Behavioral Signal Processing: Enabling human-centered behavioral informatics

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

University of Southern California
ISCA: International Speech Communication Association

ISCA started in 1999 jointly with

ESCA (European Speech Communication Association):

ICSLP(International Conference of Spoken Language Processing)

• Purpose:
  to promote Speech Communication Science and Technology, both in the industrial and academic areas, covering all the aspects of Speech Communication: acoustics, phonetics, phonology, linguistics, natural language processing, artificial intelligence, cognitive science, signal processing, pattern recognition, etc.

• ISCA offers a wide range of services;
  in particular Interspeech, ISCA workshops, SIGs (special interest groups) , and Distinguished Lectures.
ISCA Objectives:

• to stimulate scientific research and education,
• to organize conferences, courses and workshops,
• to publish, and to promote publication of scientific works,
• to promote the exchange of scientific views in the field of speech communication,
• to encourage the study of different languages,
• to collaborate with all related associations,
• to investigate industrial applications of research results,
• and, more generally, to promote relations between public and private, and between science and technology.
Seeking a window into the human mind and mental state through engineering advances
BEHAVIORAL SIGNAL PROCESSING:

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

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

Complex and multifaceted

- Intricate mind-brain-body interplay
- Effect of environment and interaction with others

Reflected in and influenced by

- Communication, Social interaction, Personality, Affect,..
- Generation and processing of multimodal cues
- Typical, Atypical, Anomalous and Disordered characterizations

ROLE OF ENGINEERING?
BEHAVIOR ANALYSIS CENTRAL TO MANY ENDEAVORS

.. BOTH IN BASIC RESEARCH AND ACROSS APPLICATION DOMAINS

BEHAVIORAL INFORMATICS

SOCIAL SCIENCE SCHOLARSHIP

USER MODELING, CUSTOMER CARE

LEARNING & TRAINING

DAILY LIFE ACTIVITIES

INTEL, SECURITY & DEFENSE

HEALTH & WELL BEING

ROLE OF ENGINEERING?
Many facets & perspectives: “informatics”

.. what data we need, how to process them, derive constructs for decision making

• **Purpose**
  • Scientific understanding, Technology development, Application design e.g., clinical translation

• **The phenomenon of interest: human data**
  • Behavior Expression, Experience and Judgment

• **Use context**
  • Laboratory to Field studies to Real world environments
  • Contend with associated human and physical constraints
Customer care

Escalating frustration?  (only customer side played)
“Uncertainty” manifests itself through combination of vocal, language, and visual behavioral cues
Distressed couple interactions: marital therapy
Characterizing affective dynamics, humor, blame patterns
Autism Spectrum Disorders
Technologies for Rich Understanding of Expressive Behavior and Interaction

Example: Parent and child creating a story together

Computational Targets
Joint Attention
Turn-taking
Shared enjoyment
Behavioral Synchrony

Speech

Electrodermal activity
Multimodal Behavior Signals

• Provide a window into internal state & processes
  Some overtly expressed and directly observable
  e.g., vocal and facial expressions, body posture

Others, covert
  e.g., heart rate, electrodermal response, brain activity

• Implications for understanding
  ‣ Human information encoding and decoding
  ‣ “Mind-Body” relations
  ‣ People’s judgment of others behavior
Operationally defining Behavioral Signal Processing (BSP)

COMPUTATIONAL METHODS THAT MODEL HUMAN BEHAVIOR SIGNALS

• Manifested In Both Overt And Covert Signals
• Processed And Used By Humans Explicitly Or Implicitly
• Facilitate Human Analysis And Decision Making

OUTCOME OF BSP: "BEHAVIORAL INFORMATICS"

QUANTIFYING HUMAN EXPRESSED BEHAVIOR AND HUMAN "FELT SENSE"
How is technology helping already?

- Significant advances in foundational aspects of behavior modeling: detect, classify and track
  - Audio & Video diarization: who spoke when; doing what,
  - Speech recognition: what was spoken
  - Visual Activity recognition: head pose; face/hand gestures,
  - Physiological signal processing with EKG, GSR,
Example: A whole range of speech/language technology possibilities

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

With varying degrees of technology maturity
So ‘n’ your chest pains have been going on just for two days is that right

What more can we infer beyond words?

Words: So ‘n’ your chest pains have been going on just for two days is that right

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

Rich information beyond words
How is technology helping already?

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

SHIFT TO MODELING MORE ABSTRACT, DOMAIN-RELEVANT HUMAN BEHAVIORS
......NEEDS NEW MULTIMODAL & MODELING APPROACHES
Ongoing Advances: Multifaceted

• Sensing: From Smartrooms to Body area networks (instrumented people in the wild)

• Rich speech/spoken language and video understanding
  • who said what to whom, how and when & where

• Affective computing & Emotion modeling
  • Modeling affective behavior in acted and natural scenarios

• Social signal processing
  • Modeling individual and group social behavior: turn taking, non verbal cues such as smiles, laughters and sighs, head nods, proxemics, ...

ALL THESE ARE ESSENTIAL BSP BUILDING BLOCKS:
“LOW & MID LEVEL BEHAVIORAL DESCRIPTORS”
Behavior Coding: Humans in the loop

• Support-than supplant-human (expert) analyses

HUMAN BEHAVIOR OR INTERACTION OF INTEREST
(E.G., CHILD INTERACTING WITH A TEACHER)

Direct Observation

AVAILABLE DATA (E.G., AUDIO, VIDEO, TEXT, PHYSIOLOGICAL)

HUMAN EVALUATOR

JUDGMENTS
(E.G., WHEN IS THE CHILD UNCERTAIN?)

DATA CODING

SIGNAL PROCESSING
(E.G., FEATURE EXTRACTION)

COMPUTATIONAL MODELING
(E.G., MACHINE LEARNING)

FEEDBACK

SCREENING, DIAGNOSTICS, INTERVENTIONS

BEHAVIORAL INFORMATICS
Behavioral signal processing: Human centered

COMPUTING

OF human action and behavior data

FOR meaningful analysis: timely decision making & intervention (action)

BY collaborative integration of human expertise with automated processing: *support not supplant*

TALK OUTLINE

Some behavioral informatics building blocks

- Focus on multimodal data processing
- Affective Computing as an example

Some Case Studies

- Dyadic Improv interactions: Active analysis
  - Multimodal study of conflictive vs. friendly cases
- Dyadic interaction of distressed couples
  - Marital therapy: Blame patterns; positiveness/negativeness; humor
- Autism Spectrum Disorders
  - Characterizing and quantifying interaction patterns
  - Technology interfaces for personalized interventions
- Addiction
  - Understanding psychotherapy: mechanisms, quality
Multimodal data & processing techniques crucial for computational studies of behavior

affective computing as an example......
speech, language, face, body language, interaction patterns
Multimodal Signal Acquisition & Processing

**Instrumented Environments:** arrays of sensors—microphones, cameras, depth sensors, mocap,..

**Sense the user:**
- Identity/location of speaker; Speech, Visual activity
- Interaction, Turn taking, Back channels, Proxemics

**And the environment:**
- Lab, classroom, clinic, home, playground,..

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

The Call center Corpus

Human-Computer Agent telephone interactions

- Spoken dialog, emotions research
- Natural, spontaneous interactions; limited domain
- Categorical, dimensional ratings
Computing Emotions?

Characterizing *Expression* versus *Experience* versus *Judgment*

Representations for computation:
- Categorical (e.g., happy, sad), Dimensional (arousal, valence, dominance)
- Emotion Profiles to handle non prototypical, blended emotions
- Dynamic descriptions to capture changes in time

LEE & NARAYANAN, TOWARD RECOGNIZING EMOTIONS IN SPOKEN DIALOOGS, IEEE TRANS. SPEECH&AUDIO PROCESSING, 13(2):293-302, 2005
Recognizing frustration

- Call center data
- Classification method
  - Linear discriminant classifier for each information modality
- Modalities
  - Acoustic features
    - prosody, spectral features
  - Language features
    - emotional salience of words
  - Discourse
    - dialog acts

The VAM Corpus
A multimodal corpus of talk show interactions
(Karlsruhe, USC)

• Computational modeling, different annotation perspectives
• Incidental: (Sort of) natural, human interaction based, spontaneous
• Categorical, dimensional ratings
VAM Corpus details

Freely available: [http://emotion-research.net/download/vam](http://emotion-research.net/download/vam)

• Unscripted discussions between talk-show guests
  – German; 47 speakers (11 m/36 f)
  – 893 utterances, average duration: 3.0 s
  – Audio, Video, Faces

• More authentic emotions
  – Many negative emotions
  – Text-free, icon-based evaluation using Self Assessment Manikins
  – Categorical & Dimensional evaluation by 6 German and 17 non German evaluators

MICHAEL GRIMM, KRISTIAN KROSCHEL, AND SHRIKANTH NARAYANAN. THE VERA AM MITTAG GERMAN AUDIO-VISUAL EMOTIONAL SPEECH DATABASE. IN PROC. INTERNATIONAL CONFERENCE ON MULTIMEDIA AND EXPO (ICME), JUNE 2008.
Emotion “Primitives”: Valence, Activation, Dominance

**CATEGORICAL EMOTIONS**

- Emotions are described in terms of... emotion categories
- Usually 2 to 6 classes are distinguished
- Often emotions portrayed by actors

**GRADIENT EMOTIONS**

- Emotions are described as points in a 3D emotion space
- Emotions are estimated on a continuous-valued scale
- Spontaneous emotions are used

---

**Emotion “Primitives”: Valence, Activation, Dominance**

- **Valence:** Positive ↔ Negative
- **Activation:** Excited ↔ Calm
- **Dominance:** Weak ↔ Strong

---

Enriching behavior descriptions further....
“Situated” Interactions & Conversational Computing

- Multimodality
- Interaction dynamics
The USC IEMOCAP Corpus
A multimodal corpus of affective dyadic interactions

• Computational modeling, multimodal perspective
• Laboratory elicited (human interaction based), acted, spontaneous
• Categorical, dimensional ratings

Freely available: http://sail.usc.edu/iemocap
The USC IEMOCAP Database

http://sail.usc.edu/iemocap

- Database of actors’ affective interactions
- Rich variety of emotions and multimodal manifestations in a dyadic interaction setting

- Goals include:
  - Expressive facial expression & speech analysis
  - Emotion recognition
  - Study dyadic interaction dynamics

The USC IEMOCAP Data Description

http://sail.usc.edu/iemocap

- Facial motion capture
  - 63 markers distributed on one actor’s face and hand
  - 3 Vicon Motion Capture Cameras
- Microphone speech
  - Shot gun directional microphones
- Video
  - 2 HD cameras directed at each actor
- Data collection settings
  - Spontaneous improvisations
  - Scripted improvisation based on plays
- 10 participants (5 females, 5 males)
- ~12 hours of data
Modeling gestures/speech interrelation

VOCAL AND VISUAL FEATURES

• Speech
  • Prosodic features: Pitch, energy
  • MFCC coefficients (vocal tract)

• Gestures
  • Head motion
  • Eyebrow
  • Lips
  • Different face regions

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

---

### USING SVM

<table>
<thead>
<tr>
<th></th>
<th>Anger</th>
<th>Sadness</th>
<th>Happiness</th>
<th>Neutral</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anger</td>
<td>0.68</td>
<td>0.05</td>
<td>0.21</td>
<td>0.05</td>
</tr>
<tr>
<td>Sadness</td>
<td>0.07</td>
<td>0.64</td>
<td>0.06</td>
<td>0.22</td>
</tr>
<tr>
<td>Happiness</td>
<td>0.19</td>
<td>0.04</td>
<td>0.70</td>
<td>0.08</td>
</tr>
<tr>
<td>Neutral</td>
<td>0.04</td>
<td>0.14</td>
<td>0.01</td>
<td>0.81</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Anger</th>
<th>Sadness</th>
<th>Happiness</th>
<th>Neutral</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anger</td>
<td>0.79</td>
<td>0.18</td>
<td>0.00</td>
<td>0.03</td>
</tr>
<tr>
<td>Sadness</td>
<td>0.06</td>
<td>0.81</td>
<td>0.00</td>
<td>0.13</td>
</tr>
<tr>
<td>Happiness</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Neutral</td>
<td>0.00</td>
<td>0.04</td>
<td>0.15</td>
<td>0.81</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Anger</th>
<th>Sadness</th>
<th>Happiness</th>
<th>Neutral</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anger</td>
<td>0.95</td>
<td>0.00</td>
<td>0.03</td>
<td>0.03</td>
</tr>
<tr>
<td>Sadness</td>
<td>0.00</td>
<td>0.79</td>
<td>0.03</td>
<td>0.18</td>
</tr>
<tr>
<td>Happiness</td>
<td>0.02</td>
<td>0.00</td>
<td>0.91</td>
<td>0.08</td>
</tr>
<tr>
<td>Neutral</td>
<td>0.01</td>
<td>0.05</td>
<td>0.02</td>
<td>0.92</td>
</tr>
</tbody>
</table>

---

*BUSSO ET AL, ANALYSIS OF EMOTION RECOGNITION USING FACIAL EXPRESSIONS, SPEECH AND MULTIMODAL INFORMATION, ICMI, 2004*
Profile-based Representations of Emotions
Characterizing Ambiguous Emotion Displays

Handling non-prototypical, blended emotions

TRAIN EPs ON:
ANGRY, HAPPY, NEUTRAL, SAD

ANALYZE:
ANGRY, HAPPY, NEUTRAL, SAD, FRUSTRATED

SUPERVISED OR UNSUPERVISED LEARNING VIA CLUSTERING OF THE EMOTION SPACE

EMILY MOWER, MAJA MATARIC AND SHRIKANTH NARAYANAN. A FRAMEWORK FOR AUTOMATIC HUMAN EMOTION CLASSIFICATION USING EMOTIONAL PROFILES. IEEE TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING. 19(5): 1057-1070, 2011
Robust Arousal Estimation: A simple tool

• Simple framework: compares favorably to cross-corpus supervised classification systems

• Tool generalizes and is simple: researchers can use to investigate behavioral hypotheses

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

• Largely unsupervised, only requires “neutral” labels from each speaker
Multimodal turn taking dynamics

Problem

• Incorporate “mutual influence” of interactants in the model

Approach

• Dynamic Bayesian Network: Joint modeling both speakers

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

The USC CreativeIT Corpus
A multimodal corpus of improv dyadic interactions

Freely available: http://sail.usc.edu/improv

• Computational modeling, multimodal (full body mocap) perspective
• Laboratory elicited (human interaction based), acted, spontaneous
• Categorical, dimensional ratings; continuous-time affect ratings

ANGELIKI METALLINOU, ZHAOJUN YANG, CHI-CHUN LEE, CARLOS BUSSO, SHARON CARNICKE AND SHRIKANTH NARAYANAN. THE USC CREATIVEIT DATABASE OF MULTIMODAL DYADIC INTERACTIONS: FROM SPEECH AND FULL BODY MOTION CAPTURE TO CONTINUOUS EMOTIONAL ANNOTATIONS. JOURNAL OF LANGUAGE RESOURCES AND EVALUATION. 2015
The USC Creative IT database

- Multimodal emotional database of theatrical improvisation
- Collaboration between engineering and theater
  - Active Analysis methodology
  - Goal driven improvisations
- Rich variety of emotions and multimodal manifestations in an interaction setting
- Goals include:
  - Expressive body language and speech analysis
  - Emotion recognition
  - Study interaction dynamics
  - Animation of affective full-body virtual agents
  - Study actor’s creativity/quality of performance


**S. M. CARNICKE, STANISLAVSKY IN FOCUS: AN ACTING MASTER FOR THE TWENTY-FIRST CENTURY, 1998**
Creative IT: Data Description

http://sail.usc.edu/improv

- Full body Motion capture
  - 45 markers distributed on each actor’s body
  - 12 camera VICON Motion Capture system
- Microphone speech
  - Close-talking microphones
- Video
  - 2 HD cameras at each end of the room
- Data collection settings
  - 2sentence exercises, varying actor goals
  - Paraphrases of theatrical plays, varying actor goals
- 16 participants (9 females, 7 males)
- 50 improvisation recordings ranging 2-5mins long
Creative IT: Continuous Annotations

- Annotators agree on the **trends**
  - rather than actual attribute values
- Easier to rate emotions in relative rather than absolute terms
- We define agreement in terms of correlation metrics
  - \( \sim 0.6 \) median inter-annotator correlation
  - Challenging rating task
- We mostly focus on trends of attributes

---

**ANGELIKI METALLINOU, ATHANASSIOS KATSAMANIS, YUN WANG, AND SHRIKANTH NARAYANAN.** TRACKING CHANGES IN CONTINUOUS EMOTION STATES USING BODY LANGUAGE AND PROSODIC CUES. IN PROCEEDINGS OF IEEE INTERNATIONAL CONFERENCE ON AUDIO, SPEECH AND SIGNAL PROCESSING (ICASSP), MAY, 2011

**ANGELIKI METALLINOU AND S. NARAYANAN,** ANNOTATION AND PROCESSING OF CONTINUOUS EMOTIONAL ATTRIBUTES: CHALLENGES AND OPPORTUNITIES, IN: 2ND INTERNATIONAL WORKSHOP ON EMOTION REPRESENTATION, ANALYSIS AND SYNTHESIS IN CONTINUOUS TIME AND SPACE (EMOSPACE), 2013
The USC Creative IT database
Freely available: http://sail.usc.edu/improv

Continuous rating by three different annotators
Activation of Male Actor
Body Language Feature Extraction

Front View  Back View

Person A
hand position
(xh,yh,zh)
local system

Person B

|V| : absolute velocity
V' : relative velocity of A towards B

Person A

angle

global system

x

y

z

45
ANGELIKI METALLINOU, ATHANASIOS KATSAMANIS AND SHRIKANTH NARAYANAN. TRACKING CONTINUOUS EMOTIONAL TRENDS OF PARTICIPANTS DURING AFFECTIVE DYADIC INTERACTIONS USING BODY LANGUAGE AND SPEECH INFORMATION. JOURNAL IMAGE AND VISION COMPUTING. 31(2): 137-152, FEBRUARY 2013

ZHAOJUN YANG, ANGELIKI METALLINOU AND SHRIKANTH S. NARAYANAN. ANALYSIS AND PREDICTIVE MODELING OF BODY LANGUAGE BEHAVIOR IN DYADIC INTERACTIONS FROM MULTIMODAL INTERLOCUTOR CUES. IEEE TRANSACTIONS ON MULTIMEDIA. 16(6): 1766-1778, 2014.
TALK OUTLINE

Some behavioral informatics building blocks

• Focus on multimodal data processing
• Affective Computing as an example

Some Case Studies

• Dyadic Improv interactions: Active analysis
  • Multimodal study of conflictive vs. friendly cases

• Dyadic interaction of distressed couples
  • Marital therapy: Blame patterns; positiveness/negativeness; humor

• Autism Spectrum Disorders
  • Characterizing and quantifying interaction patterns
  • Technology interfaces for personalized interventions

• Addiction
  • Understanding psychotherapy: mechanisms, quality
Modeling of Body Language Behavior from Multimodal Interaction Cues

• Focus on interactions where friendly versus conflictive stances are taken
• Expression of private internal state of attitude in multimodal cues

ZHAOJUN YANG, ANGELIKI METALLINOU, SHRIKANTH NARAYANAN, TOWARD BODY LANGUAGE GENERATION IN DYADIC INTERACTION SETTINGS FROM INTERLOCUTOR MULTIMODAL CUES, IN: PROCEEDINGS OF ICASSP, 2013

ZHAOJUN YANG, ANGELIKI METALLINOU AND SHRIKANTH S. NARAYANAN. ANALYSIS AND PREDICTIVE MODELING OF BODY LANGUAGE BEHAVIOR IN DYADIC INTERACTIONS FROM MULTIMODAL INTERLOCUTOR CUES. IEEE TRANSACTIONS ON MULTIMEDIA. 16(6): 1766-1778, 2014.

ANGELIKI METALLINOU, ATHANASIOS KATSAMANIS AND SHRIKANTH NARAYANAN. TRACKING CONTINUOUS EMOTIONAL TRENDS OF PARTICIPANTS DURING AFFECTIVE DYADIC INTERACTIONS USING BODY LANGUAGE AND SPEECH INFORMATION. JOURNAL IMAGE AND VISION COMPUTING. 31(2): 137-152, 2013
Objectives

• Uncover the coordination patterns of dyad’s behavior

Friendly interactions

Conflictive interactions

• Computationally model the dyadic coordination

Body language

Speech

Body language
Dyad’s Behavior Pair Setup

- The window context of the interlocutor
- The target feature of an interaction participant

The interlocutor

- Speech features
- Body language

The target participant

- Target body language feature

multimodal feature window

Context (PCA)

Dimensionality reduction with PCA

t

t
Feature Extraction

- Audio features
  - Pitch, Energy, MFCC (30 ms frames with 16.67 ms shift)

- Body language features
  - Using MoCap data
  - Motivated by psychology literature [Harrigan et al.,’05]
  - Computed by defining local and global coordinate systems
  - Of 24 total features 8 selected as target: body posture, velocity, orientation and relative motion toward the interlocutor

Local system

Global system

Face angle computation

Creative-IT Multimodal Database Subset
Body Language Modeling: GMM-based Statistical Mapping

Predicting target body language features from interlocutor’s multimodal cues

\( x_t \): the interlocutor feature vector at time \( t \)

\( y_t \): the participant target feature at time \( t \)

GMM modeling the dyad’s behavior pair:

\[
P(x_t, y_t | \theta^{(x,y)}) = \sum_{m=1}^{M} \pi_m N(x_t, y_t; \mu_m^{(xy)}, \Sigma_m^{(xy)})
\]

Captures generative relationship of feature streams

Statistical mapping using MLE [Toda et al.,’08]:

Estimated Target feature \( \hat{y}_t = \arg \max_{y_t} P(y_t | x_t, \theta^{(x,y)}) \)

GMM: Gaussian Mixture Models
MLE: Maximum Likelihood Estimation
Predicted Body Language Trajectories

![Graphs showing predicted body language trajectories](image)

- Ground truth
- SVR
- MLE
- Fisher

**Fisher Kernel Approach:** Good trend tracking & Estimation of values
Prediction Results: Correlation
how well the trends of body language trajectories are captured

Friendly interactions

Conflictive interactions

velocity features of arms and body better predicted

Lower than friendly interactions
higher predictability of body language orientation in the friendly case → increased adaptation to interlocutor behavior

Increased behavior coordination in positive affect (also seen in vocal entrainment)
Summary

• Quantitative analysis of coordination between dyad’s body language

• Coordination patterns systematically differ for friendly and conflict stances in interactions

• Predictability of body language from interlocutor multimodal cues

• Future directions
  • Body movement animation for a virtual agent based on human user’s attitude and input
TALK OUTLINE

Some behavioral informatics building blocks

- Focus on multimodal data processing
- Affective Computing as an example

Some Case Studies

- Dyadic Improv interactions: Active analysis
  - Multimodal study of conflictive vs. friendly cases

- Dyadic interaction of distressed couples
  - Marital therapy: Blame patterns; positiveness/negativeness; humor

- Autism Spectrum Disorders
  - Characterizing and quantifying interaction patterns
  - Technology interfaces for personalized interventions

- Addiction
  - Understanding psychotherapy: mechanisms, quality
Couple therapy
Characterizing affective dynamics, blame patterns
Corpus

- Real couples in 10-minute problem-solving interactions
- Longitudinal study at UCLA and UW [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
Manual Coding by Human Experts

- Each spouse evaluated by 3-4 trained coders
- 33 session-level codes (all on 1 to 9 scale)
- No utterance- and turn-level ratings
- Social Support Interaction Rating System
- Couples Interaction Rating System
- All evaluators underwent a training period to standardize the coding process

- Analyzed 6 codes for initial studies
  - Level of acceptance ("acc")
  - Level of blame ("bla")
  - Global positive affect ("pos")
  - Global negative affect ("neg")
  - Level of sadness ("sad")
  - Use of humor ("hum")

---

**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. … "

---

**Code Correlation**

<table>
<thead>
<tr>
<th>Code</th>
<th>acc</th>
<th>bla</th>
<th>pos</th>
<th>neg</th>
<th>sad</th>
</tr>
</thead>
<tbody>
<tr>
<td>acc</td>
<td>0.647</td>
<td>0.751</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bla</td>
<td>-0.80</td>
<td>0.470</td>
<td>0.788</td>
<td></td>
<td></td>
</tr>
<tr>
<td>pos</td>
<td>0.670</td>
<td>-0.54</td>
<td>0.667</td>
<td>0.740</td>
<td></td>
</tr>
<tr>
<td>neg</td>
<td>-0.77</td>
<td>0.72</td>
<td>0.690</td>
<td>0.798</td>
<td></td>
</tr>
<tr>
<td>sad</td>
<td>-0.18</td>
<td>0.19</td>
<td>-0.18</td>
<td>0.36</td>
<td>0.315</td>
</tr>
<tr>
<td>hum</td>
<td>0.330</td>
<td>-0.20</td>
<td>0.47</td>
<td>-0.29</td>
<td>-0.15</td>
</tr>
</tbody>
</table>
Automatic Behavior Coding: Estimate behavioral codes from data

Multimodal Signals

- Acoustic
  - Voice activity
  - Pitch
  - Energy
- Lexical
  - Words
  - Topic
  - Word boundaries
  - Fragments
- Visual
  - Head orientation
  - Body orientation
  - Velocity of arms
  - Openness of posture
Focus on extreme cases of session-level judgments

Sample codes:

- acceptance
- blame
- positive affect
- negative affect
- sadness
- humor

M. Black, et al  “Automatic classification of married couples’ behavior using audio features” - Interspeech 2010

(Very) Simple Acoustic-feature based Behavior Estimation

- **Use of acoustic low-level descriptors (LLDs)**
  - Binary classification task
  - Linear-SVM
  - Global speaker-dependent cues capture evaluators’ codes well
  - Capture relevant speech properties of spouses: every 10 ms:
    - Prosody (pitch, energy), spectral (MFCCs), voice quality (jitter, shimmer)
    - Separate features for each spouse (wife, husband)
## Lexical-information based Behavior Code Estimation

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

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
Informing experts

- Automated lexical analysis can inform experts
  - Example: Words that contributed to (correct) classification of a partner as “blaming”

<table>
<thead>
<tr>
<th>Word</th>
<th>High Blame</th>
<th>Low Blame</th>
</tr>
</thead>
<tbody>
<tr>
<td>YOU</td>
<td>-9.61</td>
<td>UM</td>
</tr>
<tr>
<td>YOUR</td>
<td>-4.06</td>
<td>THAT</td>
</tr>
<tr>
<td>ME</td>
<td>-2.53</td>
<td>I</td>
</tr>
<tr>
<td>TELL</td>
<td>-1.51</td>
<td>WE</td>
</tr>
<tr>
<td>ACCEPT</td>
<td>-1.45</td>
<td>THINK</td>
</tr>
</tbody>
</table>

**Most blaming words in terms of discriminative contribution**

<table>
<thead>
<tr>
<th>Word</th>
<th>Δ log</th>
<th>Word</th>
<th>Δ log prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>YOU</td>
<td>1.14</td>
<td>UM</td>
<td>6.01</td>
</tr>
<tr>
<td>YOUR</td>
<td>1.21</td>
<td>THAT</td>
<td>2.67</td>
</tr>
<tr>
<td>ME</td>
<td>1.53</td>
<td>I</td>
<td>2.57</td>
</tr>
<tr>
<td>TELL</td>
<td>1.55</td>
<td>WE</td>
<td>2.36</td>
</tr>
<tr>
<td>ACCEPT</td>
<td>1.56</td>
<td>THINK</td>
<td>2.07</td>
</tr>
</tbody>
</table>

**Least blaming words in terms of discriminative contribution**

<table>
<thead>
<tr>
<th>Word</th>
<th>Δ log prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>YOU</td>
<td>1.14</td>
</tr>
<tr>
<td>YOUR</td>
<td>1.21</td>
</tr>
<tr>
<td>ME</td>
<td>1.53</td>
</tr>
<tr>
<td>TELL</td>
<td>1.55</td>
</tr>
<tr>
<td>ACCEPT</td>
<td>1.56</td>
</tr>
</tbody>
</table>

**Word log prob**

| YOU     | -95.49    |
| YOUR    | -85.88    |
| ME      | -9.61     |
| TELL    | -4.06     |
| ACCEPT  | -2.53     |
| ACCEPT  | -1.45     |
| CARING  | -3.26     |
| EXPECTS | -16.70    |
| CONSIDERATION | -16.11 |
| ME      | -40.27    |
| KNOW    | -4.06     |
| TELL    | -1.51     |
| TOLD    | -1.77     |
| NOT     | -40.32    |
| WHAT    | -50.77    |
| INTIMACY | -43.16   |
| IT      | -42.70    |
Example Fusion Results: Estimating “Blame”

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

<table>
<thead>
<tr>
<th>Classifier Type</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline Chance</td>
<td>50%</td>
</tr>
<tr>
<td>Language</td>
<td>75.4%</td>
</tr>
<tr>
<td>Acoustic</td>
<td>79.6%</td>
</tr>
<tr>
<td>Fusion</td>
<td>82.1%</td>
</tr>
</tbody>
</table>

• REMARKS

Lower performance of language classifier due to (our) ASR issues
Fusion advantageously uses language and acoustic information
Feasible to model high-level behaviors with automatically derived speech and language information
Some technical challenges & approaches..

• Any single feature stream offers partial, noisy code information
  ➡ Multimodal approach, Context sensitive learning

• 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

• Behavior is a part of an interaction: mutual interlocutor dependency
  ➡ Models of entrainment

• Not all human observers/evaluators are equally reliable, and reliability is data dependent
  ➡ Realistic models of human observers/evaluators
Behavior Collection Space:

A new multichannel multimodal database

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
Head motion modeling for behavior analysis

- **Head motion**
  - Important nonverbal behavior cues
  - Nods & shakes are common

- **Data driven modeling**
  - Optical flow of head motion
  - Motion segmentation
  - LSF representation
  - GMM clustering
  - Predict expert annotated behavior codes
  - Binary classification: ensemble of GMMs

---

BO XIAO, PANAYIOTIS GEORGIOU, BRIAN BAUCOM, SHRIKANTH NARAYANAN, HEAD MOTION SYNCHRONY AND ITS CORRELATION TO AFFECTIVITY IN DYADIC INTERACTIONS, IN PROCEEDINGS OF THE IEEE INTERNATIONAL CONFERENCE ON MULTIMEDIA & EXPO (ICME), 2013
Low Level Descriptor Features

- **Audio features**
  - pitch, energy, MFCCs
  - speaker segmentation using VAD/Mic array
- **Motion capture based features**
  - Head/body orientation relative to the other subject
  - 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**
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

• Ordinal SVM captures inherent order information of A-A codes
• Difference of confusion matrices show the advantage of ordinal regression method

V. ROZGić, B. XIAO, A. KATSAMANIS, B. BAUCOM, P. G. GEORGIou, AND S. NARAYANAN. ESTIMATION OF ORDINAL APPROACH-AVOIDANCE LABELS IN DYADIC INTERACTIONS: ORDINAL LOGISTIC REGRESSION APPROACH. IN PROCEEDINGS OF ICASSP 2011
Some technical challenges & approaches..

- Any single feature stream offers partial, noisy code information
  ➡  Multimodal approach, Context sensitive learning

- 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

- Behavior is a part of an interaction: mutual interlocutor dependency
  ➡  Models of entrainment

- Not all human observers/evaluators are equally reliable, and reliability is data dependent
  ➡  Realistic models of human observers/evaluators
Multiple Instance Learning

EACH SPEAKER-TURN IS AN INSTANCE (OF BEHAVIOR)

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

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

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

Preprocessing
Raw Acoustic Features
Speech-text Alignment

Representative Acoustic and Lexical Features Selection and Extraction

Lexical Features Extraction

Vocal Features Extraction

Compute Saliency

Classify

Estimate distance of each session from the salient prototypes

Identify speaker turns that appear to make the difference, i.e., maximize diverse density

multiple instance learning
Saliency Detection with Multiple Instance Learning

SALIENT PROTOTYPES: INSTANCES CLOSE TO POSITIVE BAGS AND FAR AWAY FROM NEGATIVE BAGS


Audio & Visual Salient Features

Classification accuracy (%) using audio, visual, and audio-visual fusion

<table>
<thead>
<tr>
<th>behavior</th>
<th>audio</th>
<th>visual</th>
<th>fusion</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>early</td>
<td>late</td>
<td></td>
</tr>
<tr>
<td>acceptance</td>
<td>70.5</td>
<td>62.5</td>
<td>64.3</td>
</tr>
<tr>
<td>blame</td>
<td>69.4</td>
<td>57.4</td>
<td>70.4</td>
</tr>
</tbody>
</table>

Late fusion improves accuracy for classification of both behaviors

JAMES GIBSON, BO XIAO, PANAYIOTIS GEORGIOU, SHRIKANTH NARAYANAN, AN AUDIO-VISUAL APPROACH TO LEARNING SALIENT BEHAVIORS IN COUPLES’ PROBLEM SOLVING DISCUSSIONS, IN PROCEEDINGS OF THE IEEE INTERNATIONAL CONFERENCE ON MULTIMEDIA & EXPO (ICME), 2013
Some technical challenges & approaches..

• Any single feature stream offers partial, noisy code information
  ➡ Multimodal approach, Context sensitive learning

• 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

• Behavior is a part of an interaction: mutual interlocutor dependency
  ✓ 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

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]

Quantification of Prosodic Entrainment

Signal-derived quantitative measure

“How do two people sound alike as they interact in a conversation?”
Computing Vocal Entrainment: A novel measure
“HOW MUCH DO TWO PEOPLE SYNCHRONIZE IN A CONVERSATION?”

Preprocessing
Raw Acoustic Features
Speech-text Alignment

Vocal Features Extraction

PCA Vocal Characteristics Space

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

Compute Similarity Metric

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

Vocal Entrainment Measures

Predict Affect

Representative Acoustic Features Identification and Parameterization

CHI-CHUN LEE, ATHANASIOS KATSAMANIS, MATTHEW P BLACK, BRIAN R BAUCOM, ANDREW CHRISTENSEN, PANAYIOTIS G GEORGIOU AND SHRIKANTH S NARAYANAN. COMPUTING VOCAL ENTRAINMENT: A SIGNAL-DERIVED PCA-BASED QUANTIFICATION SCHEME WITH APPLICATION TO AFFECT ANALYSIS IN MARRIED COUPLE INTERACTIONS. COMPUTER, SPEECH, AND LANGUAGE. 28(2): 518-539, MARCH 2014
Computing Multi/Cross-modal Entrainment & Synergy

- Computational models of synchrony between head, hand and body gestures and vocal patterns, physiological signals
- Use to
  - characterize behavioral constructs e.g., approach-avoidance, positive affect, empathy,..
  - predict the behavior of the other interactant
Some technical challenges & approaches:

- Any single feature stream offers partial, noisy code information
  ➡ Multimodal approach, Context sensitive learning

- 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

- Behavior is a part of an interaction: mutual interlocutor dependency
  ➡ Models of entrainment

- Not all human observers/evaluators are equally reliable, and reliability is data dependent
  ✓ Realistic models of human observers/evaluators
Fusion of Labels from Multiple Diverse Experts Without Knowing Reference Label

**GLOBALLY VARIANT LOCALLY CONSTANT (GVLC) MODEL**

Model captures data-dependent expert reliability

Experts could be human evaluators and/or machine classifiers.

Developed MAP-EM algorithm for learning unknown parameters.

---

KARTIK AUDHKHASI AND SHRIKANTH NARAYANAN. A GLOBALLY-VARIANT LOCALLY-CONSTANT MODEL FOR FUSION OF LABELS FROM MULTIPLE DIVERSE EXPERTS WITHOUT USING REFERENCE LABELS. IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE. 2013
TALK OUTLINE

Some behavioral informatics building blocks

- Focus on multimodal data processing
- Affective Computing as an example

Some Case Studies

- Dyadic Improv interactions: Active analysis
  - Multimodal study of conflictive vs. friendly cases

- Dyadic interaction of distressed couples
  - Marital therapy: Blame patterns; positiveness/negativeness; humor

- Autism Spectrum Disorders
  - Characterizing and quantifying interaction patterns
  - Technology interfaces for personalized interventions

- Addiction
  - Understanding psychotherapy: mechanisms, quality
Autism Spectrum Disorders (ASD)

• 1 in 68 US children diagnosed with ASD (CDC, 2014)

• 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
  • Stratify phenotyping with quantifiable and adaptable metrics
  • Track, quantify children’s progress during interventions

  Interfaces/systems to elicit, encourage, analyze behavior
  • Complex, but phased; Structured; Naturalistic
Analyzing Interaction in ASD

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

ASD

Language and Interaction
- Echolalia
- Gestures
- Conversation
- Social Response
- Prosodic Abnormalities
- Unusual Preoccupation

Stereotyped Behaviors

Intonation
- Volume
- Rate
- Voice Quality

87
Our Case study Setup

Approach

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

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

Hypotheses

1. Children with ASD will demonstrate correlation between acoustic-prosodic cues and severity of ASD-related impairment
2. Psychologist’s speech is also informative of rated severity (both participant and evaluator)
Quantifying Atypical Prosody

Qualitative descriptions are general and contrasting

“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
Atypical Prosody & Interaction

Spearman’s Correlation between rated severity and prosodic cues

**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 \( p < 0.01 \)

The psychologists may be varying their engagement strategies

### ASD Severity Regression

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

Spearman’s $\rho$ between prediction and labels. [**,****]$\equiv a=[0.01,1e-4]$. $N=28$.

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

**Modeling Interaction Dynamics Critical**

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

---

Quantifying Qualitative Social Perceptions
Atypicality in Facial Expressions of ASD children

Tanaya Guha, Zhaojun Yang, Anil Ramakrishna, Ruth Grossman, Darren Hedley, Sungbok Lee, Shrikanth Narayanan. ON QUANTIFYING FACIAL EXPRESSION-RELATED ATYPICALITY OF CHILDREN WITH AUTISM SPECTRUM DISORDER. In Proceedings of IEEE International Conference on Audio, Speech and Signal Processing (ICASSP), Brisbane, Australia, April, 2015

Understanding the expression and perception of social cues in ASD: What makes the difference?

*Example*: Production of Affective Facial Expressions During Smile Imitation Task

**Computational Targets**
*Quantify atypicality of smile*
Region-based activation
Synchrony & symmetry

---

Tanaya Guha, Zhaojun Yang, Anil Ramakrishna, Ruth Grossman, Darren Hedley, Sungbok Lee, Shrikanth Narayanan. *ON QUANTIFYING FACIAL EXPRESSION-RELATED ATYPICALITY OF CHILDREN WITH AUTISM SPECTRUM DISORDER*. In ICASSP, 2015

ASD and facial expressions

- ASD linked to production of atypical facial expressions and prosody [Asperger, 1944] [Kanner 1968]
- Asynchrony in coordinating speech and gestures [DeMarchena et al., 2010]
- Facial expressions often perceived as ‘atypical’ or awkward [Grossman et al., 2012] [Yirmiya et al., 1989]
  - Awkwardness impression is hard to quantify

- Use MoCap technology and statistical methods to computationally quantify:
  - What causes this impression of awkwardness?
  - What differentiates neurotypical from ASD populations?
- Gain insights that can’t be obtained otherwise
Data and Methodology

- 37 Children (21 with autism, 16 typical), aged 9-14 years
- Emotion Mimicry Tasks
- Detailed Facial Motion Capture
- Functional Data Analysis and Statistical testing
  - Quantify facial expression properties

Example video stimuli

Example subject response
Methods and Findings

• Motion Capture (MoCap)
  • Detailed capture of facial expressions
• Functional Data Analysis (FDA) and statistical testing
  • Model, analyze, discover properties
• Multidimensional Scaling
  • Visualization of subject variability and similarity

• Findings
  • Increased **facial motion asynchrony** for ASD
  • Increased **facial motion roughness**
  • Consistently **greater expression variability**
    • Idiosyncratic face gestures
Child-Parent Synchrony

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

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

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

**APPROACH**

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

**FINDINGS**

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


Affective Synchrony

- Synchrony of vocal arousal during child-psychologist interactions

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

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

- Autism Interventions:
  - Target core domains of the disorder
  - Large heterogeneity results in variability of outcomes
  - **Child therapist fit is important**

- **Emotion regulation:** extrinsic (co-regulation) and intrinsic (self-regulation) processes responsible for monitoring, evaluating and modifying emotional reactions [Thomson 94’, Gulsrud et al. ’10].

- Simultaneous **monitoring** of child’s & therapist **physiology** permits:
  - Exploration of a person’s internal state
  - How internal states interact with observable behavior
  - Using psycho-physiological signals [Goodwin et al. ’06]
APPROACH AND FINDINGS

- EDA is modeled as a sequence of SCRs.
- SCRs form a spike train affected by observable co-existing events
  - Non-homogeneous Poisson Process (PP)
  - Rate function incorporates self/co-regulation instances

Findings within the context of JASPER (Kasari'10):
Regulatory behaviors related to child and therapist EDA.
Results indicate that child and therapist physiological states are
  - interdependent
  - associated with observable behavior
  - different across the various regulation types
TALK OUTLINE

Some behavioral informatics building blocks

• Focus on multimodal data processing
• Affective Computing as an example

Some Case Studies

• Dyadic Improv interactions: Active analysis
  • Multimodal study of conflictive vs. friendly cases

• Dyadic interaction of distressed couples
  • Marital therapy: Blame patterns; positiveness/negativeness; humor

• Autism Spectrum Disorders
  • Characterizing and quantifying interaction patterns
  • Technology interfaces for personalized interventions

• Addiction
  • Understanding psychotherapy: mechanisms, quality
Motivational Interviewing for Addiction

Widely used in psychotherapy
• Client's (interviewee) own will of making a change
• Therapist (interviewer): understand, facilitate, do not dictate
• Goal-oriented, highly-structured
• Non-confrontational, non-judgmental, dialog setting

COMPUTATIONAL BEHAVIOR MODELING POSSIBILITIES

Interview efficacy: Modeling constructs such as “reflections”
Interaction dynamics of interviewer-interviewee: e.g., Empathy:
• Computational language modeling provides useful insights into the expressed empathy behavior of therapists
• Use speech, spoken language, nonverbals, body language cues
• Data from several clinical intervention studies, coded by experts

BO XIAO, DOGAN CAN, PANAYIOTIS G. GEORGIOU, DAVID ATKINS AND SHRIKANTH S. NARAYANAN. ANALYZING THE LANGUAGE OF THERAPIST EMPATHY IN MOTIVATIONAL INTERVIEW BASED PSYCHOTHERAPY. PROCEEDINGS OF APSIPA 2012

DOGAN CAN, PANAYIOTIS GEORGIOU, DAVID ATKINS AND SHRIKANTH NARAYANAN. A CASE STUDY: DETECTING COUNSELOR REFLECTIONS IN PSYCHOTHERAPY FOR ADDICTIONS USING LINGUISTIC FEATURES. IN PROCEEDINGS OF INTERSPEECH. 2012
Addiction Psychotherapy Corpora

• Originally collected for psychotherapy process research
• ~800 audio/video sessions from 5 different, brief intervention studies:
  • HMCBI, ESP21, ESPSB, iCHAMP, ARC
• ~10% (155 sessions) manually transcribed and annotated by trained coders
  • Utterance Level Behavioral Codes (MISC)
  • Session Level Behavioral Codes (MISC, MITI)
  • Outcomes

Motivational Interviewing
Sample (training) video
https://www.youtube.com/watch?v=EvLquWI8aqc
Automatic Empathy Prediction

- Model therapist language when in high vs. low empathy:
  - How is language in these 2 cases different?

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

Bo Xiao, et al., Novel Prototype System for Rating Therapist Empathy from Audio Recordings in Addiction Counseling, in prep
Acoustic Prosodic Patterns

- **Prosody patterns and empathy**
  - Extract, quantify, and model the distribution of prosodic cues
  - *Quantized features*: turn duration, energy, pitch, jitter, shimmer

<table>
<thead>
<tr>
<th>Speaker</th>
<th>S</th>
<th>T</th>
<th>S</th>
<th>P</th>
<th>S</th>
<th>P</th>
<th>S</th>
<th>T</th>
<th>S</th>
<th>P</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duration</td>
<td>M</td>
<td>L</td>
<td>H</td>
<td></td>
<td></td>
<td></td>
<td>L</td>
<td>M</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Energy</td>
<td>H</td>
<td>L</td>
<td>M</td>
<td></td>
<td></td>
<td></td>
<td>M</td>
<td>L</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pitch</td>
<td>H</td>
<td>M</td>
<td>L</td>
<td></td>
<td></td>
<td></td>
<td>M</td>
<td>L</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shimmer</td>
<td>L</td>
<td>L</td>
<td>H</td>
<td></td>
<td></td>
<td></td>
<td>M</td>
<td>M</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jitter</td>
<td>M</td>
<td>M</td>
<td>L</td>
<td></td>
<td></td>
<td></td>
<td>H</td>
<td>L</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Results**: lower perceived empathy of therapist when:
  - Therapist has higher energy values
  - Therapist has higher pitch values
Entrainment Measures

- Link between entrainment measures and perceived empathy
  - Behavior of interlocutors become similar
  - Define similarity metrics on speech-derived properties
  - Found significant correlation: higher entrainment/similarity implies more empathy
Behavioral signal processing: Human centered

COMPUTING

OF

human action and behavior data

FOR

meaningful analysis: timely decision making & intervention (action)

BY

collaborative integration of human expertise with automated processing: support not supplant

HUMANS

TALK SUMMARY:

Open Challenges → RICH R&D Opportunities

- Robust capture and processing of multimodal signals
- Capturing natural 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 provenance, integrity, sharing, and management
- Developing productive partnerships between various domain experts
Concluding Remarks: Enabling Behavioral informatics

- **Human behavior can described from a variety of perspectives**
  - Both challenges *and* opportunities for R&D
  - Multimodal data integral to derive and model these constructs

- **Computational advances: sensing, processing and modeling**
  - Signals and systems approach to human interaction studies
  - Support **BOTH** human and machine decision making

- **Exciting technological and societal possibilities**
  - Opportunities for interdisciplinary and collaborative scholarship
  - Enable broader access, and directly impact directly various walks of life

---

**BEHAVIORAL SIGNAL PROCESSING:**

- ✓ HELP DO THINGS WE KNOW TO DO WELL MORE EFFICIENTLY, CONSISTENTLY
- ✓ HELP HANDLE NEW DATA, CREATE NEW MODELS TO OFFER UNIMAGINED INSIGHTS
  ★ CREATE TOOLS FOR DISCOVERY
Work reported represents collaborative efforts with numerous colleagues and collaborators.

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


http://sail.usc.edu/