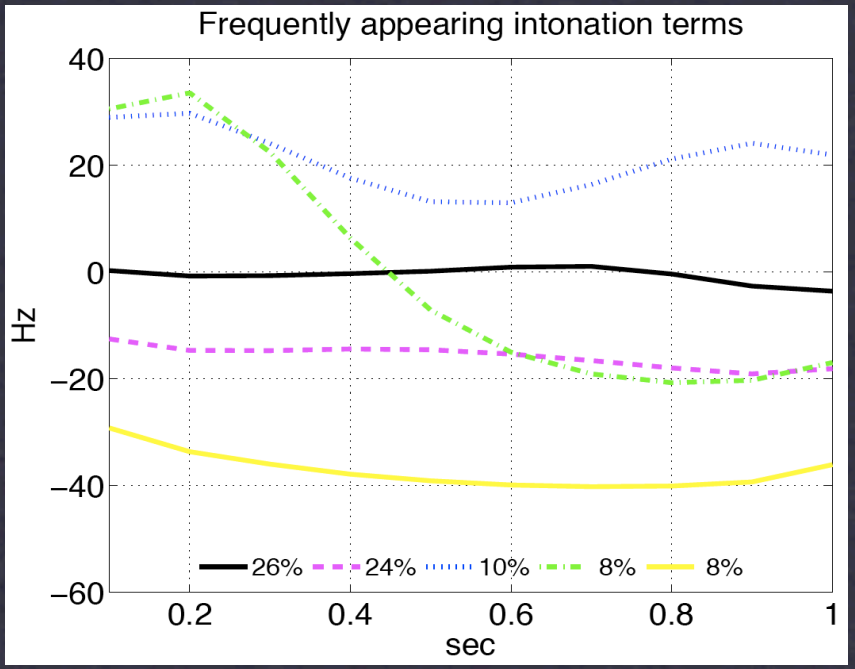
TEXT

REPRESENTATIVE (DISCRIMINATIVE) WORDS SELECTION: MAXIMIZING INFORMATION GAIN

Behavior	Informative words
acceptance	UM, TOLD, NOTHING, MM, YES, EVERYTHING, ASK, MORE, (LAUGH), CAN'T
blame	NOTHING, EVERYTHING, YOUR, NO, SAID, ALWAYS, CAN'T, NEVER, MM, TOLD
humor	(LAUGH), TOPIC, GOOD, MISSING, COOL, TREAT, SEEMED, TRULY, ACCEPT, CASE
negative	TOLD, KIND, MM, MAYBE, NOTHING, UM, YOUR, NEVER, CAN'T, (LAUGH)
positive	UM, KIND, NOTHING, MM, GOOD, (LAUGH), TOLD, CAN'T, MEAN, WHY
sadness	ACTUALLY, ONCE, WEEK, GO, OKAY, STAND, CONSTANTLY, UP, ALREADY, WENT

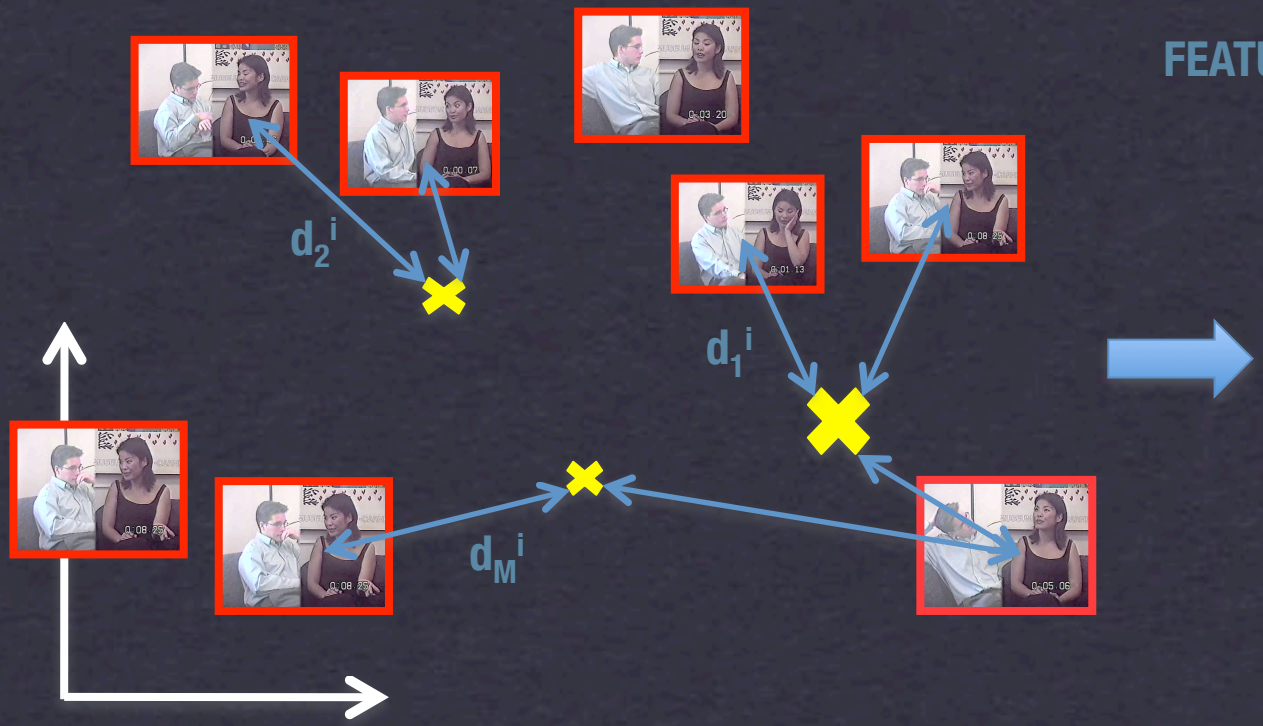
AUDIO

REPRESENTATIVE INTONATION PATTERNS: FOUND BY K-MEANS CLUSTERING.



Session Representation

EACH SESSION IS REPRESENTED BY THE VECTOR OF DISTANCES FROM THE SALIENT PROTOTYPES

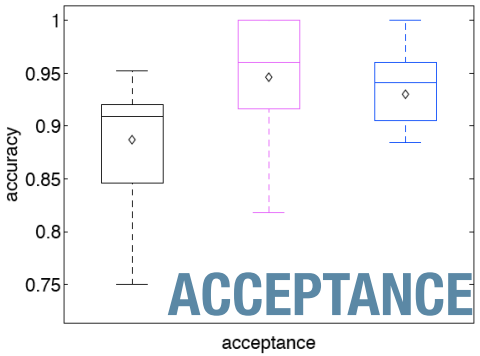


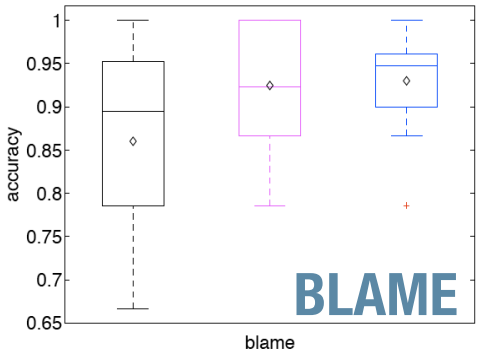
FEATURE VECTOR FOR SESSION

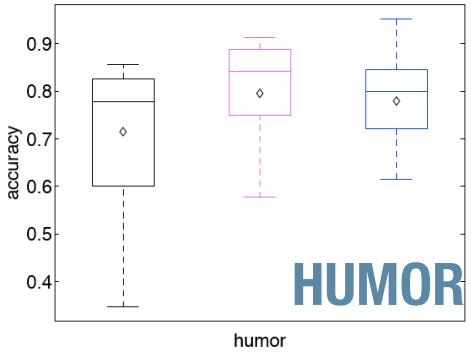
$$\begin{bmatrix} \min_i d_1^i \\ \min_i d_2^i \\ \vdots \\ \min_i d_m^i \\ \vdots \end{bmatrix}$$

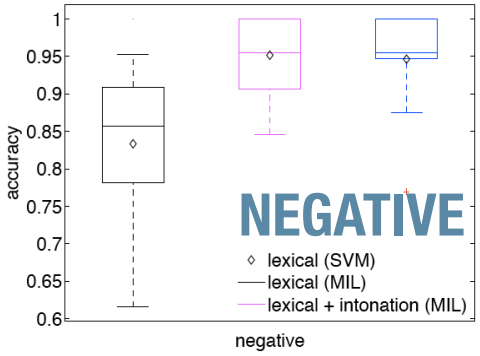
TO VALIDATE THE REPRESENTATION, WE RUN CLASSIFICATION EXPERIMENTS USING SVMS

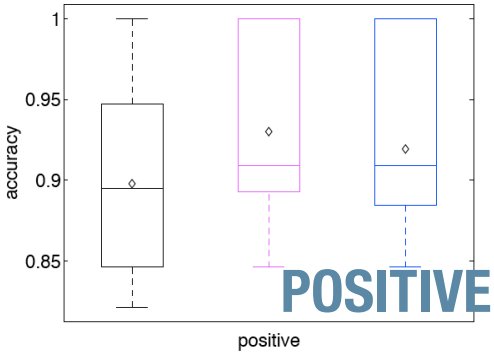
Classification Results

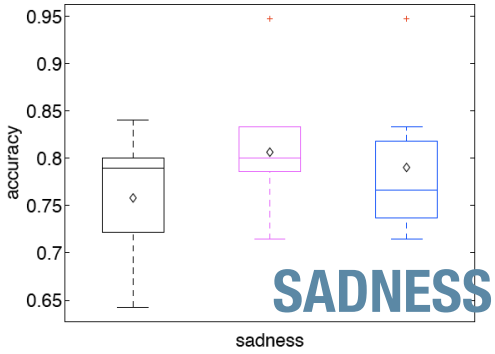












10-FOLD CROSS-VALIDATED RESULTS FOR SIX BEHAVIORAL CODES (HIGH VS LOW).

BLACK BOXES -- BASELINE: BAG-OF-WORDS REPRESENTATION OF THE WHOLE SESSION (WITHOUT EXPLOITING SALIENCY ESTIMATES)

- SIGNIFICANT PERFORMANCE IMPROVEMENT WITH MULTIPLE INSTANCE LEARNING

Remarks

- **Saliency currently based on discriminative criteria and fully data-driven**
  - Validate and improve estimates based on human expert saliency annotations
  - Derive reliability of our saliency estimates to increase their usability
  - Integrate machine processing and expert knowledge in a closed loop: Active learning possibilities
- **Initial experiments used only audio**
  - Exploit visual data: facial expressions and body language

Some challenges (and preliminary) approaches

- Any single feature stream offers partial, noisy code information

  ➡ **Multimodal approach**

- Behavior ratings are relative, often on an ordered scale

  ➡ **Ordinal regression**

✓ Not all portions of the feature stream are equally relevant in explaining an overall behavior description

➡ **Salient instances: Multiple instance learning**

- Behavior is a part of an interaction: mutual dependency between interlocutors

  ➡ **Models of entrainment**

- Not all human observers/evaluators are equally reliable, and reliability is data dependent

  ➡ **Realistic models of human observers/evaluators**

Interaction Models

- **Interaction Synchrony / Entrainment** [Kimura 2006]
  - Mutual adaptation of verbal/nonverbal behaviors in dyadic interactions
- **Quantification of Prosodic Entrainment**
  - Signal-derived quantitative measure
- **Positive vs. Negative valence in interactions**
  - Higher degree of entrainment in positive interactions [Kimura 2006, Warner 1987]
  - Entrainment measures as features for automatic classification [Margolin 1998]

Hypotheses

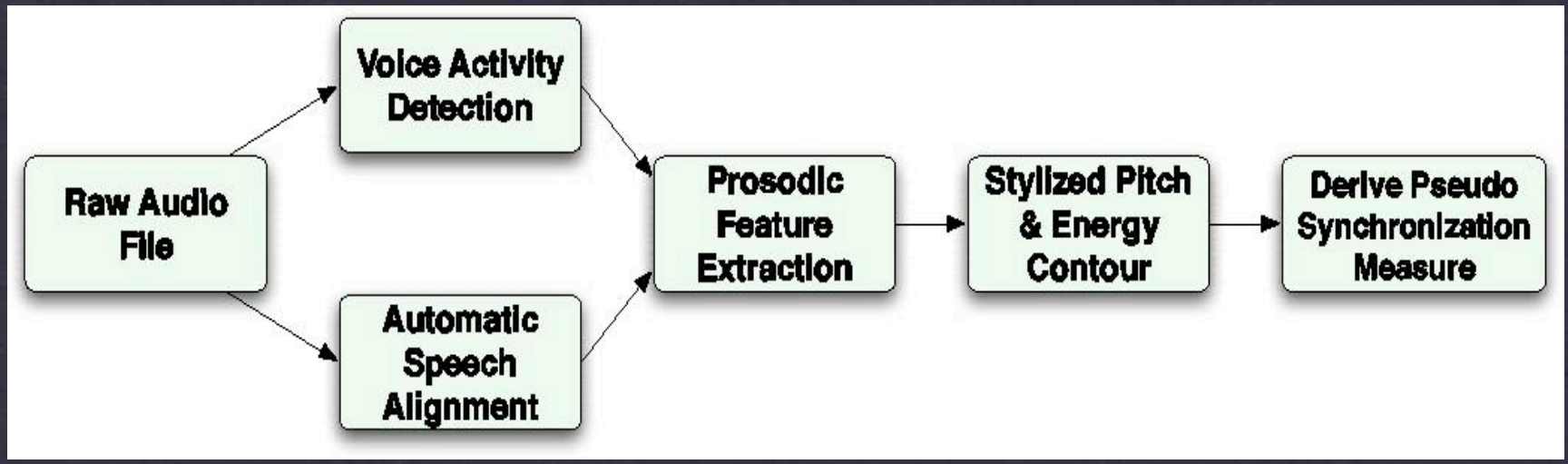
Signal-derived measures quantify degrees of prosodic entrainment

Entrainment measures discriminate interactions with positive vs. negative emotions

CHI-CHUN LEE, MATTHEW BLACK, ATHANASIOS KATSAMANIS, ADAM LAMMERT, BRIAN BAUCOM, ANDREW CHRISTENSEN, PANAYIOTIS G. GEORGIOU, SHRIKANTH NARAYANAN. QUANTIFICATION OF PROSODIC ENTRAINMENT IN AFFECTIVE SPONTANEOUS SPOKEN INTERACTIONS OF MARRIED COUPLES. IN PROCEEDINGS OF INTERSPEECH, 2010.

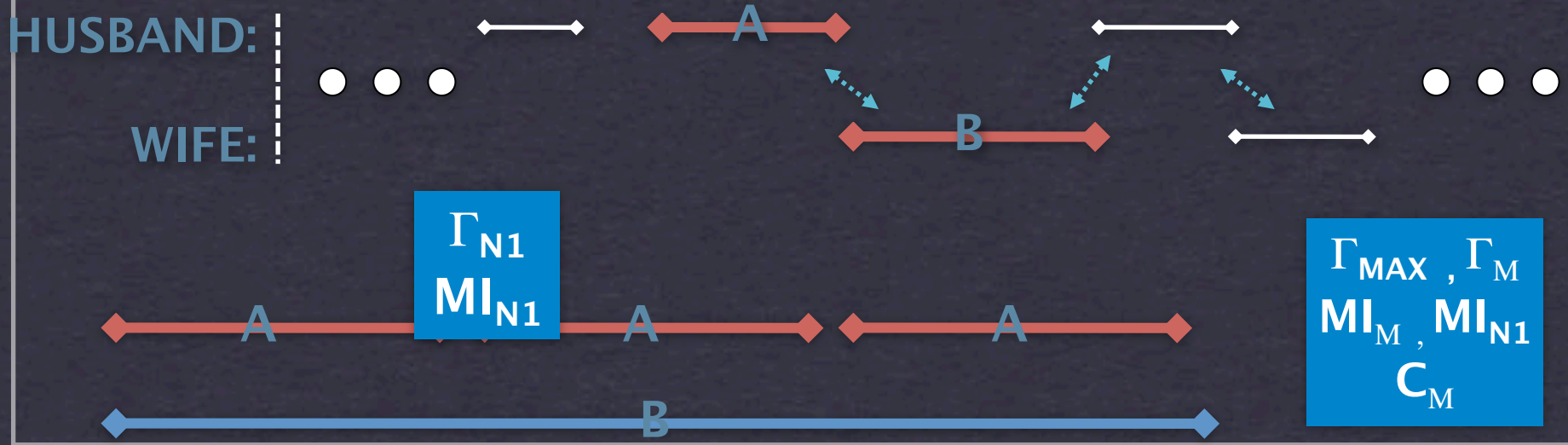
Procedure of Deriving Prosodic Entrainment Measure

ACOUSTIC ANALYSIS **ENTRAINMENT MODELING**



Entrainment Measures

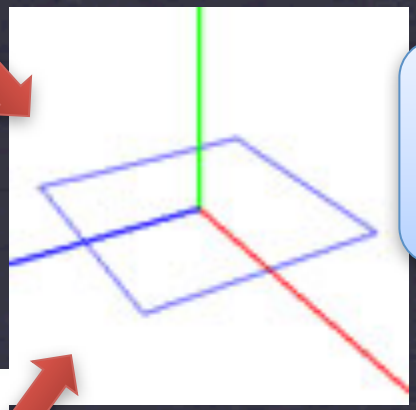
- **Square of correlation coefficients (Γ_{μ}, Γ_{N1}, Γ_{MAX})**
  - Linear dependency between two random variables
- **Mutual information (*mi*$_{\mu}$, *mi*$_{n1}$, *mi*$_{max}$)**
  - Information theoretic mutual dependency measure between two random variables
- **Coherence (c_{μ}): Mean spectral coherence across all frequency bands**
  - A measure of degree of causality between a system's input and output relationship



Quantifying Vocal Entrainment using Principal Component Analysis

HUSBAND:

(1) CONSTRUCTING PCA VOCAL CHARACTERISTIC SPACE



(3) COMPUTE PRESERVED VARIANCE (ENTRAINMENT)

WIFE:

(2) PROJECTING ONTO CONSTRUCTED PCA SPACE

- Two directions of entrainment per spouse per turn: Toward (as shown), From (swap (1) and (2))
- Mitigates issues of non co-occurring signals (turn-taking nature of conversation)
- Mitigates issues of variable length of turns

Quantifying Entrainment and Affective State Classification

- **Analysis of PCA-based entrainment measures**
  - **Positive Emotion > Negative Emotion** (consistent with psychology studies)
  - **Pair of Turns in Dialog > Random Turns** (consistent with notions of naturally occurring entrainment in interpersonal interaction)
  - **Varying dynamics across different married couples interactions** (differences in the value of entrainment in entraining *toward* and *from*)
- **Affective Classification:** engineering utility of entrainment measure
  - Accuracy using vocal entrainment ***solely*** is 53.93%
  - Multiple-length of salient turns provide the best accuracy
  - Multiple instance learning based on local saliency outperforms others

CHI-CHUN LEE, MATTHEW BLACK, ATHANASIOS KATSAMANIS, ADAM LAMMERT, BRIAN BAUCOM, ANDREW CHRISTENSEN, PANAYIOTIS G. GEORGIOU, SHRIKANTH NARAYANAN. QUANTIFICATION OF PROSODIC ENTRAINMENT IN AFFECTIVE SPONTANEOUS SPOKEN INTERACTIONS OF MARRIED COUPLES. IN PROCEEDINGS OF INTERSPEECH, 2010.

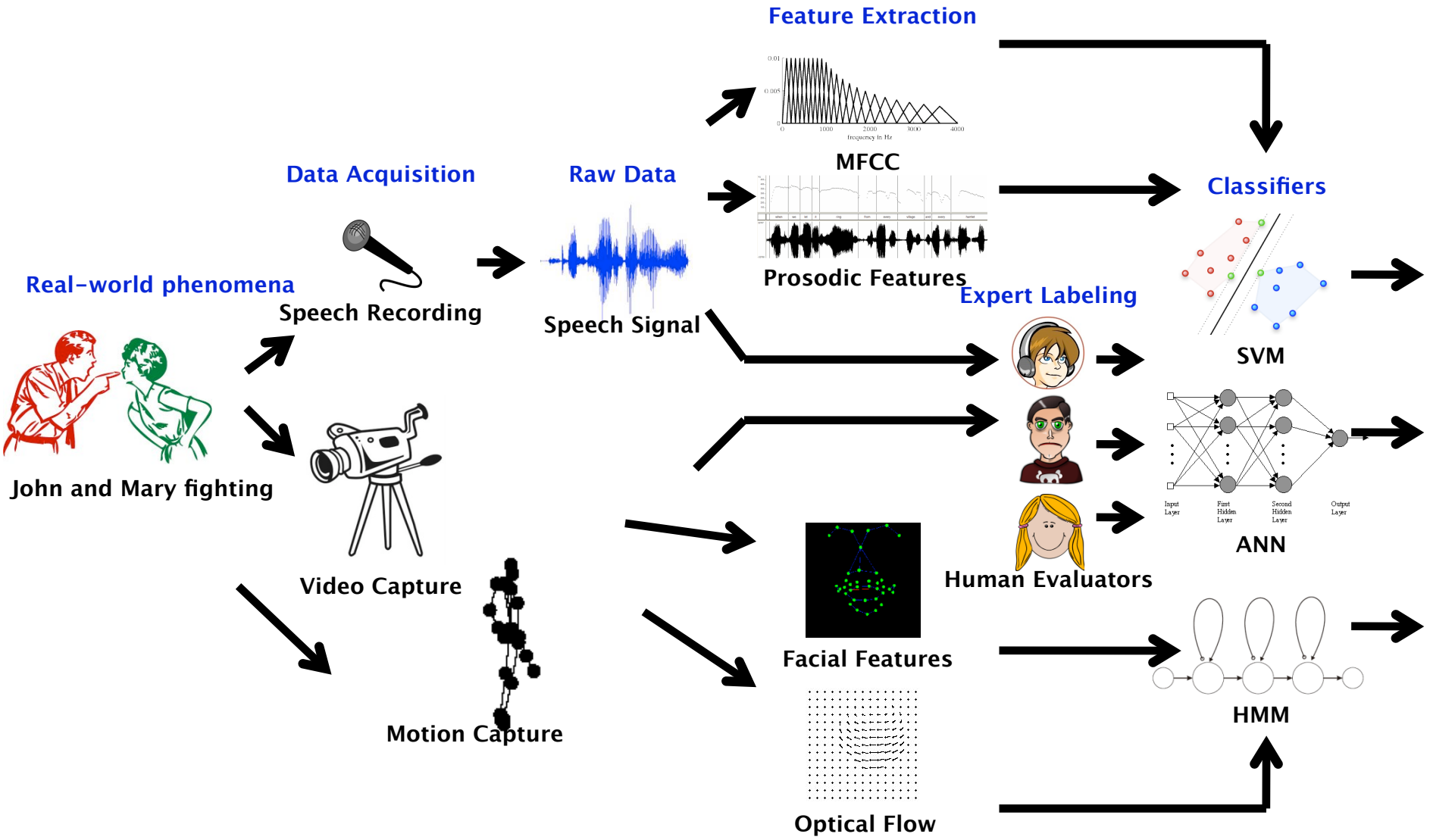
Interaction behavior: Summary

- **Signal-derived prosodic entrainment measures**
  - Capture notions of interaction synchrony
  - Consistent with existing psychology studies
  - Provide discriminative power in classifying emotional valence

- **Numerous open questions...**
  - Other modalities (body posture, turn-taking patterns)
  - Propose other ways of quantifying entrainment
  - A better dynamic framework for classification
  - Visualization of the flow/change of entrainment measures

Modeling the human observer

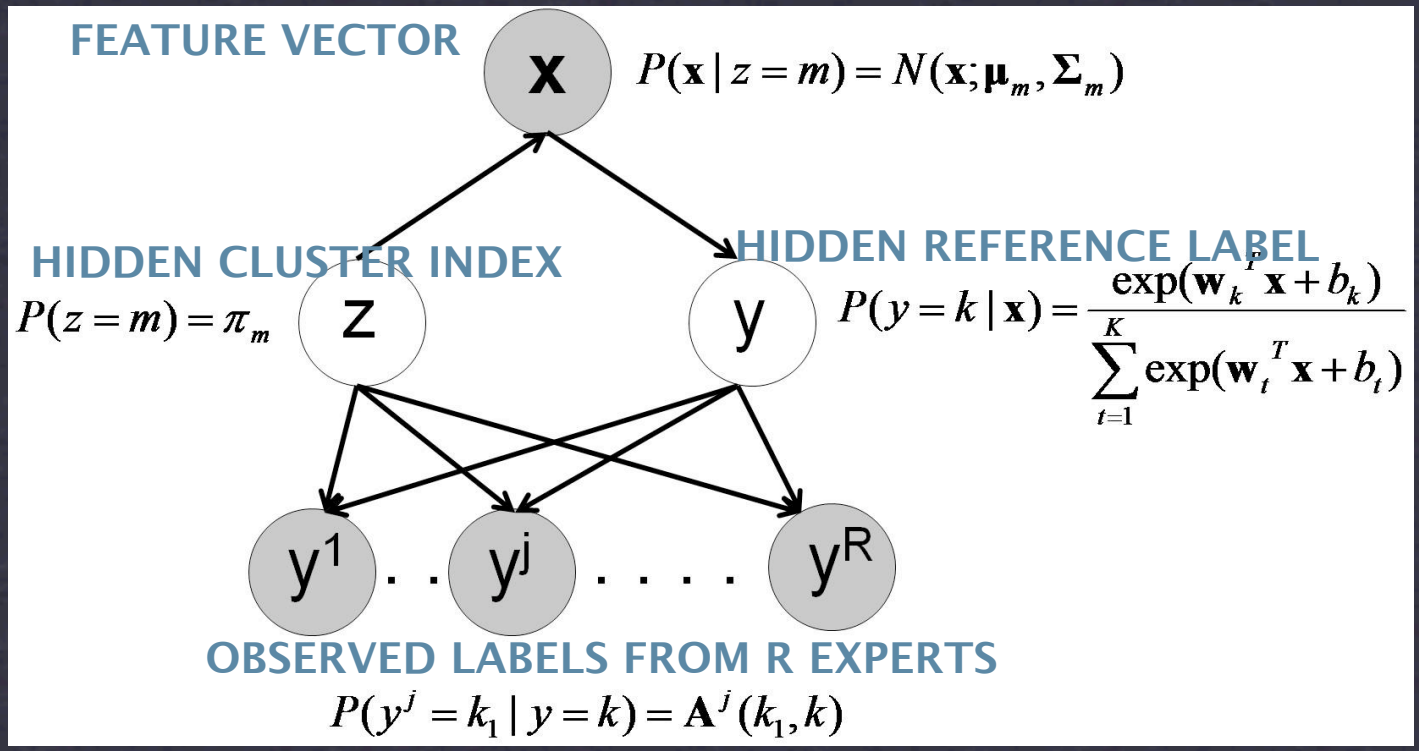
Real – World Pattern Recognition



There are multiple perspectives at each stage in the real-world tasks.

Fusion of Labels from Multiple Diverse Experts Without Knowing Reference Label

GLOBALLY VARIANT LOCALLY CONSTANT (GVLC) MODEL



- Experts could be human evaluators and/or machine classifiers.
- Model captures data-dependent expert reliability.
- Developed MAP-EM algorithm for learning unknown parameters.

KARTIK AUDHKHASI AND S. NARAYANAN. EMOTION CLASSIFICATION FROM SPEECH USING EVALUATOR RELIABILITY-WEIGHTED COMBINATION OF RANKED LISTS. IN PROCEEDINGS OF ICASSP, 2011

Results on Emotional Speech Classification

- Acted emotional speech database – label of acted emotion used as reference for evaluation.
- Four human labelers.

Model	Emotion Classification	Valence Classification	Activation Classification
Simple Plurality	78.4 %	61.54 %	94.12 %
Constant reliability based fusion *	80.39 %	61.54 %	90.20 %
GVLC	82.35 %	63.46 %	96.08 %

- Performance benefit is greater when expert reliability is highly variable over data instances.

* V. C. RAYKAR ET AL., "LEARNING FROM CROWDS", JMLR, VOL. 11, MARCH 2010

Open Questions and Opportunities

- **Studying *diversity* of expert ensembles.**
  - How to quantify it?
  - How does it affect ensemble performance?
  - Can it be controlled while parameter estimation?
  - Which expert fusion method exploits diversity the most?
- **Studying diversity of feature sets.**
- **Exploring other real-world problems involving multiple expert labeling**:
  - time series data, e.g., body gesture sequence in videos.
  - Bayesian network structure learning, e.g., learning dependence between codes in couples therapy.

THIS TALK

- Some BSP building blocks

Example Behavioral Analysis Studies

- Family Studies: Marital couples
  - Blame patterns; positiveness/negativeness; humor/sarcasm
- Autism Spectrum Disorders
  - Quantitatively characterizing behavior phenotyping
  - Technology interfaces for personalized interventions
- Metabolic Health Monitoring
  - Characterizing physical behavior in context

MULTIMEDIA INTERFACES
MOBILE SETTINGS
BEYOND A-V.

65

Autism Spectrum Disorders (ASD)

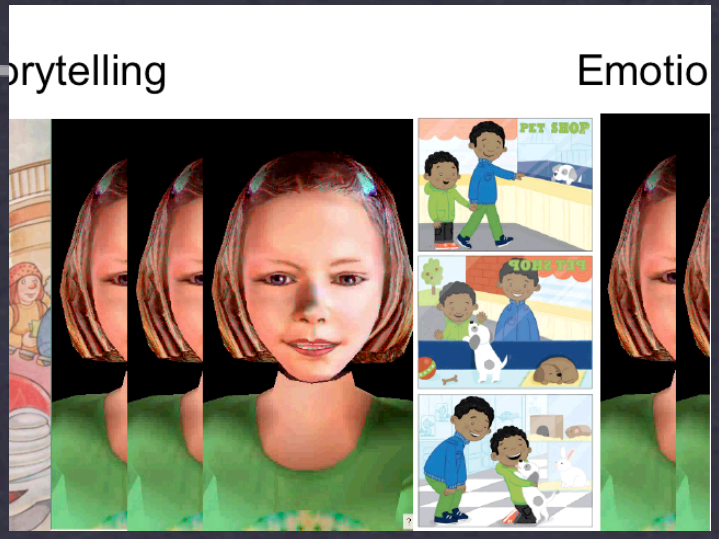
- **1 in 110 children diagnosed with ASD (CDC, 2010)**
- **ASD characterized by**
  - Difficulties in social communication, reciprocity
  - Repetitive or stereotyped behaviors and interests

Technology possibilities in ASD include

- Computational techniques to
  - Better understand communication and social patterns of children
  - Help support diagnoses with quantifiable and adaptable metrics
  - Automatically track children's progress during interventions
- Interfaces/systems to elicit, encourage, analyze behavior:
  - Complex, but phased; Structured; Naturalistic

Data

- **Automatic data logging**
  - ECA behavior
  - Wizard flag
- **Recorded data: Clinic**
  - Three Sony Handycams
  - Two shotgun microphones
- **Extracted data**
  - Manual transcription – Word usage statistics
  - Audio feature extraction

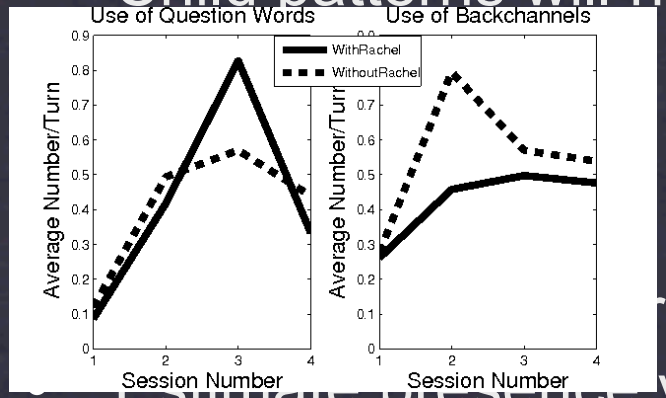


EMILY MOWER, MATTHEW BLACK, ELISA FLORES, MARIAN WILLIAMS AND SHRIKANTH NARAYANAN. DESIGN OF AN EMOTIONALLY TARGETED INTERACTIVE AGENT FOR CHILDREN WITH AUTISM. IN PROCEEDINGS OF ICME 2011

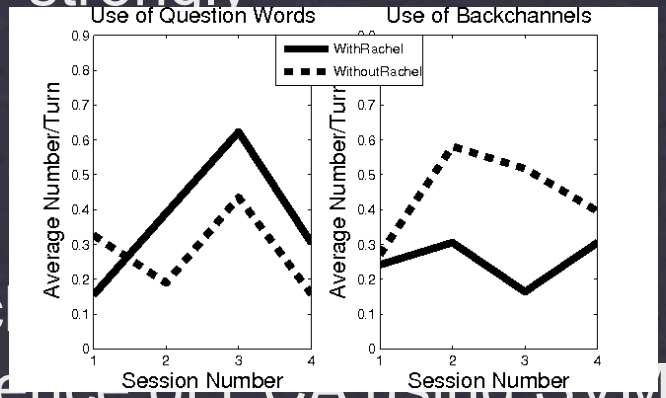
ECA Effect on Speech Behavior

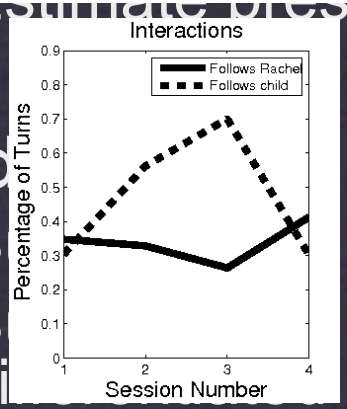
- **Speech pattern hypotheses (without ECA):**
  - Child patterns will not differ strongly
  - differ
  - res (pitch
  - Estimate presence vs. absence of ECA using SVM
- **Find**
  - S ld speech less diffe an parent
  - S ld speech not differ rent di 50% of the time

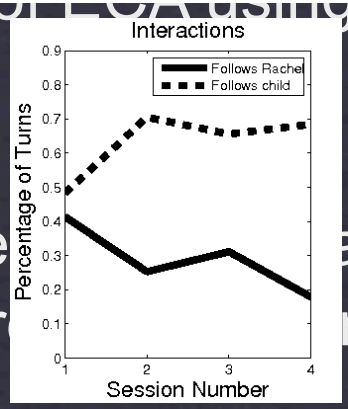
SUBJECT 1: PARENT BEHAVIOR



SUBJECT 2: PARENT BEHAVIOR







Remarks

- **Given the experimental setup, it is practical to extract:**
  - Prosodic information (pitch, energy of speech)
  - Spectral information (measure of frequency content of speech)
  - Word Usage statistics (e.g., number of backchannels)

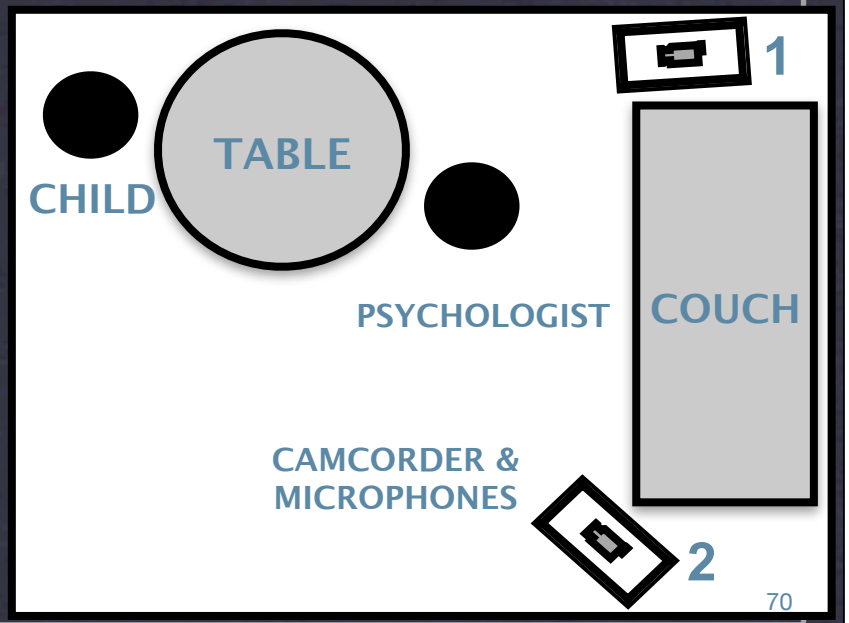
- **The results demonstrate that in general:**
  - The parent's audio data provides discriminatory information regarding the ECA presence or absence
  - The child's audio data does not

Suggests: similarity in the child's communicative behavior across the two interaction conditions, with and without the ECA present

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.

USC CARE Corpus

- **Child-psychologist interactions of children with autism in the context of Autism Diagnostic Observation Schedule (ADOS)**
  - ADOS is a widely used clinical research instrument, comprised of semi-structured social activities to provide psychologist with sample of behavior (30-60 minutes)
  - Psychologist then rates child on autism-relevant symptoms and provides final classification of autism/ASD
- **USC CARE Corpus**
  - *CARE*: Center for Autism Research in Engr.
  - *Collaboration*: USC Keck School of Medicine
  - Recorded ADOS of 60+ children (goal = 100)
  - Controlled, clinical environment
  - 2 HD camcorders + 2 far-field microphones
  - Access to psychologists' notes and codes



USC CARE Corpus

- **Challenges of using the ADOS for psychologists**
  - Qualitative descriptions of assessments
  - Lack of continuous quantitative measures
  - Subjective nature inherent to rating system
- **Engineering-related modeling goals**
  - Investigate how technology can enhance this difficult observation rating task
    - analyze the children's language use and turn-taking trends
    - automate assessment of children's prosody
    - combining data-driven and expert-inspired knowledge

M. Black, D. Bone, M. Williams, P. Gorrindo, P. Levitt, and S. Narayanan. The USC CARE Corpus: Child-Psychologist Interactions of Children with Autism Spectrum Disorders. In Proceedings of InterSpeech, 2011.

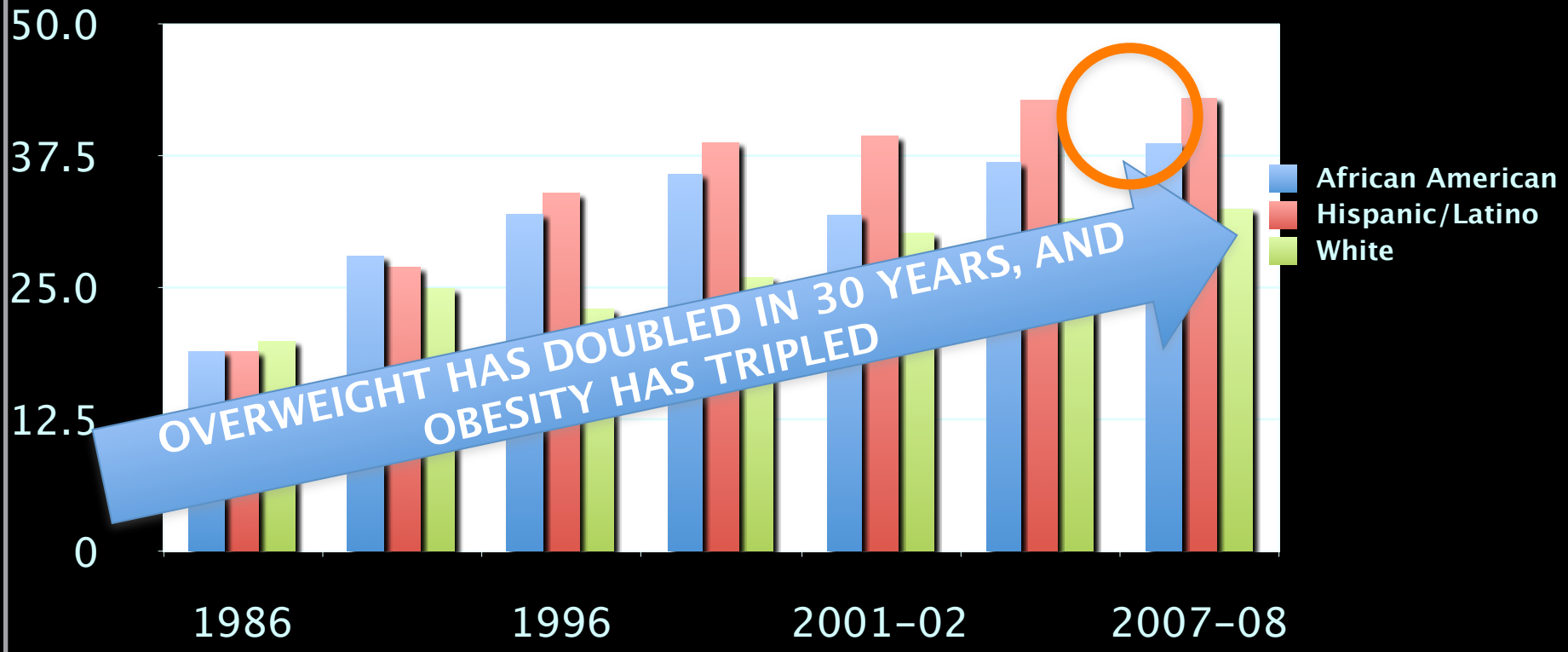
THIS TALK

- Some BSP building blocks

Example Behavioral Analysis Studies

- Family Studies: Marital couples
  - Blame patterns; positiveness/negativeness; humor/sarcasm
- Autism Spectrum Disorders
  - Characterizing joint attention; quantifying socio-emotional discourse
  - Technology interfaces for elicitation and personalized interventions
- Metabolic Health Monitoring
  - Characterizing physical behavior in context

BMI ≥ 85th Percentile (age 4-19)



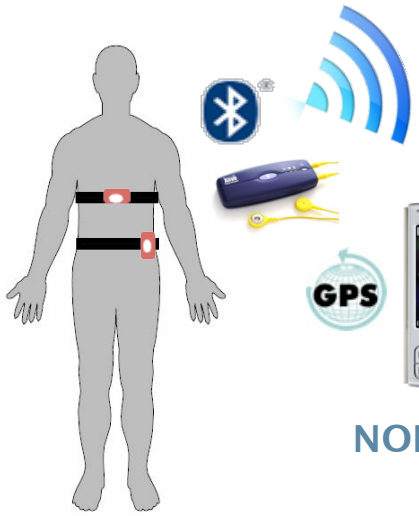
- ***PERSISTS INTO ADULTHOOD (WHITAKER ET AL. NEJM: 1997;337:869-873)***

DATA FROM THE NATIONAL LONGITUDINAL SURVEY OF YOUTH 1986-1998, NHANES 1998-2008

KNOWME Networks

- A suite of mobile, Bluetooth-enabled, wireless, wearable sensors
- That interface with a mobile phone and secure server
- To process data in real time,
- Designed specifically for use in overweight minority youth

KNOWME Networks: Current System



2 ALIVE HEART RATE MONITOR/ ACCELEROMETERS

(ECG/ACC)

END-TO-END ENCRYPTION OF SENSITIVE DATA

3G
GSM
WI-FI

NOKIA N95

WEB SERVER

CHECK RIGHT TO USE SYSTEM
FILTER NOISY UPDATES
WEB ENABLE DATA ACCESS

SQL WITH ENCRYPTED DATA

DATABASE SERVER

DECRYPTION KEYS

DATA WITH SECURE CONNECTION

DOCTOR

PHYSICAL ACTIVITY, STRESS, COGNITIVE BEHAVIOR IN REALTIME

KNOWME NETWORKS

IN-LAB & FIELD DEVELOPMENT: BEHAVIORAL SIGNAL PROCESSING

ADAR EMKEN, MING LI, GAUTAM THATTE, SANGWON LEE, MURALI ANNAVARAM, URBASHI MITRA, SHRIKANTH NARAYANAN, AND DONNA SPRUIJT-METZ. RECOGNITION OF PHYSICAL ACTIVITIES IN OVERWEIGHT HISPANIC YOUTH USING KNOWME NETWORKS. JOURNAL OF PHYSICAL ACTIVITY & HEALTH. 2011.

GAUTAM THATTE, MING LI, SANGWON LEE, ADAR EMKEN, MURALI ANNAVARAM, SHRIKANTH NARAYANAN, DONNA SPRUIJT-METZ, AND URBASHI MITRA. OPTIMAL TIME-RESOURCE ALLOCATION FOR ENERGY-EFFICIENT PHYSICAL ACTIVITY DETECTION. IEEE TRANSACTIONS ON SIGNAL PROCESSING. 2011.

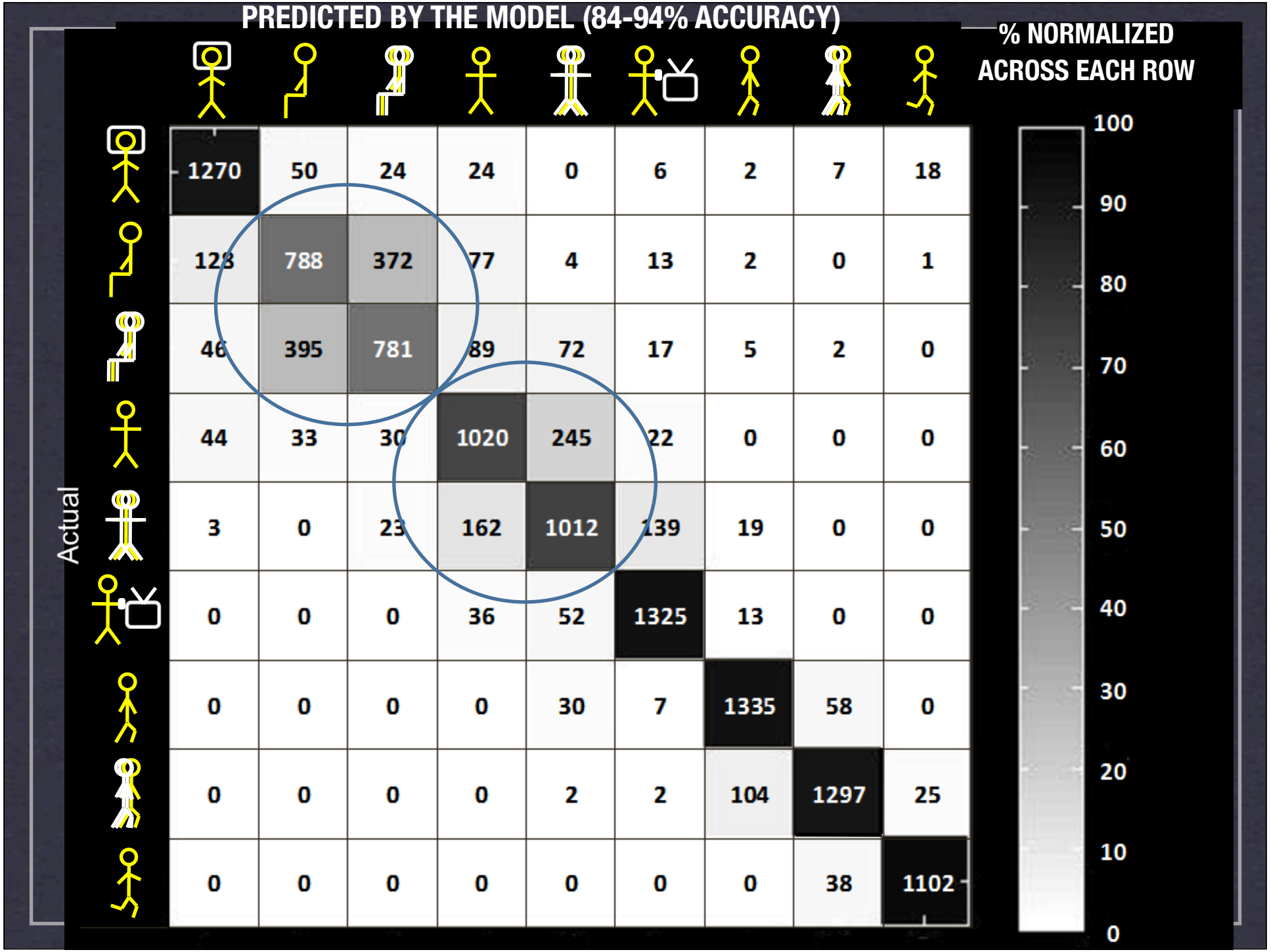
GAUTAM THATTE, MING LI, SANGWON LEE, ADAR EMKEN, SHRIKANTH NARAYANAN, URBASHI MITRA, DONNA SPRUIJT-METZ AND MURALI ANNAVARAM. KNOWME: AN ENERGY-EFFICIENT, MULTIMODAL BODY AREA NETWORK FOR PHYSICAL ACTIVITY MONITORING. ACM TRANSACTIONS ON EMBEDDED COMPUTING SYSTEMS. 2011.

MING LI, VIKTOR ROZGIC, GAUTAM THATTE, SANGWON LEE, ADAR EMKEN, MURALI ANNAVARAM, URBASHI MITRA, DONNA SPRUIJT-METZ AND SHRIKANTH NARAYANAN. MULTIMODAL PHYSICAL ACTIVITY RECOGNITION BY FUSING TEMPORAL AND CEPSTRAL INFORMATION. IEEE TRANSACTIONS ON NEURAL SYSTEMS AND REHABILITATION ENGINEERING. 18(4): 369-380, 2010.

In-Lab Physical Activity Detection

- Developed and tested algorithms to classify PA types in 20 overweight Hispanic youth
  - 10 F/10M; 14.6 ± 1.8 years old; BMI %tile 96 ± 4
- Protocol: 9 activities, 7 minutes/activity

- 3 sessions develop algorithm/1 session test algorithm

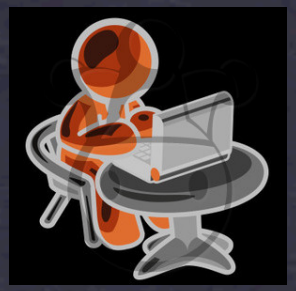

PREDICTED BY THE MODEL (84-94% ACCURACY)
% NORMALIZED ACROSS EACH ROW
Actual
1270 50 24 24 0 6 2 7 18
128 788 372 77 4 13 2 0 1
46 395 781 89 72 17 5 2 0
44 33 30 1020 245 22 0 0 0
3 0 23 162 1012 139 19 0 0
0 0 0 36 52 1325 13 0 0
0 0 0 0 30 7 1335 58 0
0 0 0 0 2 2 104 1297 25
0 0 0 0 0 0 0 38 1102
100 90 80 70 60 50 40 30 20 10 0

Will/Can youth use this system in the real?

- **Subjects: 12 overweight Hispanic youth**
  - 5F/7M; 14.8 ± 1.9 years old; BMI %tile 97 ± 3
- **Protocol:**

IN-HOME TRAINING

WEAR KNOWME FOR 2 DAYS



REMOTE MONITORING

TROUBLE-SHOOTING VIA TEXT

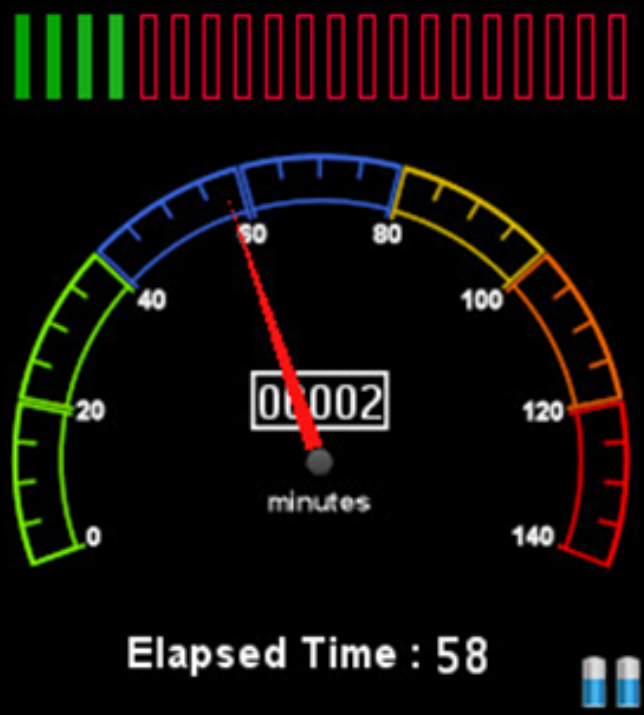
EXIT INTERVIEW

- **Results/day**
  - Wore KNOWME for 11.4 ± 2.0 hours
  - Phone battery life 9.2 ± 2.6 hours
  - 8 SMS sent to us / 9 SMS received from us

Sedentary Intervention

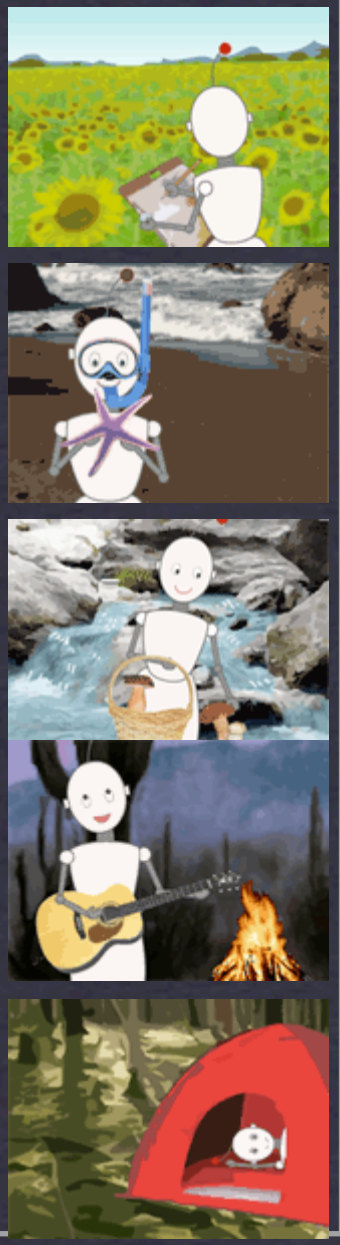
The intervention:

- When the gauge reaches 120 minutes of sedentary time, the phone will automatically begin delivering “Move!!” messages.
- The sedentary gauge will automatically reset to 0 minutes:
  - Following 10 minutes of active time within a 60 minute period; or
  - One hour after 140 minutes of sedentary time is reached if participant doesn’t respond (time out)
  - Researchers are notified at each reset.
- Personalized text messages can be sent from the website monitoring team



KNOWME Networks:
Next Steps

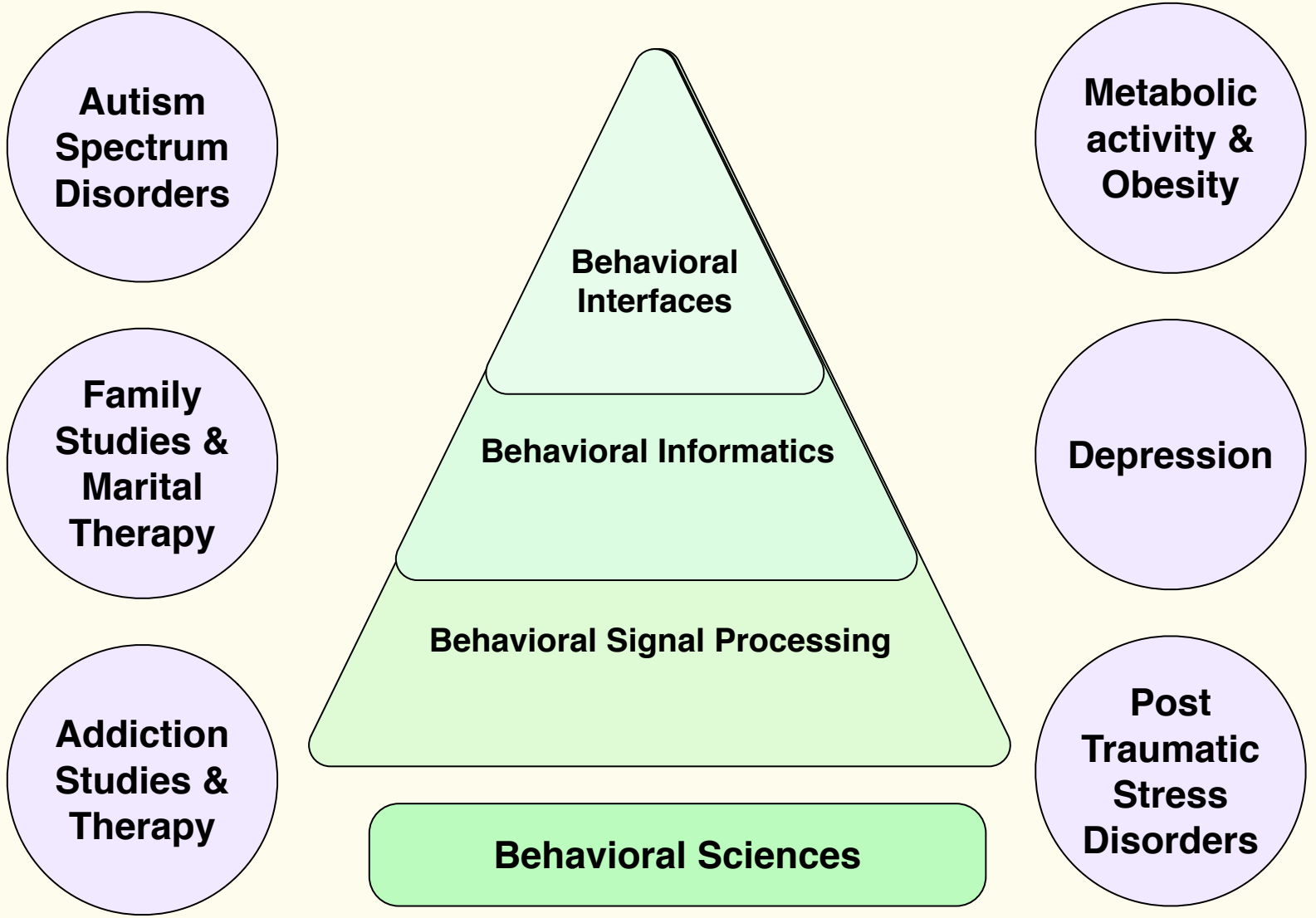
- Social Networking
- Dietary intake (SeeMe)
- Biological data such as body fat, blood glucose
- Mobile games for health (Wellness Partners)
- Web-based interactive games for health
- Immersive environments: Virtual environments for virtual reality, augmented reality, mixed reality platforms
- Move KNOWME to new health frontiers (f.i. Disability, Rehab, Elder health)



TALK SUMMARY:
Open Challenges => R&D Opportunities

- Robust capture and processing of multimodal signals
- Capturing behavior in ecologically valid ways
- Behavior representations for computing
- Reflecting multiple (diverse) perspectives and subjectivity
- Feature-behavior correspondence: human like processing
- Scientifically and Computationally principled modeling
- Reliably characterizing atypical and disordered patterns
- Data privacy, sharing, and management
- Developing productive partnerships with domain experts

HUMAN-FOCUSED SIGNALS AND SYSTEMS



Concluding Remarks

- **Human behavior can described from a variety of perspectives**
  - Both challenges ***and*** opportunities for signal processing
  - Multimodal data integral to derive and model these features
- **Computational advances: sensing, processing and modeling**
  - Support human and machine decision making
- **Exciting scientific and societal possibilities**
  - Interdisciplinary and collaborative scholarship opportunities

BEHAVIORAL INFORMATICS AND MULTIMEDIA SIGNAL PROCESSING:

✓HELP DO THINGS WE KNOW TO DO WELL MORE EFFICIENTLY, CONSISTENTLY
✓HELP HANDLE NEW DATA, CREATE NEW MODELS TO OFFER UNIMAGINED INSIGHTS

WORK REPORTED REPRESENTS COLLABORATIVE EFFORTS WITH NUMEROUS COLLEAGUES AND COLLABORATORS.

ALL THEIR SUPPORT GRATEFULLY ACKNOWLEDGED

WORK SUPPORTED BY: ONR, ARMY, DARPA, NSF AND NIH

http://sail.usc.edu/