Language Technologies Institute

Carnegie Mellon University

Multimodal Affective Computing

Lecture 15: Multimodal Applications

Louis-Philippe Morency
Jeffrey Girard

Originally developed with help from
Stefan Scherer and Tadas Baltrušaitis

Course Review

Week 1 – Lecture 1 - Introduction

- Introductions
- Human multimodal communication
  - Behaviors, multimodal and interpersonal
- Multimodal Affective Computing
  - A historical view
  - Psychological constructs
- Course syllabus and project assignments
  - Grades and course structure
  - Course project

Week 1 – Lecture 2 – Course Project

- Course project assignment
- Resources for the course project
  - Common topics in affective computing
  - Common research questions in affective computing
  - Available multimodal and social video datasets
  - Tools for annotation and feature extraction
- Project discussions
  - Identifying topics of interest
  - Begin forming project groups

Week 2 – Psychological Constructs

- What is a psychological construct?
  - Constructs, indicators, and hierarchies
  - Measurement and construct estimation
- How are constructs commonly measured?
  - Self-report questionnaires
  - Observational and judgment studies
- When is measurement trustworthy?
  - Validity and reliability of measurement
  - An introduction to measurement validation

Week 3 – Psychological Theories

- Theories of affect and emotion
  - Widely accepted aspects and controversies
- Theories of personality
- Theories of psychopathology
- Theories of interpersonal functioning

Week 4 – Visual Messages

- Interpersonal Communication
  - Encoder-Decoder Process, Lens Model
  - Elements of interpersonal communication
- Nonverbal visual messages
  - Facial expressions
  - Eye gaze and mutual contact
  - Proxemics and group formations
  - Gestures and body language
  - Practical tools for automatic sensing

Week 5 – Vocal Messages

- Multiple Layers of Vocal Messages
  - What we can convey with speech
- Fundamentals of speech production and hearing
  - Anatomy of the vocal tract and the physiology of hearing
  - Fundamental speech measures (direct vs. perceptual measures)
- Prosodic manipulation and its meaning
- Use and detection of varying voice quality
- Nonverbal vocal expressions
  - Laughter, pause filler (e.g. uh, um), and moans
- Practical tools for speech signal processing
- Automatic Techniques for visual processing

Week 6 – Verbal Messages

- Linguistics and the study of language
- Word and lexical representations
  - Sentiment and topic analysis
  - LIWC and lexicons
  - Word2vec and word embeddings
- Language structure
  - Grammar, syntax and language models
- Discourse and dialogue analysis
  - Adjacency pairs, common ground
  - Speech and dialogue acts
- Turn-taking and conversation dynamics
  - Overlaps, interruptions, Backchannel, Disfluencies
  - Multi-party floor management
- Practical tools for automatic annotation

Week 7 – Statistical Foundations

1. Exploratory data analysis
2. Statistical hypothesis testing
3. Point estimation and effect sizes
4. Interval estimation and confidence intervals

Week 8 – Statistical Modeling

1. The general linear model (LM)
2. The generalized linear model (GLM)
3. Preview of advanced frameworks
   - Multilevel modeling (MLM)
   - Structural equation modeling (SEM)
   - Regularization and prediction (GLMNET)

Week 10 – Probabilistic Predictive Models

- Basic concepts of machine learning
  - Definitions and types of algorithms
  - Linear regression and classification
- Joint probability distribution
- Probabilistic graphical models
  - Independence and Conditional independence
  - Example: modeling affect during learning
- Bayesian networks
  - Conditional probability distribution
  - Dynamic Bayesian Network
  - Naïve Bayes classifier
- Evaluation methods and error measures

Week 11 – Discriminative Prediction Models

- Dynamic Bayesian Network
  - Hidden Markov Models
  - Factorial and coupled HMMs
- Markov Random Fields
  - Unary, binary and clique potentials
  - Factor graph representation
- Multimodal Machine Learning
  - Core Challenges: Representation, Alignment, Fusion, Translation and Co-Learning
- Discriminative Graphical Models
  - Logistic classifier
  - Conditional random fields
  - L1 and L2 regularization
- Evaluation methods and error measures

Week 12 – Neural Prediction Models

- Discriminative Graphical Models
  - Logistic classifier
  - Conditional random fields
  - L1 and L2 regularization
- Neural Networks
  - Multi-layer perceptron
  - Back-propagation
  - Convolutional neural networks
- Evaluation methods and error measures

- Next week: Multimodal deep learning

Week 13 – Multimodal Deep Learning

- Multimodal core challenges - review
- Multimodal representations
  - Joint and coordinated representations
    - Multimodal autoencoder & tensor fusion
    - Deep canonical correlation analysis
- Multimodal alignment
  - Implicit and explicit alignment
    - Dynamic time warping
    - Attention models
- Multimodal fusion
  - Multi-view recurrent network
  - Memory fusion networks

Week 14 – Multimodal Behavior Generation

- Embodied conversational agents
- Media equation
- Nonverbal communication signals
- Behavior generation
  - Manually generated scripts
  - Rule-based
  - Behavior prediction
  - Joint position prediction
- Communicating with Virtual Agent

Week 15 – Multimodal Applications (Today)

- Multimodal sentiment analysis
- Public speaking training and assessment
- Multimodal agreement recognition
- Multimodal learning analytics
- Health behavior informatics
  - Depression, PTSD and suicidal ideation
- Virtual interviewer and behavior generation

Multimodal Applications

Multimodal Applications

Multimodal Sentiment Analysis

Public speaking training and assessment

(with Stefan Scherer @ USC)

Multimodal Agreement Recogntion

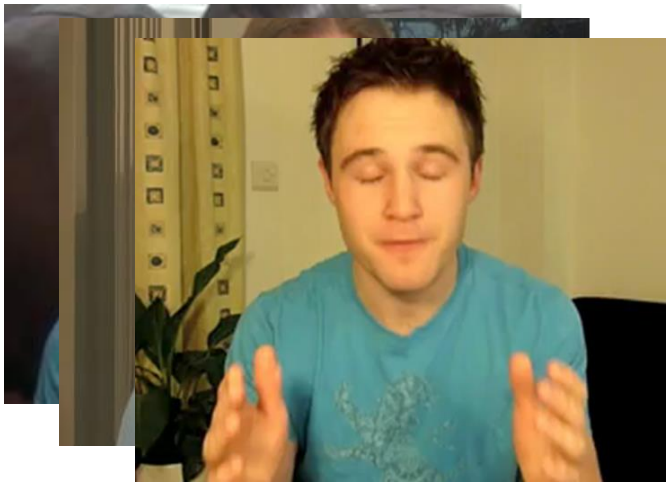
(with Konstantinos Bousmalis and Maja Pantic @ Imperial College)

Multimodal Sentiment Analysis

Learning from the Web

Multimodal

- Audio
- Visual
- Verbal

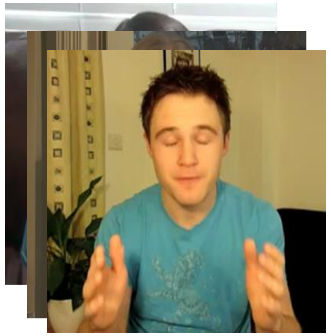


YouTube

- Multiplicity of expressions
- 10,000+ new videos per days
- Verbal, vocal and visual modalities
- Spontaneous and natural behaviors
- Limited motion range (fixed camera)

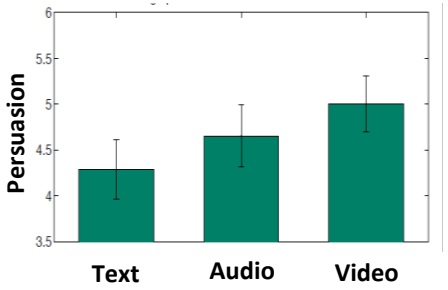
Persuasion in Social Multimedia

YouTube videos

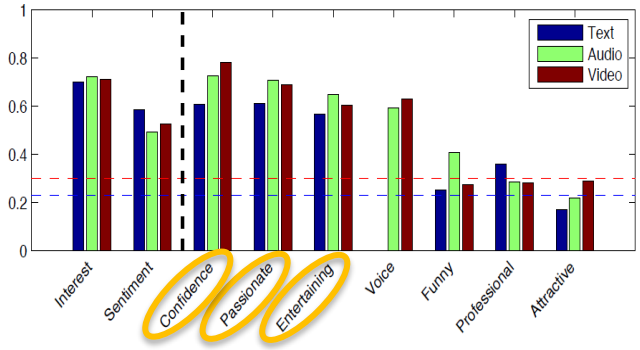


amazon mechanical turk

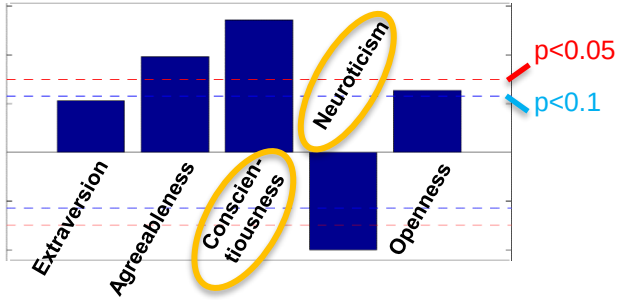
H1 – The communication modality affects persuasiveness of the speaker



H2 – Speaker traits such as confidence, passionate and humoristic correlate with persuasiveness



H3 – Perceived personality traits correlate with the persuasiveness of the speaker



Multimodal Sentiment Analysis

[ICMI 2010, IEEE Intelligent Systems 2012, ACL 2013]

Multimodal

- Audio
- Visual
- Verbal

I. **Complementarity**
II. **Nonlinear fusion**
III. **Multi-stream models**
IV. **Multimodal synchrony**

Utterance-level classification

Modality	Accuracy
Baseline	55.93%
One modality at a time	
Linguistic	70.94%
Acoustic	64.85%
Visual	67.31%
Two modalities at a time	
Linguistic + Acoustic	72.88%
Linguistic + Visual	72.39%
Acoustic + Visual	68.86%
Three modalities at a time	
Linguistic+Acoustic+Visual	74.09%

Video-level classification

Modality	Accuracy
Baseline	55.93%
One modality at a time	
Linguistic	73.33%
Acoustic	53.33%
Visual	50.66%
Two modalities at a time	
Linguistic + Acoustic	72.00%
Linguistic + Visual	74.66%
Acoustic + Visual	61.33%
Three modalities at a time	
Linguistic+Acoustic+Visual	74.66%

Spanish YouTube videos

Modality	**Accuracy**
Text only	64.94%
Visual only	61.04%
Audio only	46.75%
Text-visual	73.68%
Text-audio	68.42%
Audio-visual	66.23%
Text-audio-visual	**75.00%**

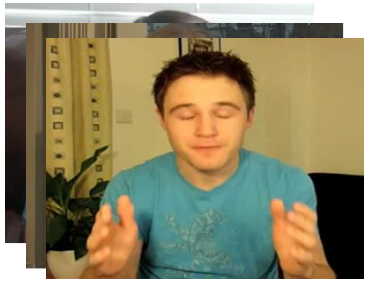
Multimodal Sentiment Analysis

[ICMI 2010, IEEE Intelligent Systems 2012, ACL 2013]

Multimodal

- Audio
- Visual
- Verbal

I. Complementarity
II. Nonlinear fusion
III. Multi-stream models
IV. Multimodal synchrony



YouTube videos

47

Audio
- Pauses
- Pitch

Visual
- Smile
- Gaze

Verbal
- Polarized words

Extracted features

?

- Positive
- Neutral
- Negative

	Positive	Neutral	Negative
Smile intensity	+	-	-
Look at camera	+	-	-
Pause length	-	+	-
Voice pitch level	+	-	+
Polarized words	+	*	-

Online Crowdsourcing Tool for Annotations of Behaviors

[ACM Multimedia 2012 CrowdMM workshop]

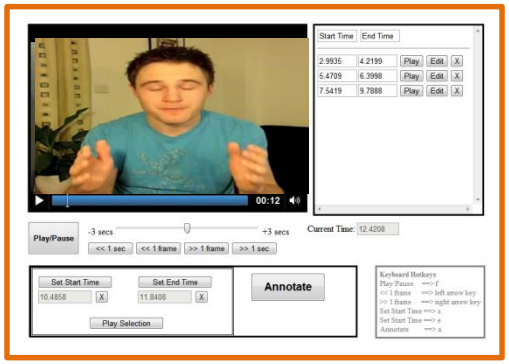
Multimedia

Event / Behavior

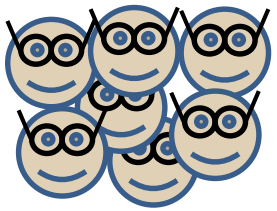
- **Eye gaze**
- **Pause fillers** (hmm, uh)

:

OCTAB



Crowdsourcing Members



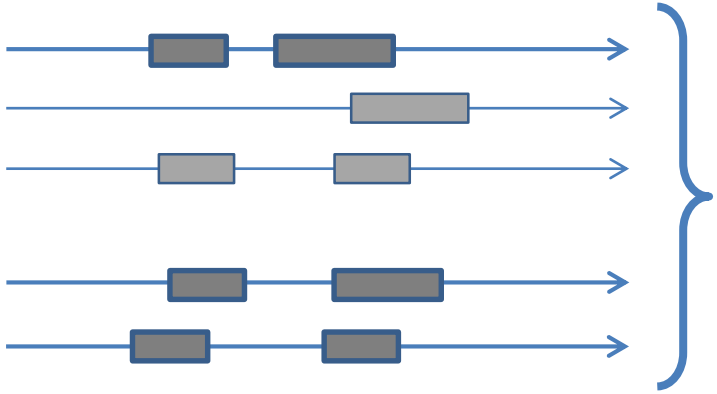
Annotation strategies

- **Unique**
- **Repeated**

Micro-level annotations

Crowd annotations (unique or repeated)

Expert Annotations

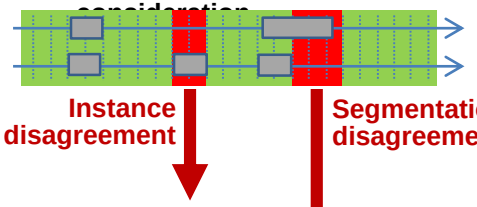


MM-Eval Procedure

(Micro-Level Multimedia Evaluation)

1) Time-Slice Krippendorff's Alpha

- **Evaluates overall agreement, taking chance into**



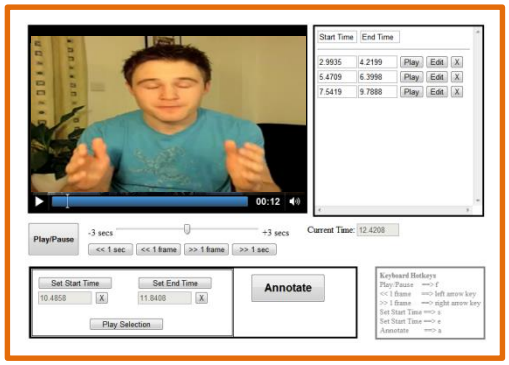
2) Instance agreement metric

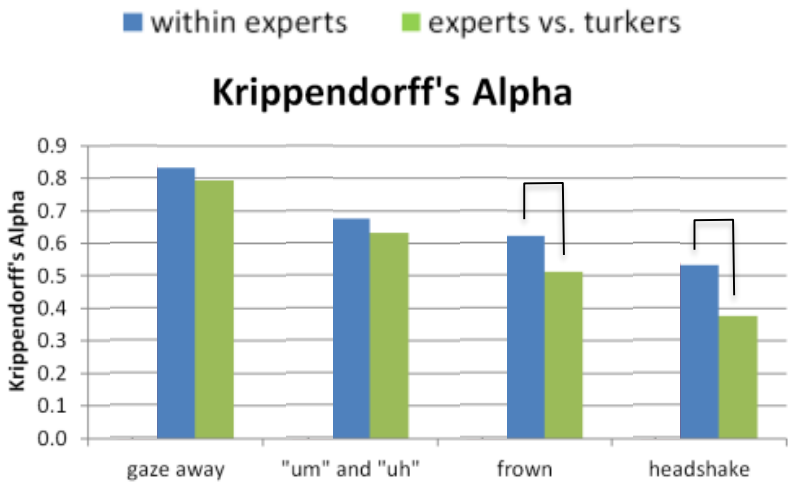
3) Segmentation agreement metric

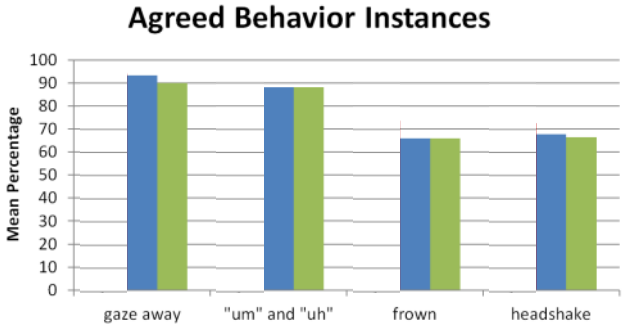
Online Crowdsourcing Tool for Annotations of Behaviors

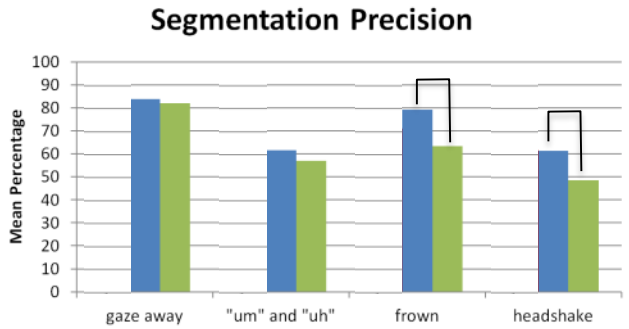
[ACM Multimedia 2012 CrowdMM workshop]

OCTAB



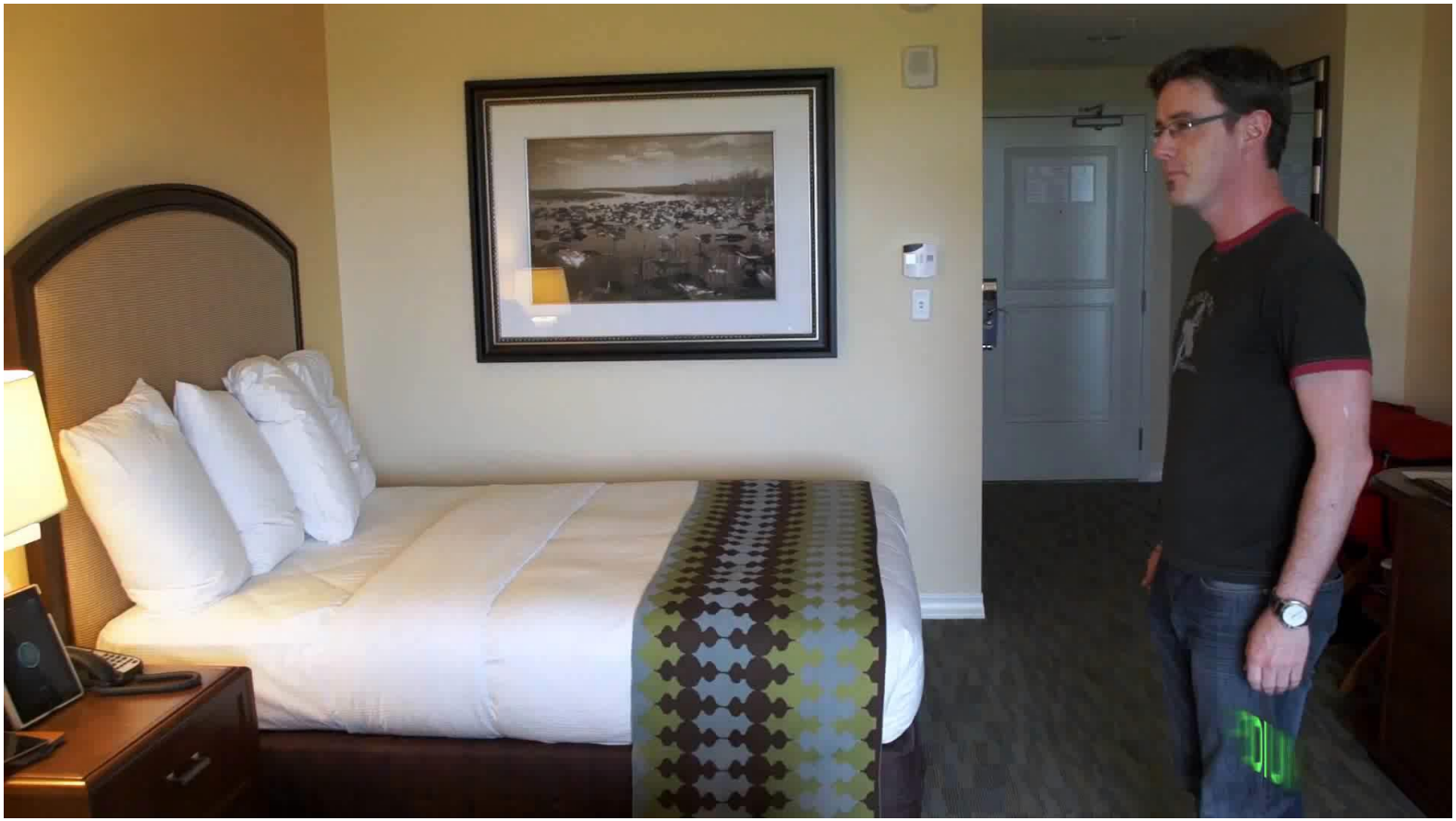






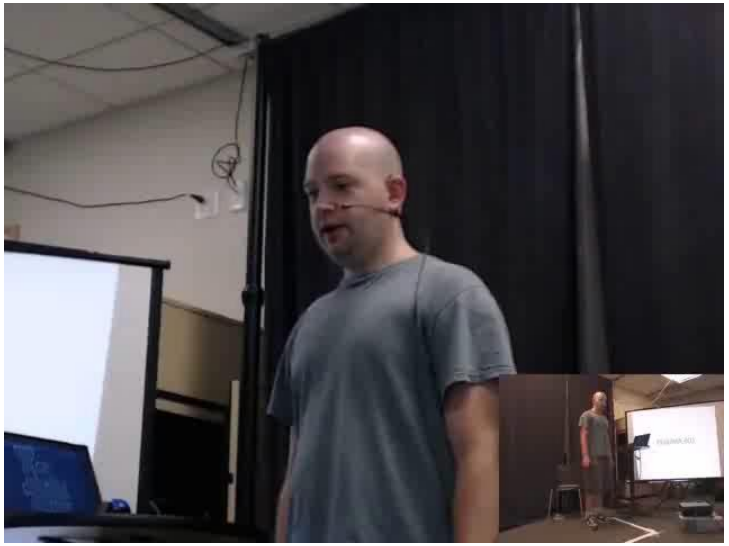
Public Speaking Competition

The Challenges of Public Speaking Training



Cicero: Multimodal Virtual Audience Platform for Public Speaking Training

Live Audience Feedback



Multimodal Sensing

Acoustic and Visual Behaviors

- Visual
  - Gaze and smile behavior (OKAO Vision)
  - Gaze behavior - head orientation (CLNF)
  - Facial expressions (FACET)
  - Orientation towards audience, simple gesture indicator (Kinect)
- Acoustic
  - Pitch, formants, voice quality (COVAREP)
  - speech rate, syllables per breath group, speech intensity (PRAAT)

Cicero: Multimodal Virtual Audience Platform for Public Speaking Training

[IVA 2013]

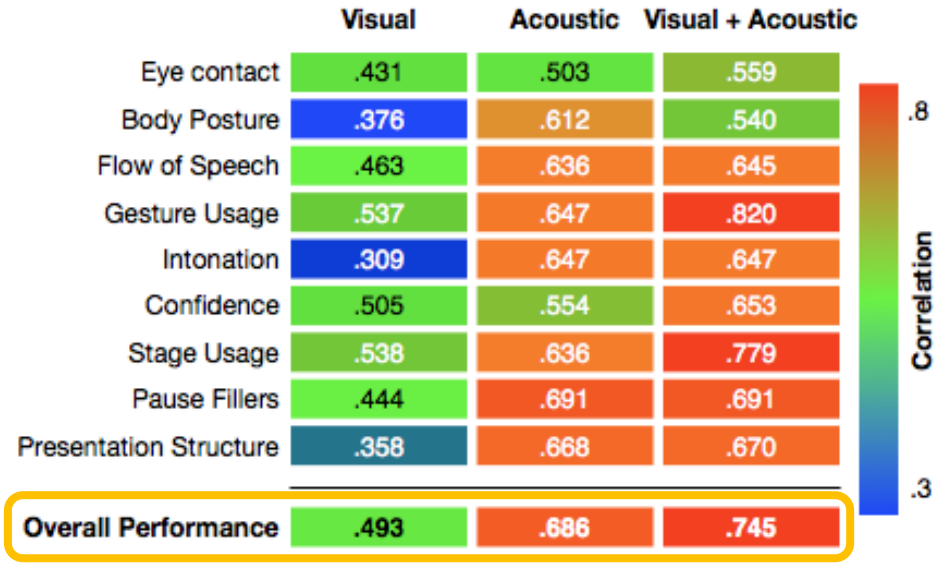
Virtual audience

Correlations between expert assessed behavior and automatically computed behavior descriptors:

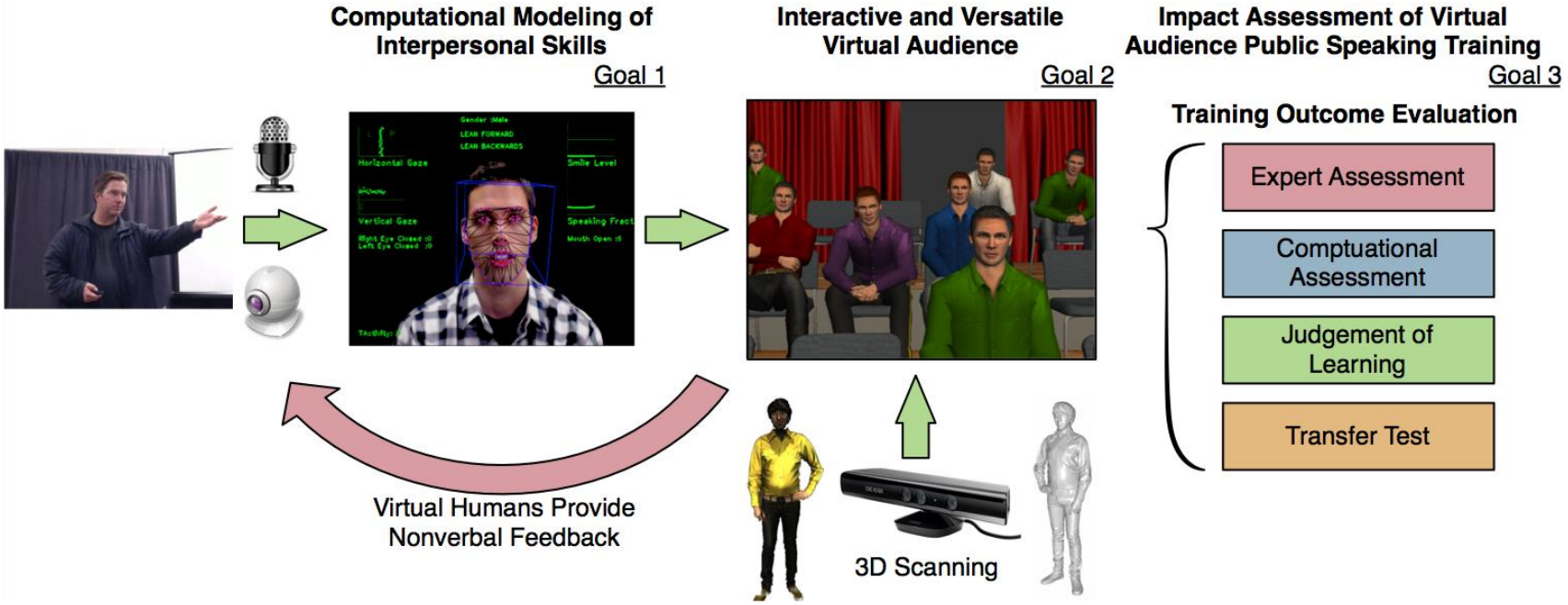
Source	Assessed behavior	Behavior descriptor	Spearman's ρ	p-value
Voice	Flow of speech	Num. pauses	-.469	.09
	Clear intonation	Avg. intensity	**.805**	.002
		Breathiness	**-.615**	.033
	Interrupted speech	Num. pause fillers	**.612**	.034
	Speaks too quietly	Avg. intensity	**-.842**	$< .001$
	Vocal variety	Std. f_0	**.709**	.010
		Spectral Stationarity	**-.586**	.045
Body	Paces too much	Leg movement	**.682**	.021
	Gestures to emphasize	Arm movement	**.710**	.014
	Gestures to much	Arm movement	.437	.179
Gaze	Gazes at audience	Face gaze towards	**.621**	.030
	Avoids audience	Face gaze towards	-.548	.065

Correlation with Expert Assessments

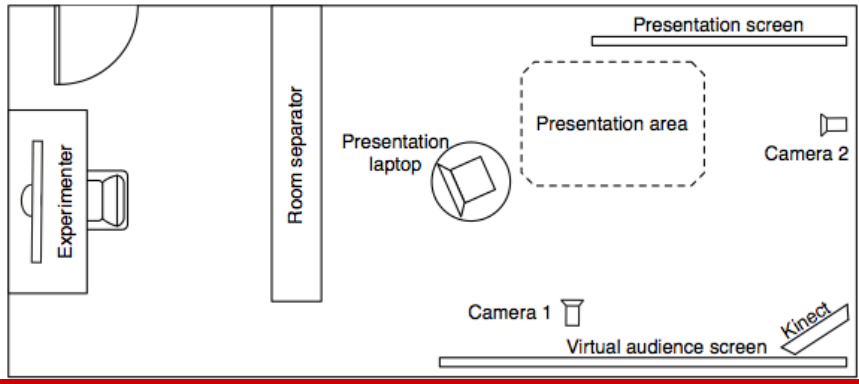
- **Assessment:**
  - Expert Ratings (baseline)
  - Automatic prediction using ensemble trees
- **Results:**
  - Strong correlations with expert assessments (> .7 for overall performance)
  - Multimodal model works best.



Cicero – Public Speaking Training

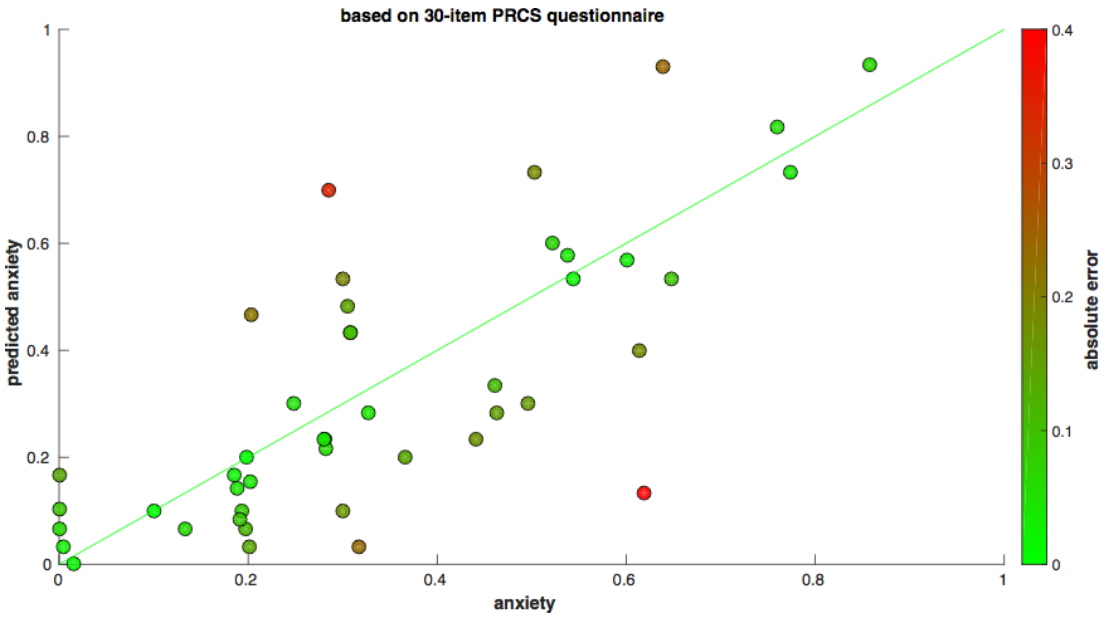


Cicero – Public Speaking Training



Automatic Anxiety Assessment

- **Assessment:**
  - PRCS rating scale
  - Automatic prediction using ensemble trees
- **Results:**
  - Strong correlations with PRCS scale
  - Multimodal model works best with 0.78 correlation and 0.13 mean absolute error



Canal 9 Dataset

Bousmalis et al., ACII 2009

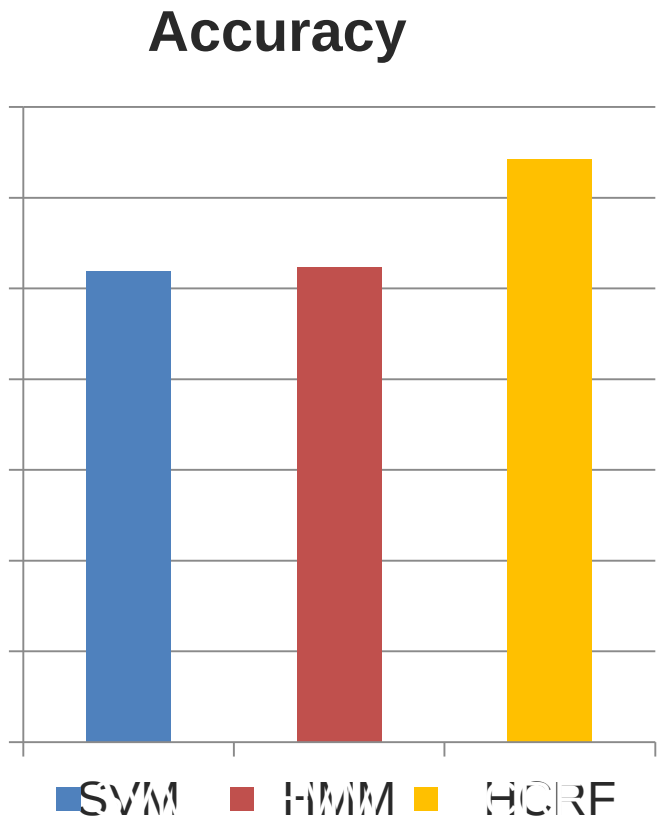
We will use this dataset to explain the models!

- Ground truth based ONLY on verbal content
- 11 debates - 28 distinct individuals
- 53 episodes of agreement
- 94 episodes of disagreement

Binary Visual Features: Presence per frame of 8 gestures

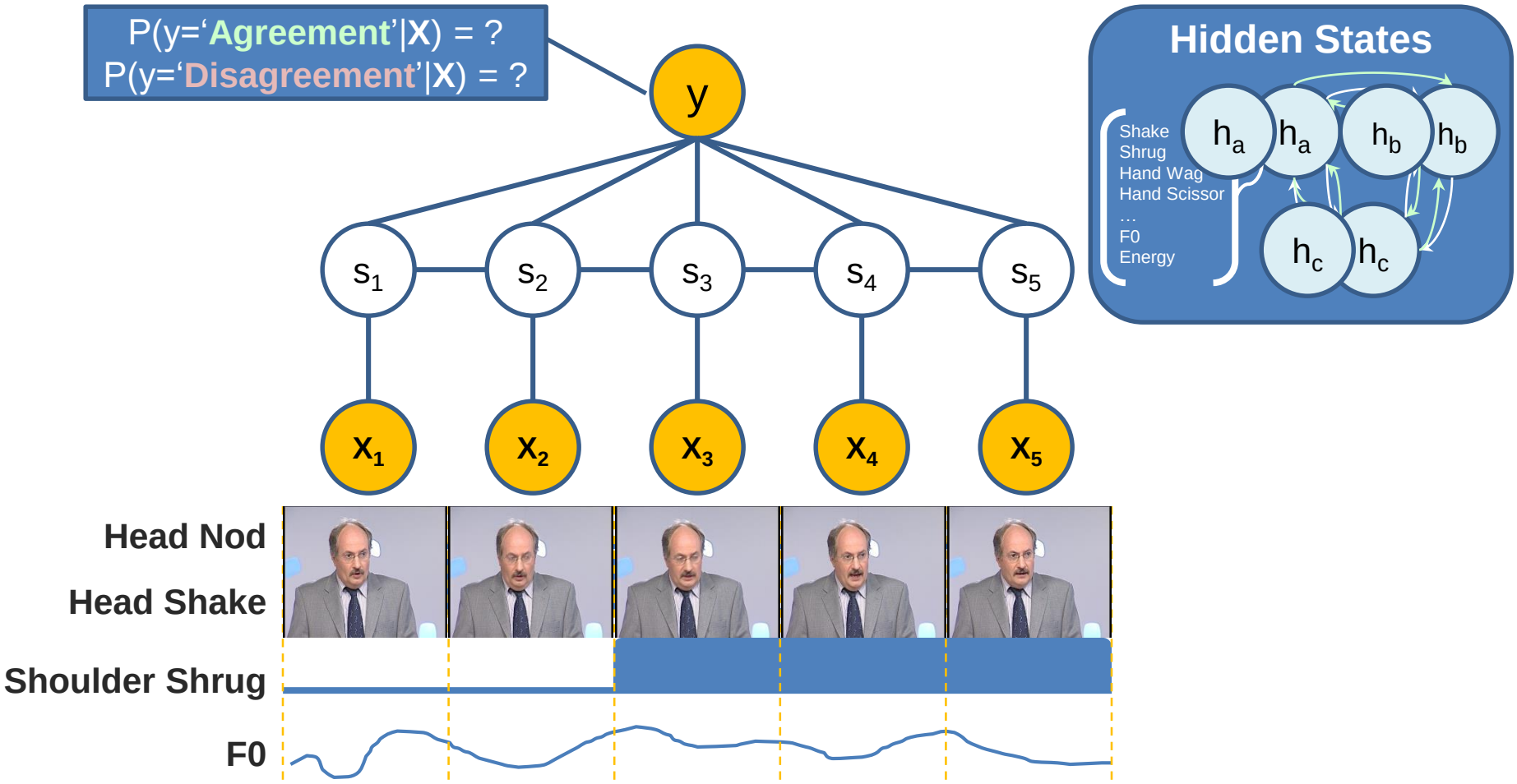
Continuous Auditory Feature: F0, Energy

Classification of Agreement/Disagreement



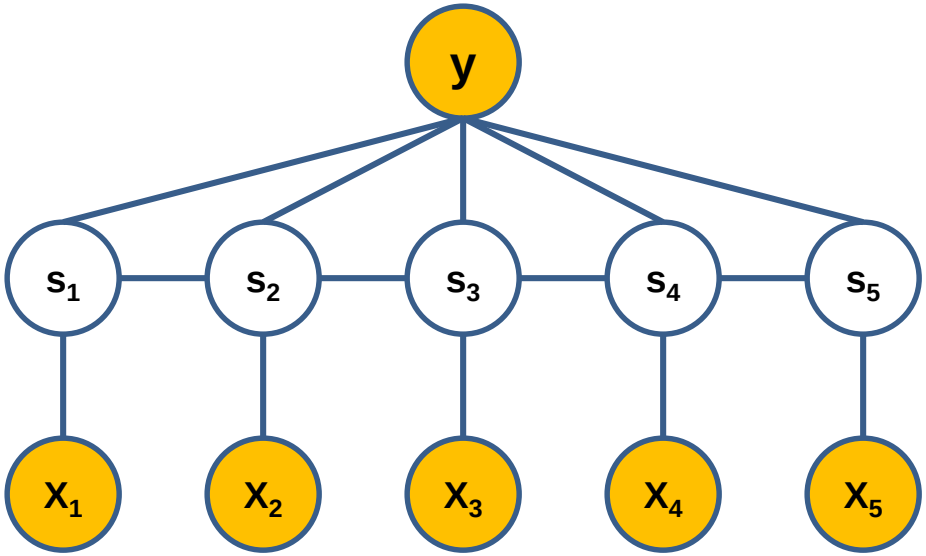
- Experiments with 2 Labels
- Support Vector Machines (SVMs)
- Hidden Markov Models (HMMs)
- Hidden Conditional Random Fields (HCRFs)

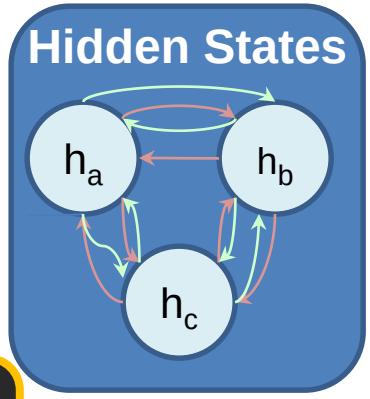
Hidden Conditional Random Field



Learned HCRF Model

- Weights and equivalent potentials for each relationship:
  - features and hidden states $\theta_x(s_t)x_t$
  - hidden states and labels $\theta_e(s_t, s_{t-1}, y)$
  - transitions among hidden states and labels $\theta_y(s_t, y)$





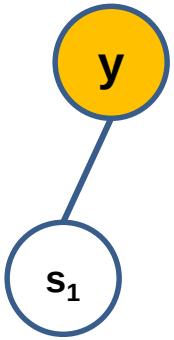
$$F(y, \mathbf{s}, \mathbf{x}; \theta) = \sum_t \theta_x(s_t)x_t + \theta_y(s_t, y) + \theta_e(s_t, s_{t-1}, y)$$

HCRF Hidden State Analysis

Compatibility between features and hidden

	h_a	h_b	h_c
Head Nod	0.01	**1.6**	**0.95**
Head Shake	**1.32**	0.21	0.34
Forefinger Raise	**2.55**	**0.53**	**0.56**
Hand Wag	0.42	0.32	**0.47**
…	…	…	…

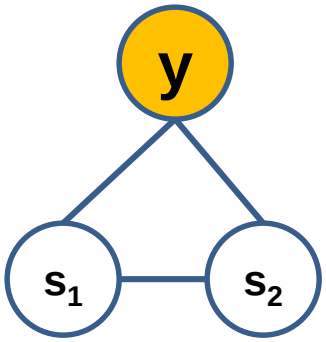
Association of hidden states with labels θ_y



	h_a	h_b	h_c
Agreement	0.33	**1.39**	**0.81**
Disagreement	**1.40**	0.32	**0.79**

HCRF Hidden State Analysis

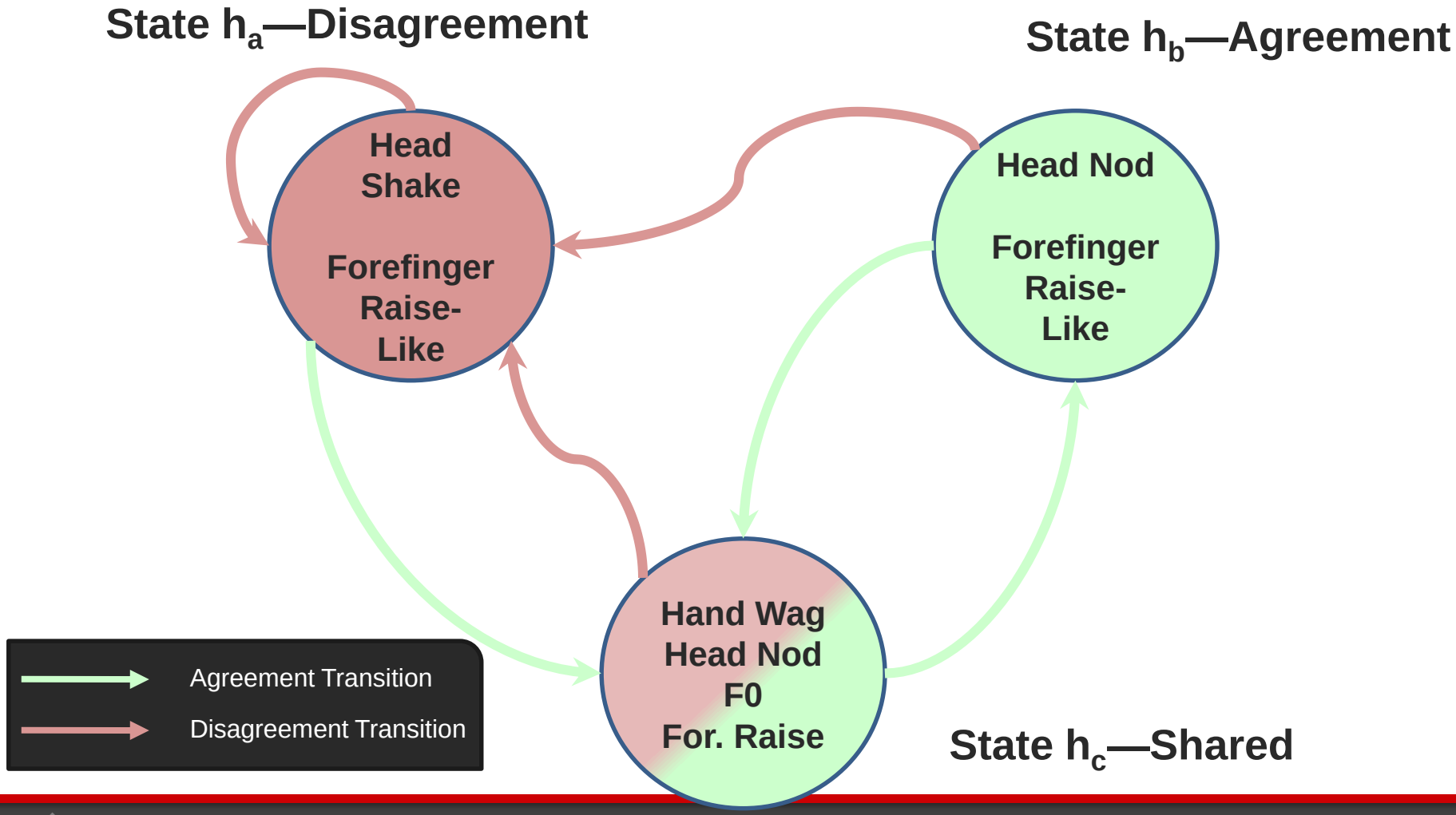
Compatibility between transitions and labels θ_e



	Agreement, h_a	Agreement, h_b	Agreement, h_c
h_a	0.08	0.04	**0.19**
h_b	0.02	0.09	**0.23**
h_c	0.11	**0.22**	**0.20**

	Disagreement, h_a	Disagreement, h_b	Disagreement, h_c
h_a	**0.21**	0.03	0.09
h_b	**0.12**	0.02	0.08
h_c	**0.14**	0.03	0.08

HCRF Hidden State Analysis

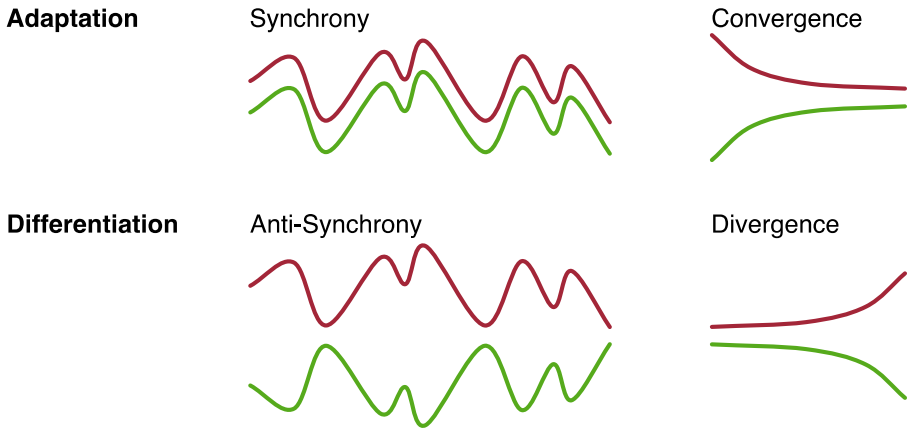


Understanding Interpersonal Dynamics

Interpersonal

I. **Predictive models**
II. **Prototypical patterns**
III. **Mutual influence**
IV. **Idiosyncrasy**

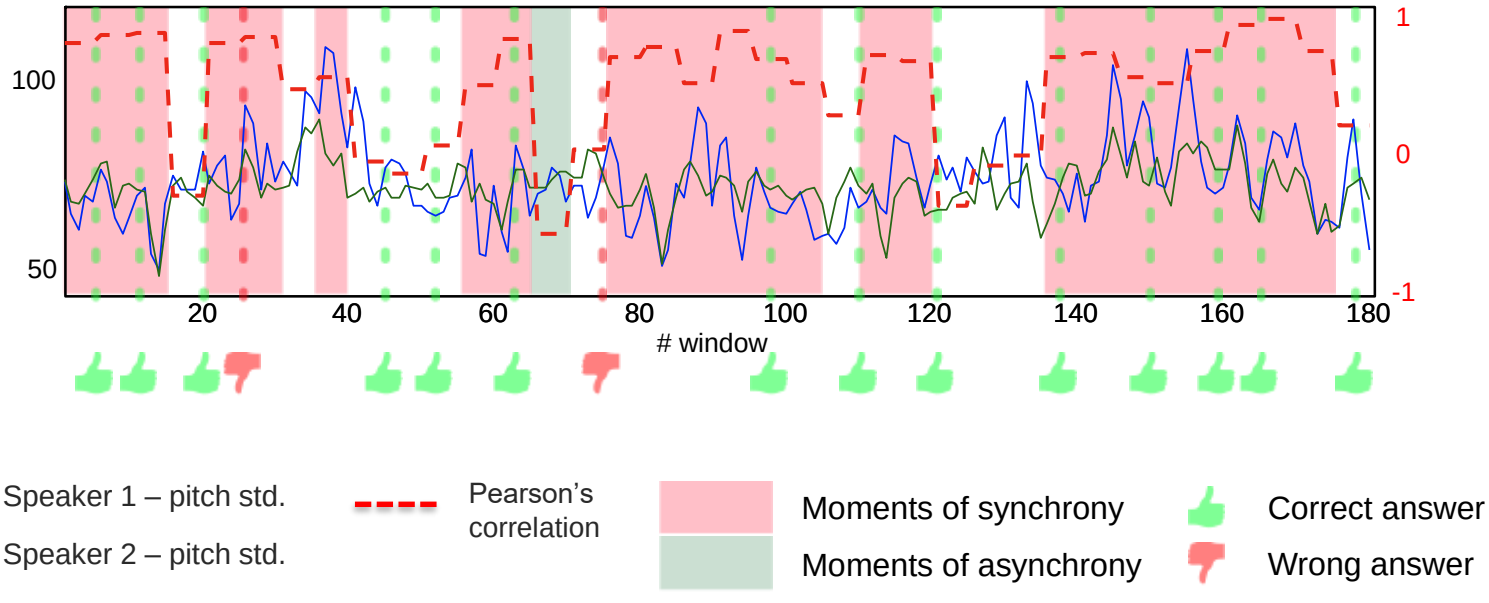
- Interlocutors adapt:
  - Lexicon (gestural and verbal)
  - Nonverbal Behavior (facial expressions, posture)
  - Prosody and speech
- High entrainment signifies:
  - Understanding
  - Flow of the conversation
  - Cooperation



Example: Audio-based Entrainment Analys

[ICMI workshop 2013]

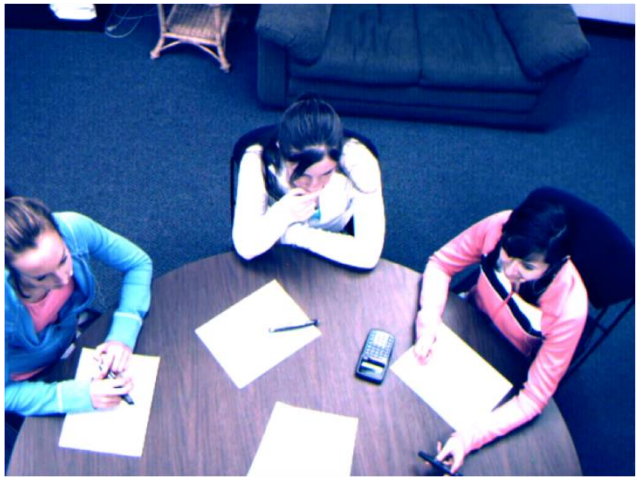
- Correlate prosodic parameters to measure moments of entrainment
- Identify performance; social dominance/expertise; teamwork

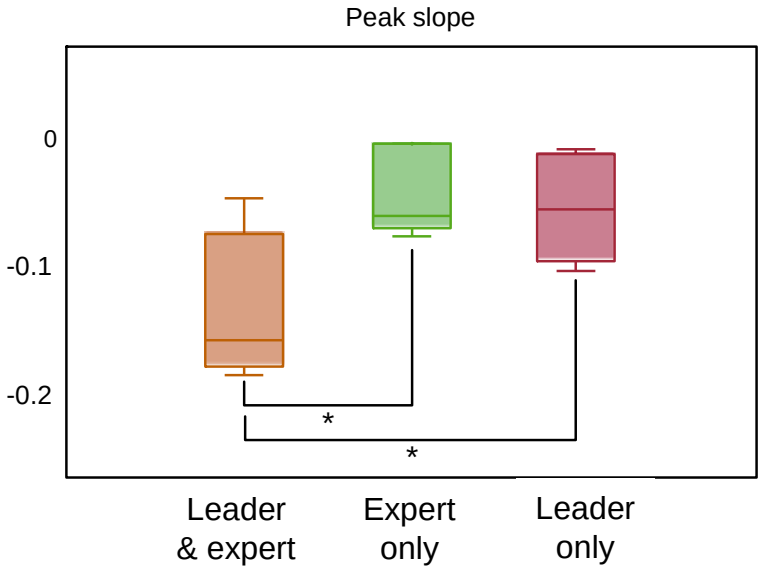


Multimodal Learning Analytics: Expert vs Leader

[ICMI workshop 2013]

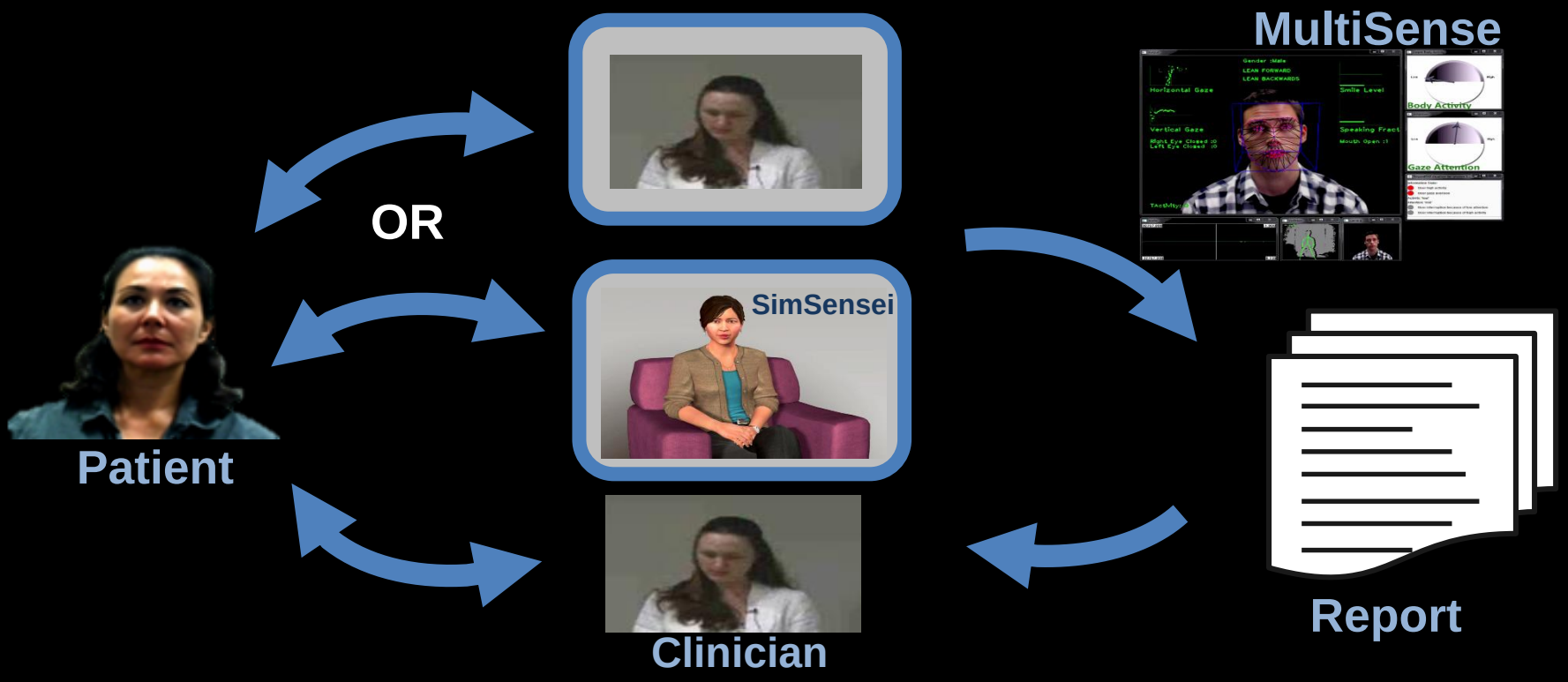
- Identify expertise and leadership in students based on observable nonverbal behavior
- Several indicators allow identification of leaders in groups (e.g. turn-taking behavior, speech intensity)





- **Voice quality (here tenser voice)** allows identification of leading expert students

Technologies for Health Behavior Informatics



Behavioral Indicators of Psychological Distress

[AAMAS 2014]

Distress Assessment Interview Corpus

DAIC

Study protocol

- **Self-Assessment of Psychological Distress**
  - Depression (PHQ-9)
  - Post-traumatic stress disorder (PCL-C)
  - Highly correlated with accepted clinical diagnosis (sensitivity & specificity > .80)
- **Interview Phases**
  - Rapport building
  - Intimate/clinical questions
  - Cool-down phase

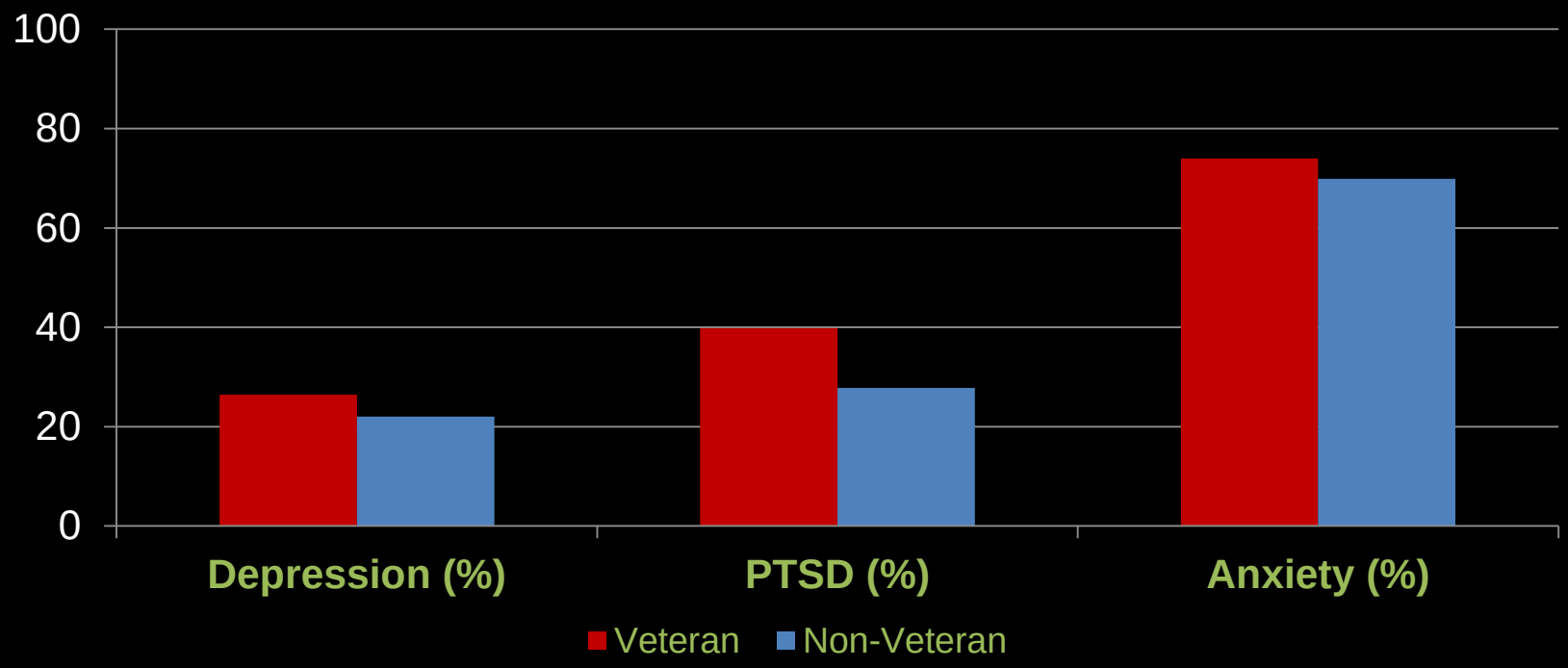
Kroenke K, Spitzer RL, Williams JB, The PHQ-9: *Validity of a brief depression severity measure*, Journal of General Internal Medicine. 2001, 16(9): pp.606-13.

Weathers, F. W., Litz, B. T., Herman, D. S., Huska, J. A. & Keane, T. M. (1993). *The PTSD Checklist (PCL): Reliablity, validity, and diagnostic utility*, Paper presented at the 9th Annual Conference of the ISTSS, San Antonio.

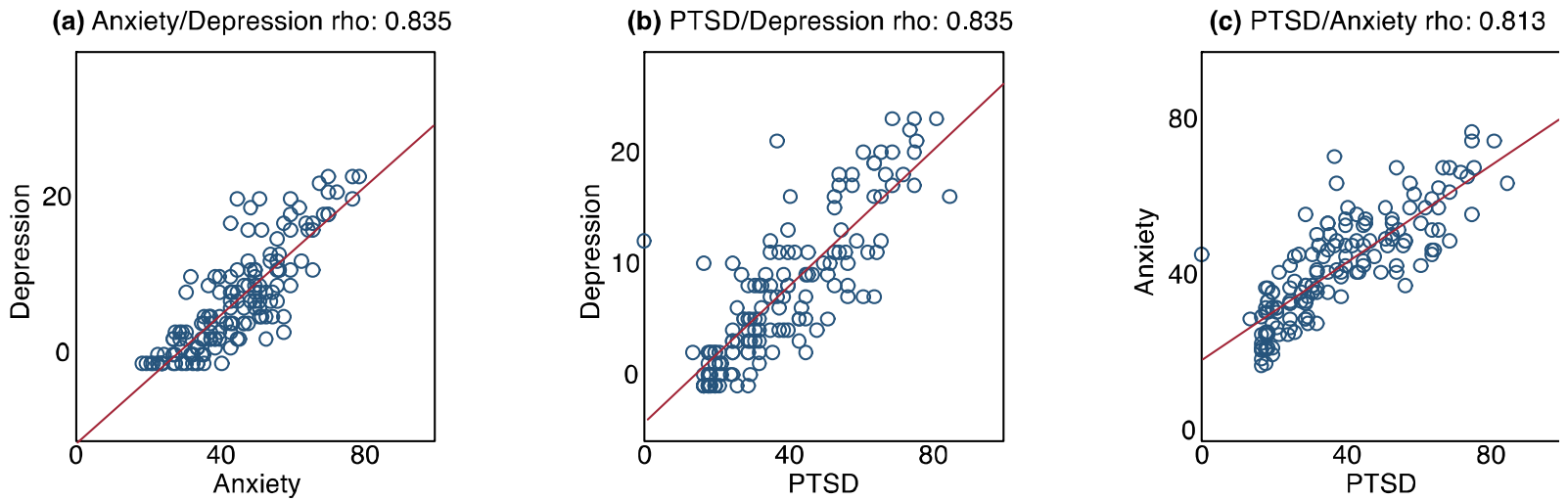
Demographics

Age	Mean	Median			
	47.93	51			
Gender	Male	Female			
	65.64%	34.36%			
Education	HS/GED	Some College	2 yr college	4 yr college	Post Graduate Degree
	13.54%	37.91%	13.54%	26.71%	8.30%
Race	African American	Asian	White/Caucasian	Hispanic	Native American
	36.38%	4.47%	43.77%	13.62%	1.75%
Military	44.32%				
Branch	Army	Navy	Marine Corps	Air Force	Coast Guard
(those who are military)	44.31%	23.57%	19.11%	11.79%	1.22%

Demographics



Co-morbidity



Especially high correlations between clinical severity (0.80)

Prevalence of Depression–PTSD Comorbidity: Implications for Clinical Practice Guidelines and Primary Care-based Interventions, D. G. Campbell, et al., J Gen Intern Med., 22(6): 711–718, 2007.

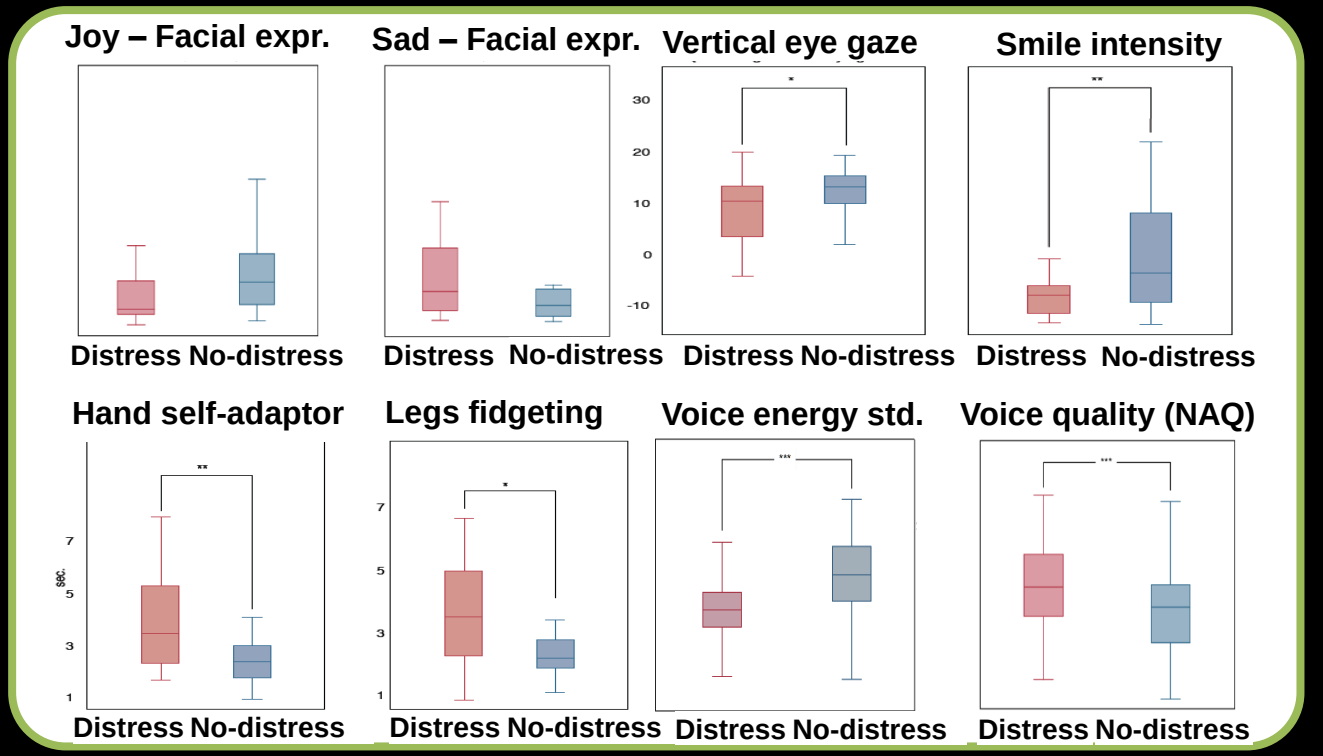
Expert Opinion Assessment and Related Work

- *Clinical experts analyzing selected samples from DCAPS interview corpuses*

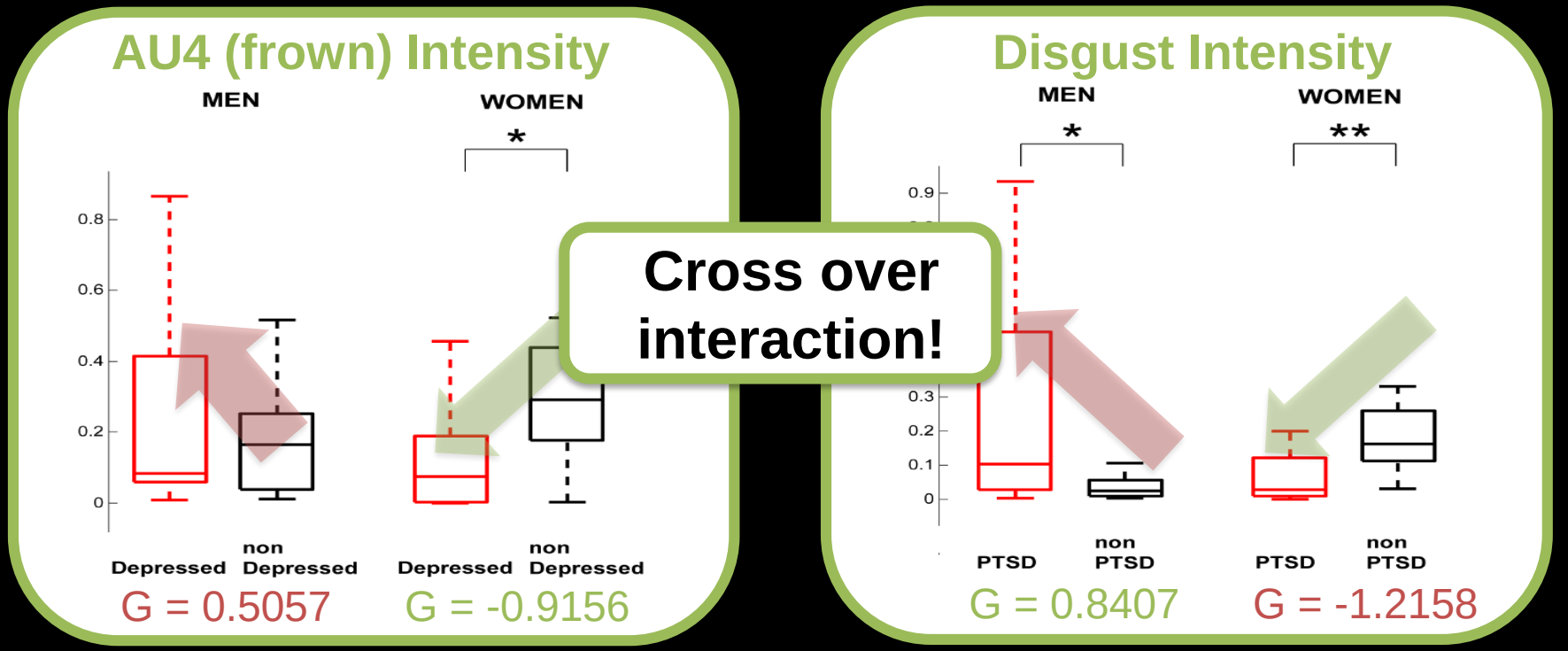
  Behaviors associated with psychological conditions

  - **Affect:** positive vs. negative affect during interactions [Perez and Riggio, 2003; Kirsch and Brunnhuber, 2007]
  - **Engagement:** lack mutual gaze, provided feedback, and slumped posture [Perez and Riggio, 2003; Schelde, 1998]
  - **Variability:** lack of gestures and overall movement, monotonous speech [Darby et al., 1984; Hall et al., 1995]
  - **Agitation:** changes in voice quality and fidgeting [Fairbanks, 1982; Flint et al., 1993]
  - **Latency:** delay of responses, reduced speech rate and nonverbal responses [Hall et al., 1995; Waxer, 1974]

Psychological Distress Indicators [IEEE FG 2013]

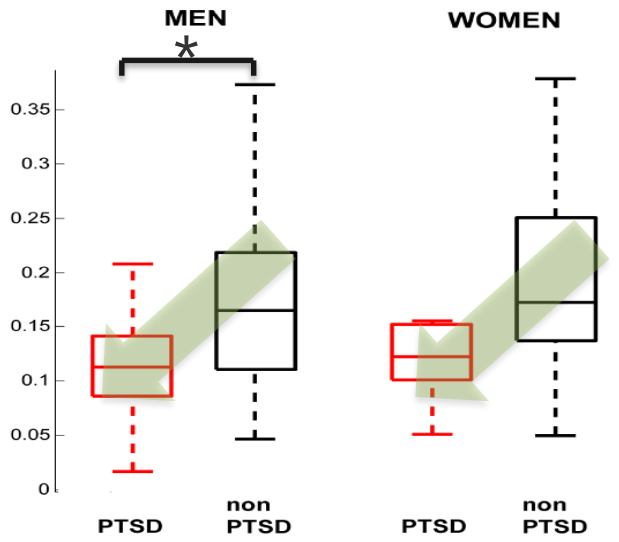


Indicators with different trends on both genders [ACII 2013]



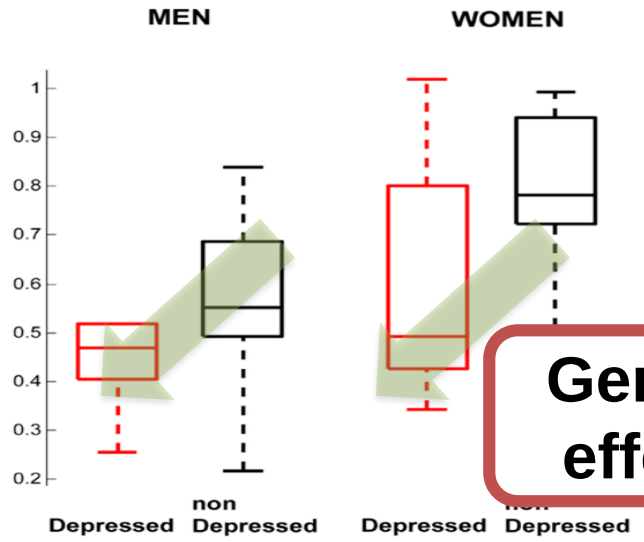
Indicators with similar trend on both genders [ACII **2013**]

Head Rotation Variation



G = -0.7426 G = -0.586

Emotional Variation



G = -0.5664 G = -0.6852

Gender effect!

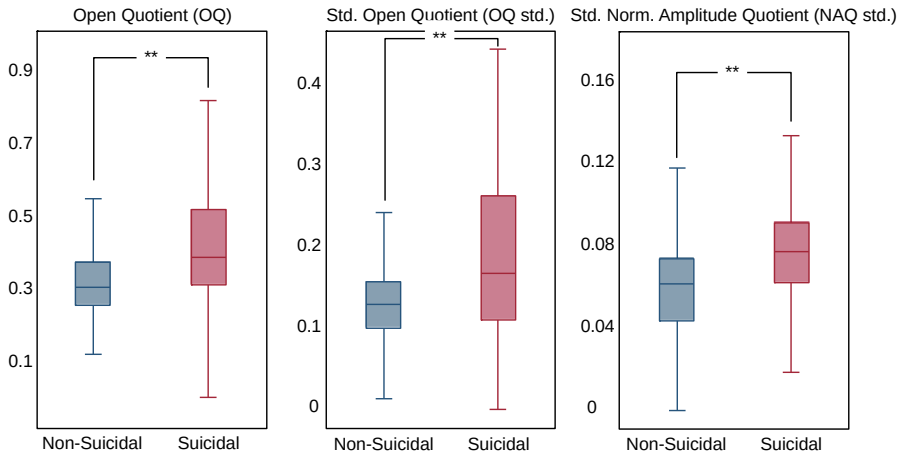
Suicide Prevention [ICASSP 2013]

	Age Groups
Rank	11-17
1	Unintentional Injury 3,924
2	Homicide 1,101
3	Suicide 823
4	Malignant Neoplasms 745
5	Heart Disease 278

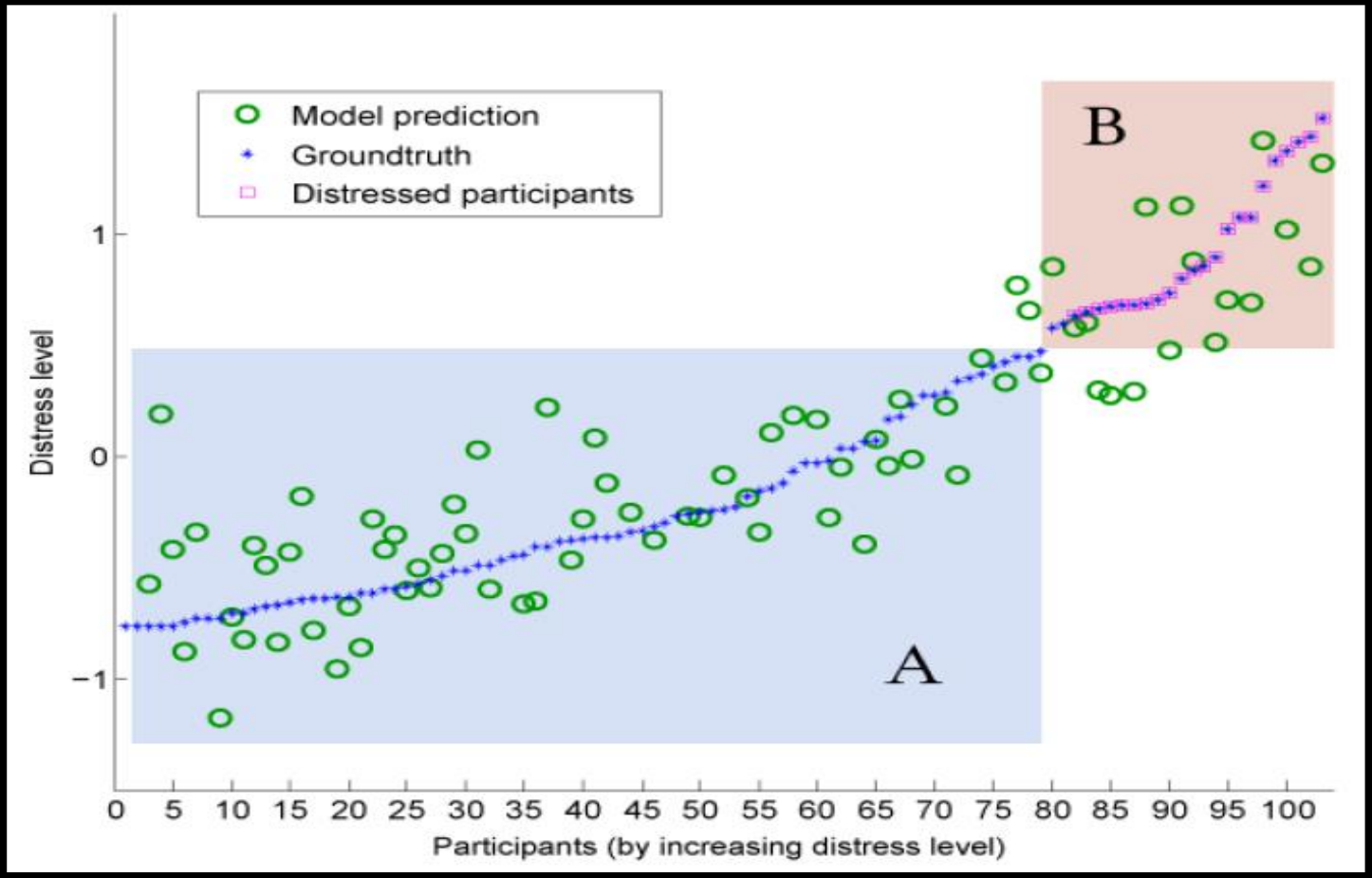
Source: CDC

Experiment

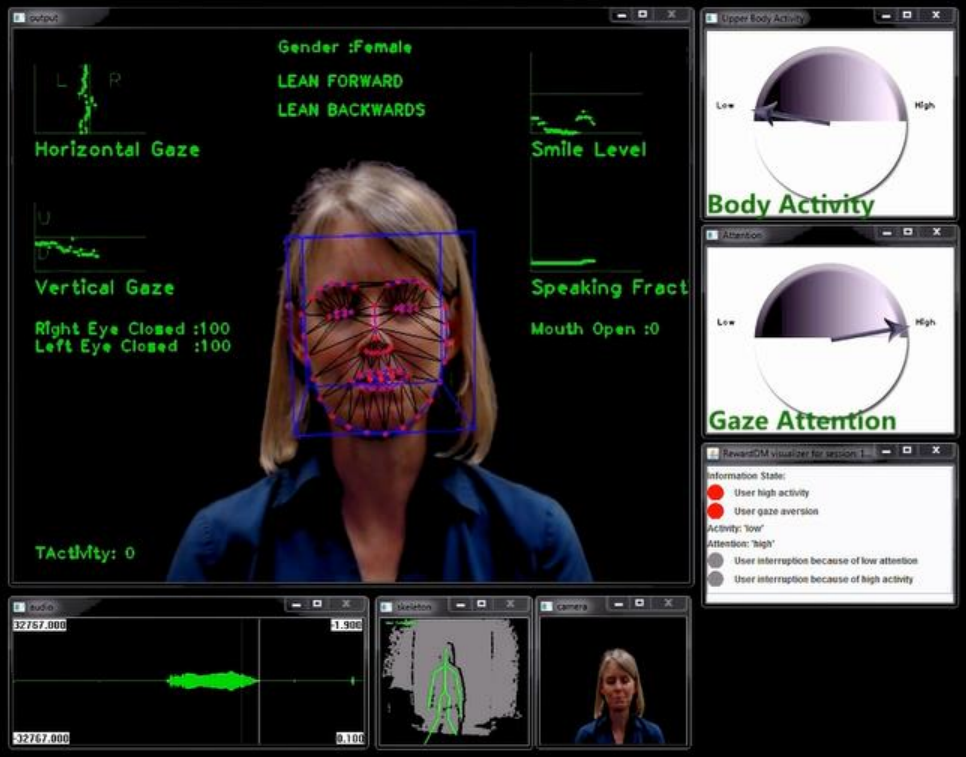
- Nonverbal indicators of suicidal ideations
- Dataset: 30 suicidal adolescents/30 non-suicidal adolescents
- Suicidal teenagers use **more breathy tones**



Automatic Distress Level Prediction



MultiSense



SimSensei

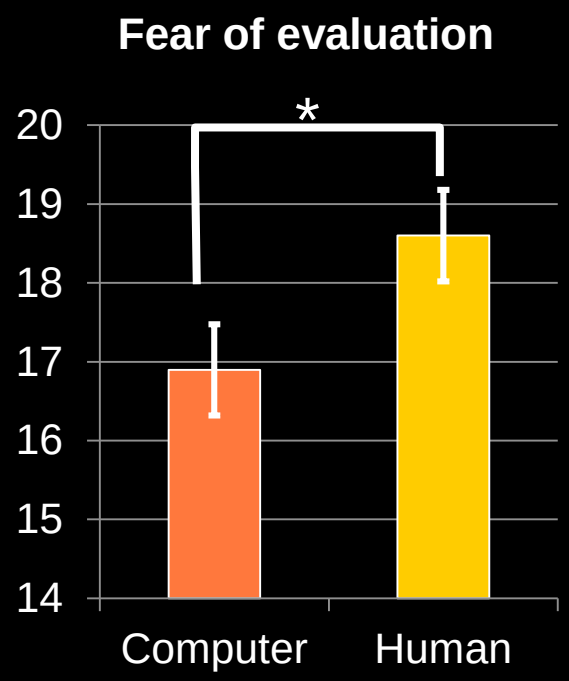
Why are we creating an AI agent (SimSensei)?

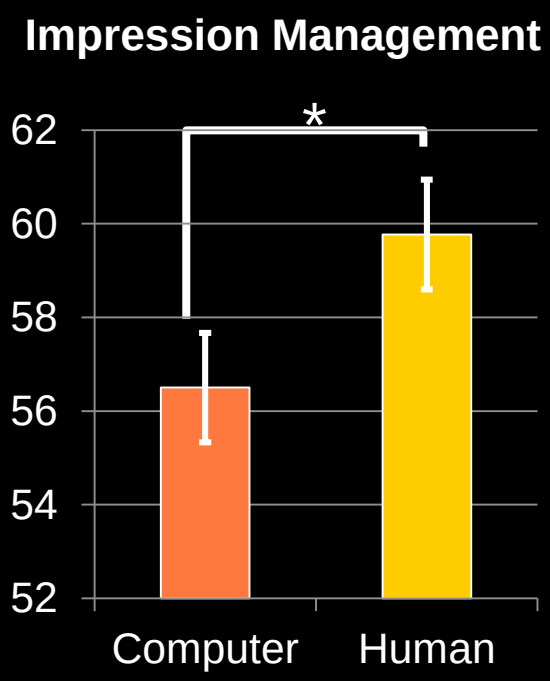
[AAMAS 2014]

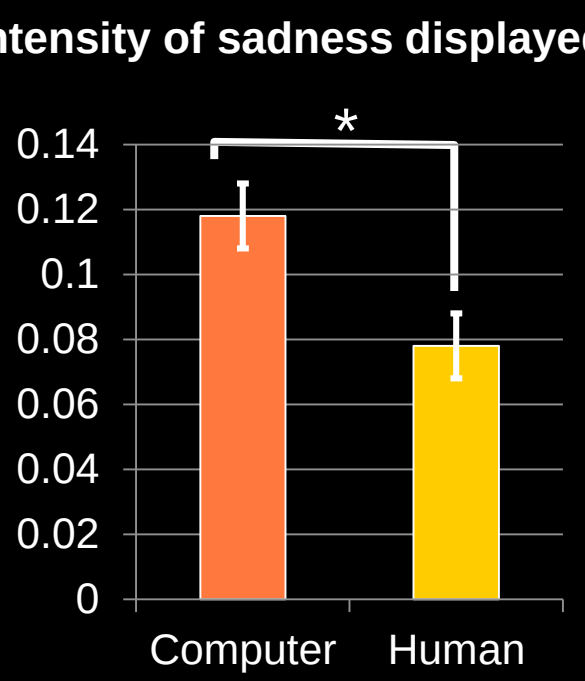
- Compare responses when participants believe the avatar is controlled by a human or by an AI
  - Computer-framed (N=77) vs.
  - Human-framed (N=77) interactions

Why are we creating an AI agent (SimSensei)?

[AAMAS 2014]



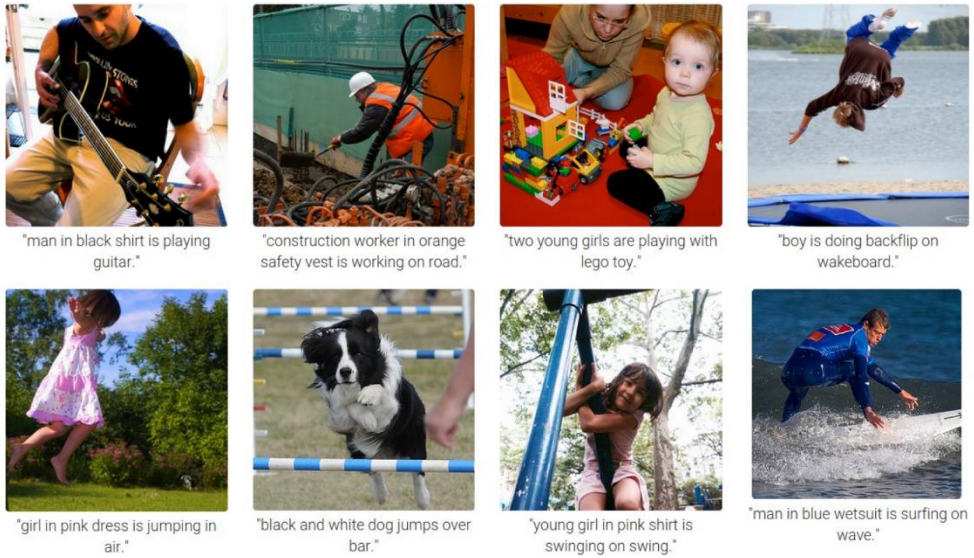




Other Multimodal Applications

Media description

- Given a piece of media (image, video, audio-visual clips) provide a free form text description
- Earlier work looked at classes/tags/etc.



Media description – MS COCO Dataset

- Microsoft Common Objects in COntext (MS COCO)
- 120000 images
- Each image is accompanied with five free form sentences describing it (at least 8 words)
- Sentences collected using crowdsourcing (Amazon Mechanical Turk)

The man at bat readies to swing at the pitch while the umpire looks on.

A large bus sitting next to a very tall building.

State-of-the-art on MS COCO

- A challenge was done with actual human evaluations of the captions (CVPR 2015)

	M1	M2	M3	M4	M5
Human[5]	0.638	0.675	4.836	3.428	0.352
Google[4]	0.273	0.317	4.107	2.742	0.233
MSR[8]	0.268	0.322	4.137	2.662	0.234
Montreal/Toronto[10]	0.262	0.272	3.932	2.832	0.197
MSR Captivator[9]	0.250	0.301	4.149	2.565	0.233
Berkeley LRCN[2]	0.246	0.268	3.924	2.786	0.204
m-RNN[15]	0.223	0.252	3.897	2.595	0.202
Nearest Neighbor[11]	0.216	0.255	3.801	2.716	0.196

State-of-the-art on MS COCO

- What about automatic evaluation?
  - Human labels are expensive…

	CIDEr-D	Meteor	ROUGE-L	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Google[4]	0.943	0.254	0.53	0.713	0.542	0.407	0.309
MSR Captivator[9]	0.931	0.248	0.526	0.715	0.543	0.407	0.308
m-RNN[15]	0.917	0.242	0.521	0.716	0.545	0.404	0.299
MSR[8]	0.912	0.247	0.519	0.695	0.526	0.391	0.291
Nearest Neighbor[11]	0.886	0.237	0.507	0.697	0.521	0.382	0.28
m-RNN (Baidu/ UCLA)[16]	0.886	0.238	0.524	0.72	0.553	0.41	0.302
Berkeley LRCN[2]	0.869	0.242	0.517	0.702	0.528	0.384	0.277
Human[5]	0.854	0.252	0.484	0.663	0.469	0.321	0.217

Video captioning

- MPII Movie Description dataset
  - A Dataset for Movie Description
- Montréal Video Annotation dataset
  - Using Descriptive Video Services to Create a Large Data Source for Video Annotation Research

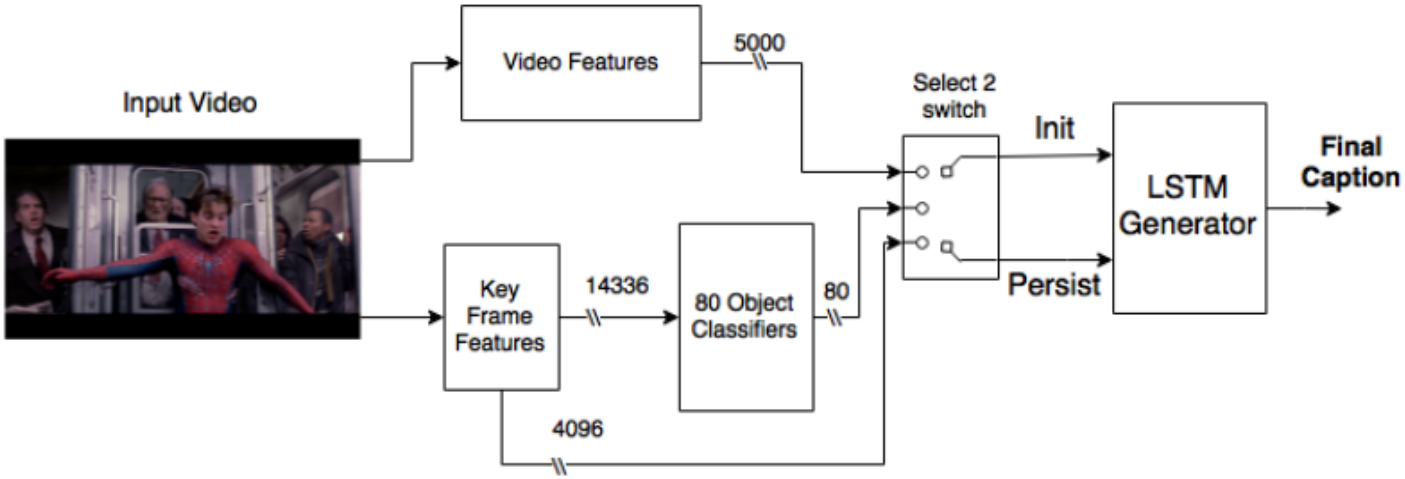
AD: Abby gets in the basket.

Mike leans over and sees how high they are.

Abby clasps her hands around his face and kisses him passionately.

Video description state-of-the-art

- Describing and Understanding Video & The Large Scale Movie Description Challenge (LSMDC), hosted at ICCV 2015
  - **Video Captioning with Recurrent Networks Based on Frame- and Video-Level Features and Visual Content Classification**
- Compared to human performance for deciding winners



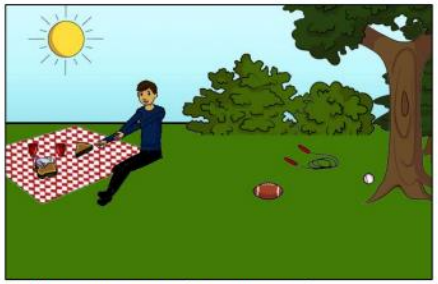
model	init	persist	perplex	beam size	avg.len	Bleu_4	CIDEr	METEOR	ROUGE_L	
9 cls+traj	cls	traj	55.14	1	5.33	**0.006**	**0.092**	**0.058**	**0.146**	(b)
				5	3.80	0.004	0.082	0.049	0.128	

Visual Question Answering

Task - Given an image and a question answer the question (http://www.visualqa.org/)

What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?



Is this person expecting company?
What is just under the tree?

Does it appear to be rainy?
Does this person have 20/20 vision?

VQA state-of-the-art

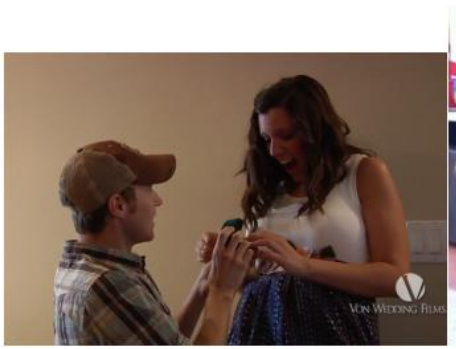
- LSTM + CNN
  - Currently held by challenge organizers
- winner is a representation/deep learning based model
- Currently good at yes/no question, not so much free form and counting

Results

	User	Team Name	By Answer Type			Overall
			Yes/No	Other	Number	
1	vqateam-deeperLSTM_NormlizeCNN		80.56 (1)	43.73 (2)	36.53 (2)	58.16 (2)
2	cxiong		80.43 (2)	48.33 (1)	36.82 (1)	60.36 (1)
3	Q.Wu	ACVT_Adelaide	79.05 (3)	40.61 (3)	36.10 (3)	55.98 (3)
4	vqateam-lstm_cnn		79.01 (4)	36.80 (4)	35.55 (5)	54.06 (4)
5	vqateam-q_lstm_alone		78.12 (5)	26.99 (5)	34.94 (6)	48.89 (5)
6	vqateam-prior_per_qtype		71.17 (6)	9.32 (6)	35.63 (4)	37.55 (6)
7	vqateam-all_yes		70.53 (7)	1.26 (7)	0.43 (7)	29.72 (7)

Multimedia event detection

- Given video/audio/ text detect predefined events or scenes
- Segment events in a stream
- Summarize videos



Multimedia event detection

- TrecVid Multimedia Event Detection (MED) 2010-2016
- One of the six TrecVid tasks
- Audio-visual data
- Event detection

Cross-media retrieval project

f media,
elevant