Carnegie Mellon University
Language
Technologies
Institute

Toward Social AI Systems that Recognize and Synthesize Human Behaviors

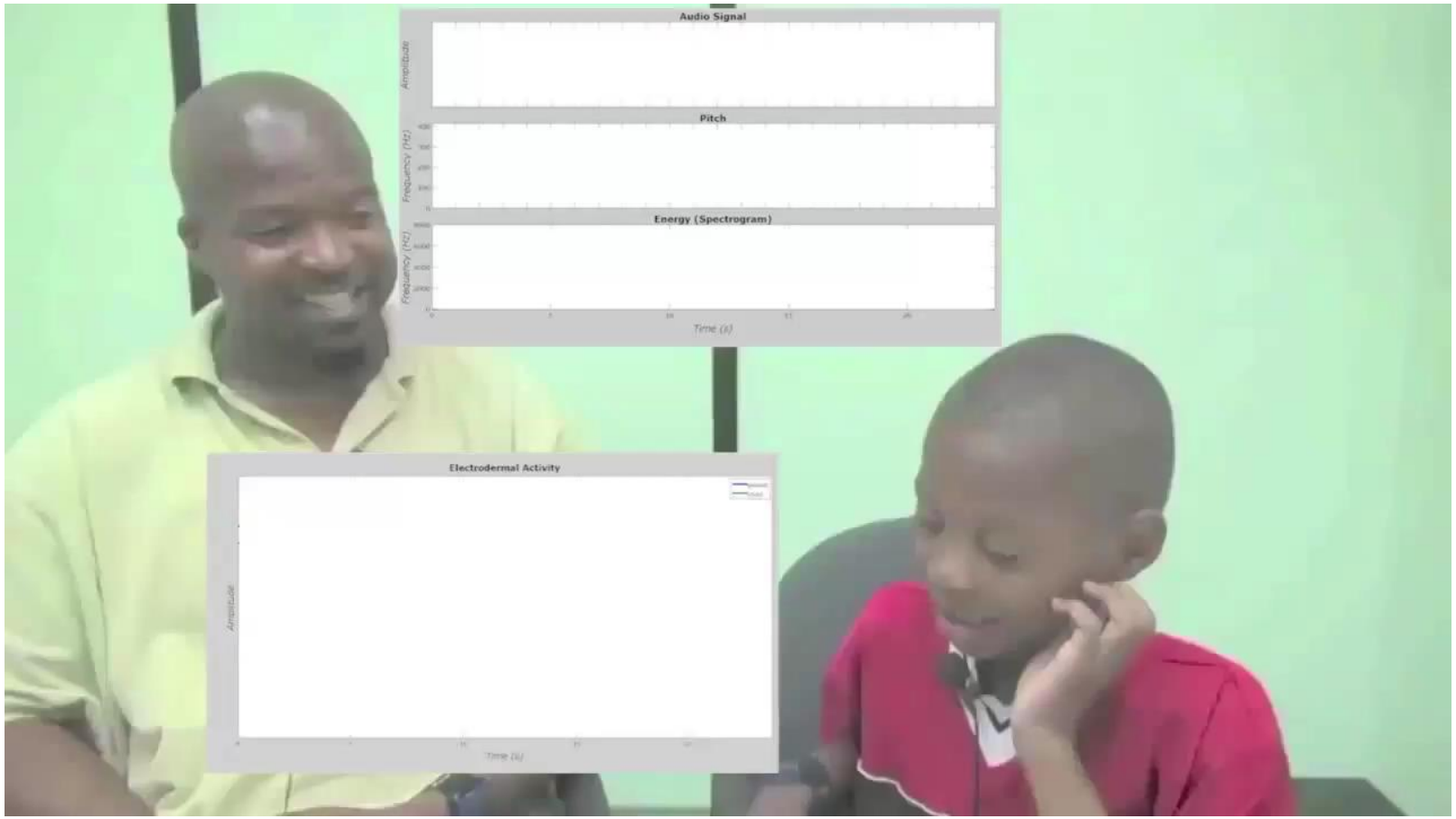
Carlos Busso

Professor at UT Dallas

Incoming Professor at CMU (Spring 2025)

THE UNIVERSITY OF TEXAS AT DALLAS

Motivation - Nonverbal Communication



Credits to SAIL @ USC
Prof. Narayanan

Human Communication – a Multimodal Process

- **Human-human interaction is a sophisticated problem**

Personality

Emotions

Intentions

Multimodal communication

Entrainment

Context

Backchannel

Turn-taking patterns

Nonverbal behaviors

Need for Multimodal Social AI Systems

- **Human Communication**
  - Spoken language carries crucial information
    - Intent, desires, emotions
  - Information resides at multiple time scales, through multiple cues
    - Inherently multimodal: natural communication involves speech, facial/hand gestures, head movement, postures, …
  - Information can be complementary (e.g., smiles while listening), or redundant (e.g., frowning while speaking in an angry tone)

Designing **Social AI** systems require careful consideration on recognizing and synthesizing multimodal human behaviors

Need for Multimodal Social AI Systems

- **Multimodal processing is an important for social AI systems**
  - Natural way for humans to interact with the environment
  - It can increase robustness and performance

- **Autonomous systems**

- **Entertainment**

- **Healthcare**

Need for Multimodal Social AI Systems

- **Recognizing social signals is not enough**
- **We should be able to synthesize social signals back to the user**
  - Gestures that are tightly synchronized
  - Gestures that mirror human behaviors
  - Increase engagement by replicating entrainment effects
- **Vision: Closing the loop by sensing and responding to users with human-like behaviors**

Recognition of Social Signals:

Human Communication – a Multimodal Process

- **Human-human interaction is a sophisticated problem**

Personality

Emotions

Intentions

Multimodal communication

Entrainment

Context

Backchannel

Turn-taking patterns

Nonverbal behaviors

Why Study Emotion or Attitude?

- **Emotions play a crucial role in human interaction**
  - Emotional (vs. cognitive) reasoning
  - Emotion is reflected in our body
  - Our emotions change the minds of others
  - People rely on emotion for making decisions

- **Knowing the user's emotional state should help to adjust system performance**
- **User can be more engaged and have a more effective interaction with the system**

Role of Emotion Recognition in Society

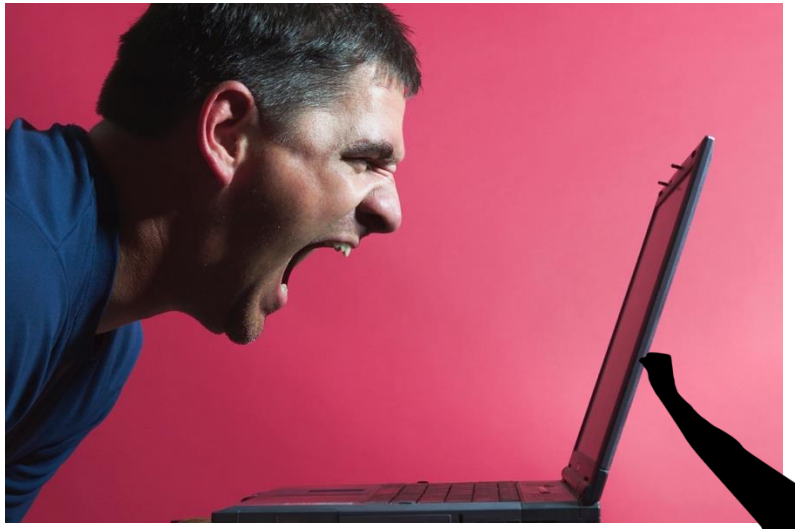
- **Call centers**
  - Quality of service
  - Coping with frustrated users
- **Education**
  - Detect frustration
- **Security and Defense**
  - Triage content
- **Healthcare**
  - Detecting emotional changes as tool to assess mood disorders
  - Empathy detection in medical training
  - Assessment of therapist adherence to cognitive processing therapy

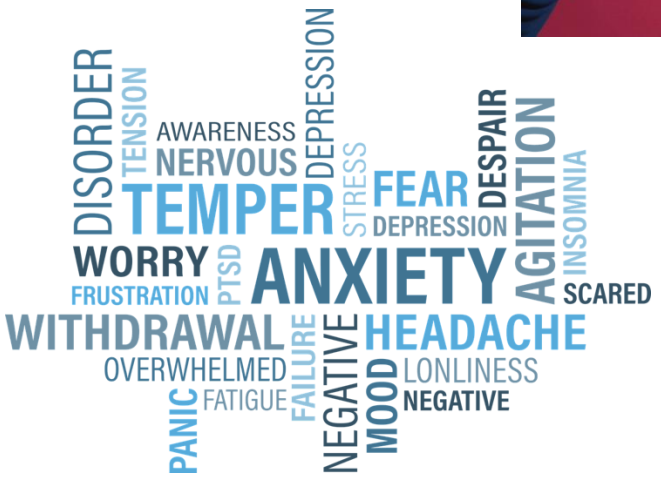
Recognition of Social Signals:

Need for multimodal solutions

Need for Multimodal Solutions

- **Several modalities**
  - Speech
  - Facial expression
  - Text
  - Body posture
  - Physiological signals





Conditions are not aways ideal!

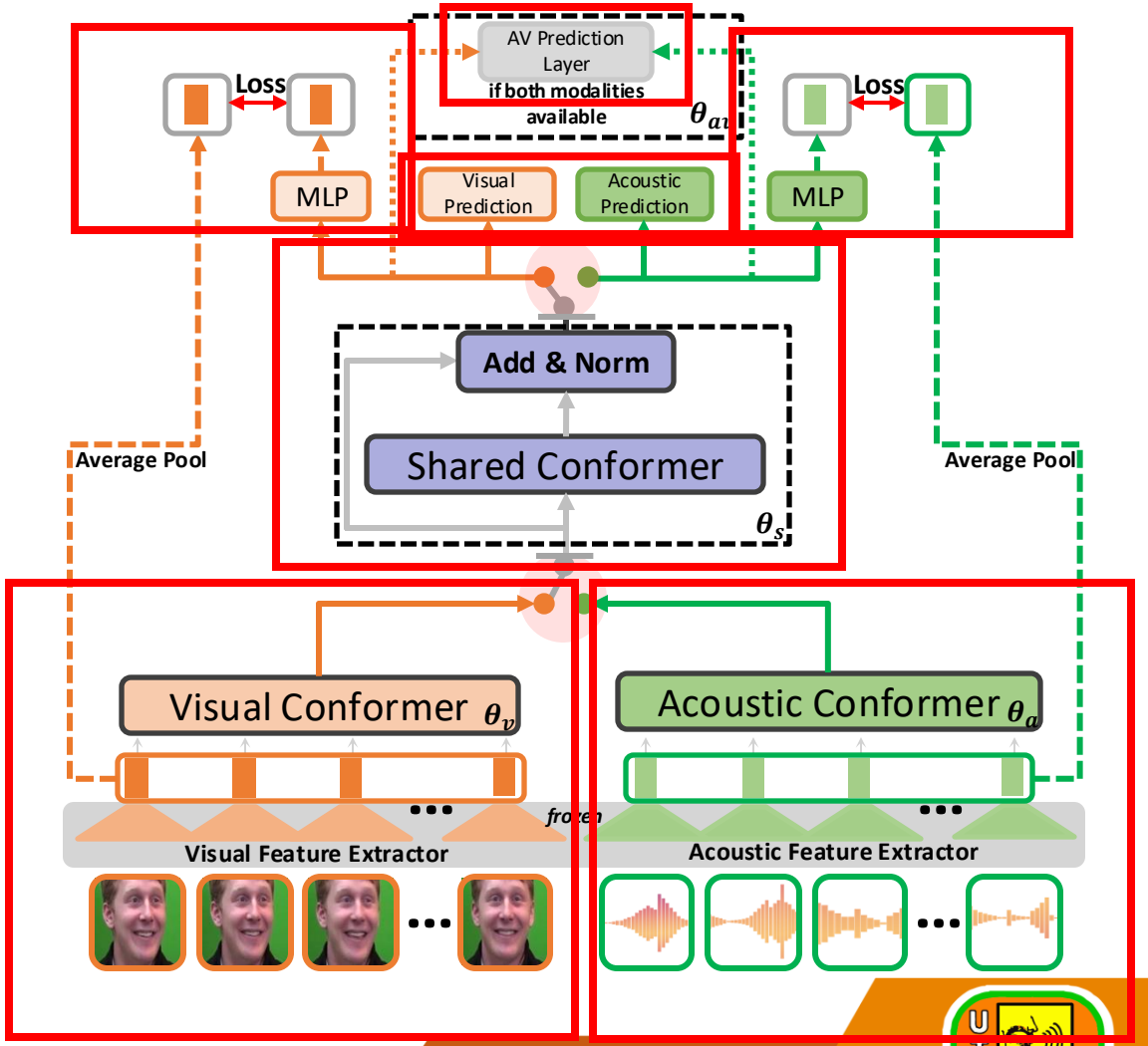
Versatile Audio-Visual Learning (VAVL)

- **Goal:**
  - To handle single and/or multi modalities
  - Framework that can be adapted for emotion regression or classification tasks
  - Need for flexible training even when audio and visual paired data is not available
    - It can use incomplete data
    - It can use unimodal databases

Lucas Goncalves, Seong-Gyun Leem, Wei-Cheng Lin, Berrak Sisman, and Carlos Busso "Versatile audio-visual learning for emotion recognition," IEEE Transactions on Affective Computing, early access, 2024.

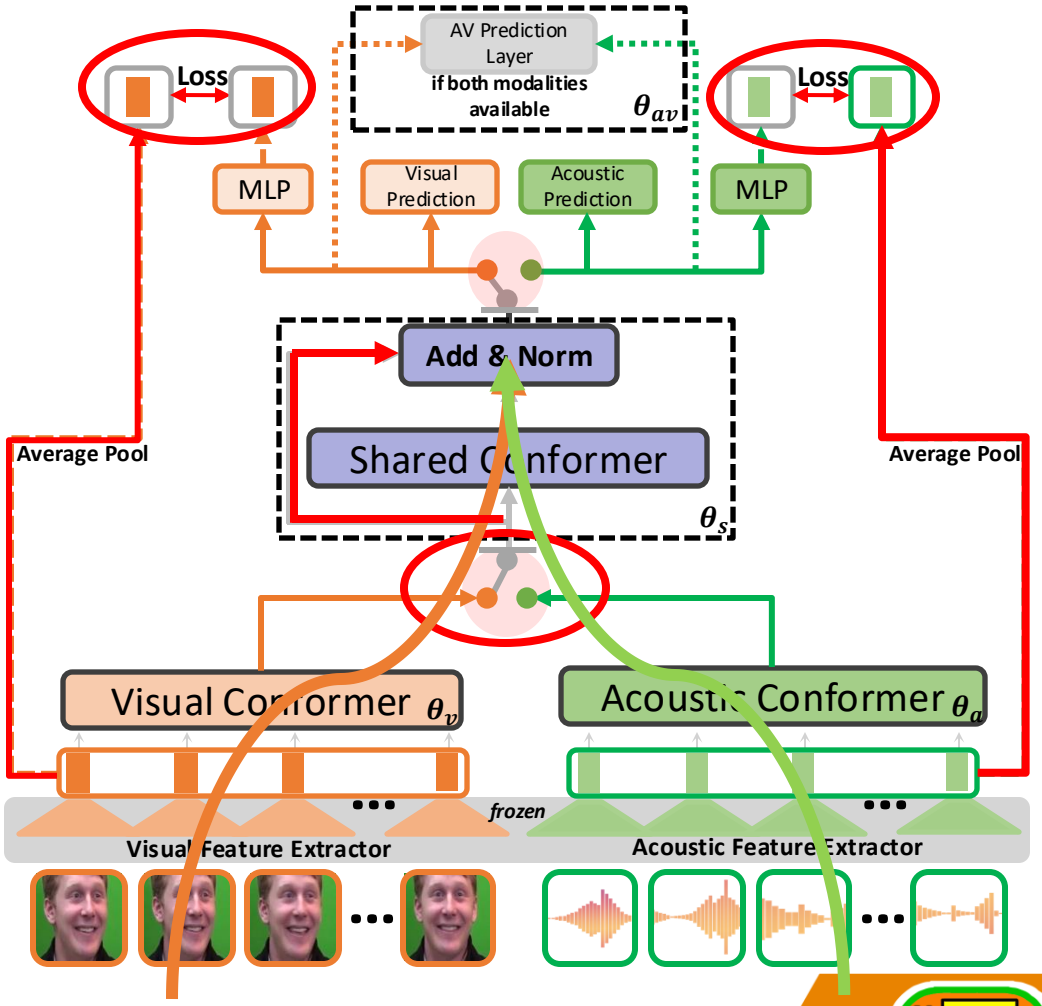
Proposed Framework

- **Versatile Audio-Visual Learning (VAVL)**
  - Visual feature extractor
  - Acoustic feature extractor
  - Shared feature representation
  - Reconstruction blocks
  - Audiovisual prediction layer
  - Unimodal prediction layers



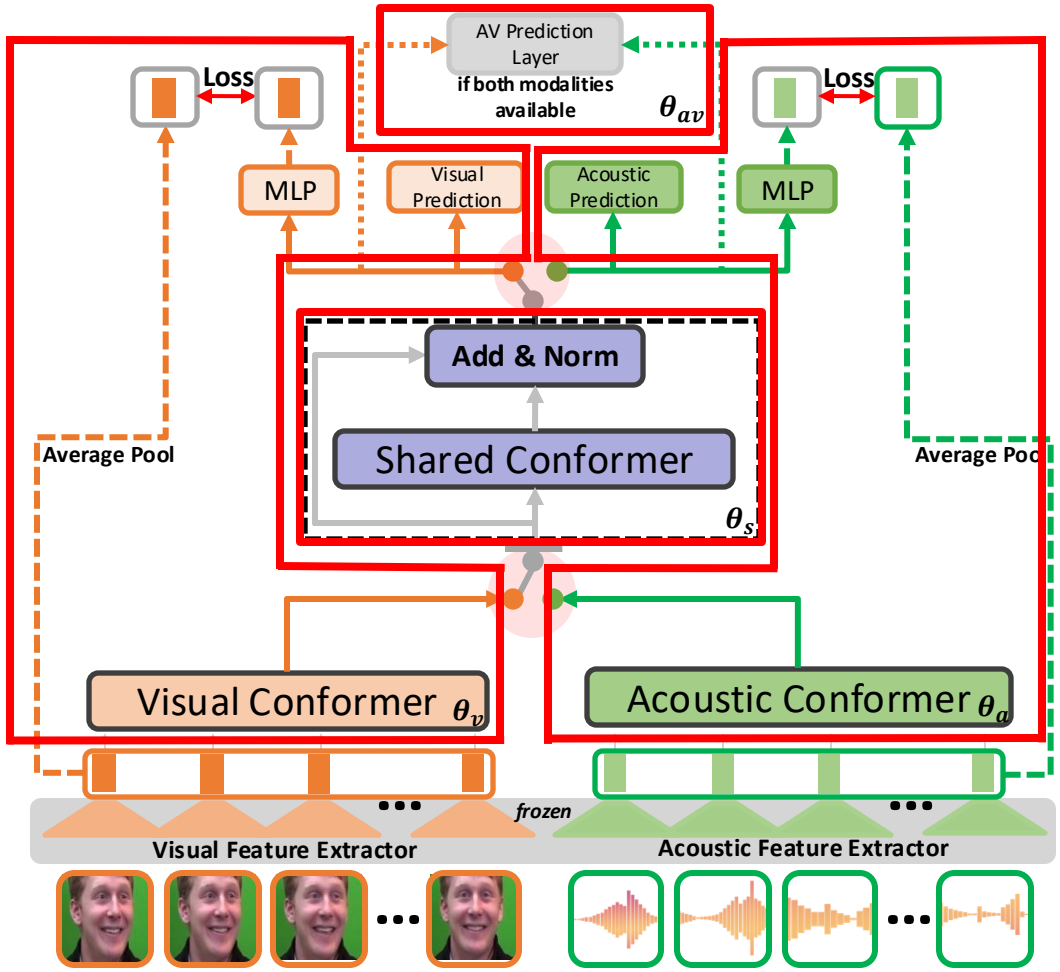
Proposed Framework

- **Versatile Audio-Visual Learning (VAVL)**
  - The shared layers take either acoustic or visual features
  - We use reconstruction layers
  - We use residual connections over:
    - Shared layers
    - Reconstruction of the inputs



Framework Overview

- **Components:**
  - Acoustic layers ($\boldsymbol{\theta_a}$) are updated when audio is presents
  - Visual layers ($\boldsymbol{\theta_v}$) are updated when vision is presents
  - Shared layers ($\boldsymbol{\theta_s}$) are updated anytime audio or visual layers are updated
  - Audio-Visual (AV) prediction layers are only updated when both modalities are present
    - The shared layers ($\boldsymbol{\theta_s}$) are frozen when training Audio-Visual (AV) prediction layers



Experimental Results – Emotional Attributes

- **Observations**
  - Proposed VAVL approach leads to the best performance for most conditions
  - **VAVL** approach achieves competitive performance when using one modality

	MSP-IMPROV								
	Audio-Visual [CCC]			Acoustic [CCC]			Visual [CCC]		
Models	**Aro.**	**Val.**	**Dom.**	**Aro.**	**Val.**	**Dom.**	**Aro.**	**Val.**	**Dom.**
VAVL	**.856***	**.814***	**.876***	**.853***	.783	.858	**.422***	**.375***	**.631***
UAVM	.471	.544	.687	.578	.637	.705	.274	.296	.522
AuxFormer	.672	.652	.820	.722	.730	.789	.363	.293	.581
MulT	.775	.761	.778	-	-	-	-	-	-
Uni. (A)	-	-	-	.841	**.820**	**.878**	-	-	-
Uni. (V)	-	-	-	-	-	-	.383	.321	.598

Baselines:

- UAVM [Gong et al., 2022],
- AuxFormer [Goncalves and Busso, 2022]
- MulT [Tsai et al., 2019] (adapted for audio and facial features only)
- Unimodal baseline model with a similar structure to our method

Experimental Results – Emotional Categories

- **Observations**
  - Similar results
  - VAVL better than other approaches
  - VAVL with unimodal features better than unimodal systems

	CREMA-D					
	Audio-Visual [F1-Score]		Acoustic [F1-Score]		Visual [F1-Score]	
Models	**F1-Ma**	**F1-Mi**	**F1-Ma**	**F1-Mi**	**F1-Ma**	**F1-Mi**
VAVL	**.779***	**.826***	**.628**	**.701***	**.738***	**.787***
UAVM	.749	.769	.614	.554	.617	.672
AuxFormer	.698	.763	.593	.648	.560	.626
MulT	.644	.692	-	-	-	-
Uni. (A)	-	-	.625	.690	-	-
Uni. (V)	-	-	-	-	.725	.783

Baselines:

- UAVM [Gong et al., 2022],
- AuxFormer [Goncalves and Busso, 2022]
- MulT [Tsai et al., 2019] (adapted for audio and facial features only)
- Unimodal baseline model with a similar structure to our method

Recognition of Social Signals:

Important role in modeling interlocutor

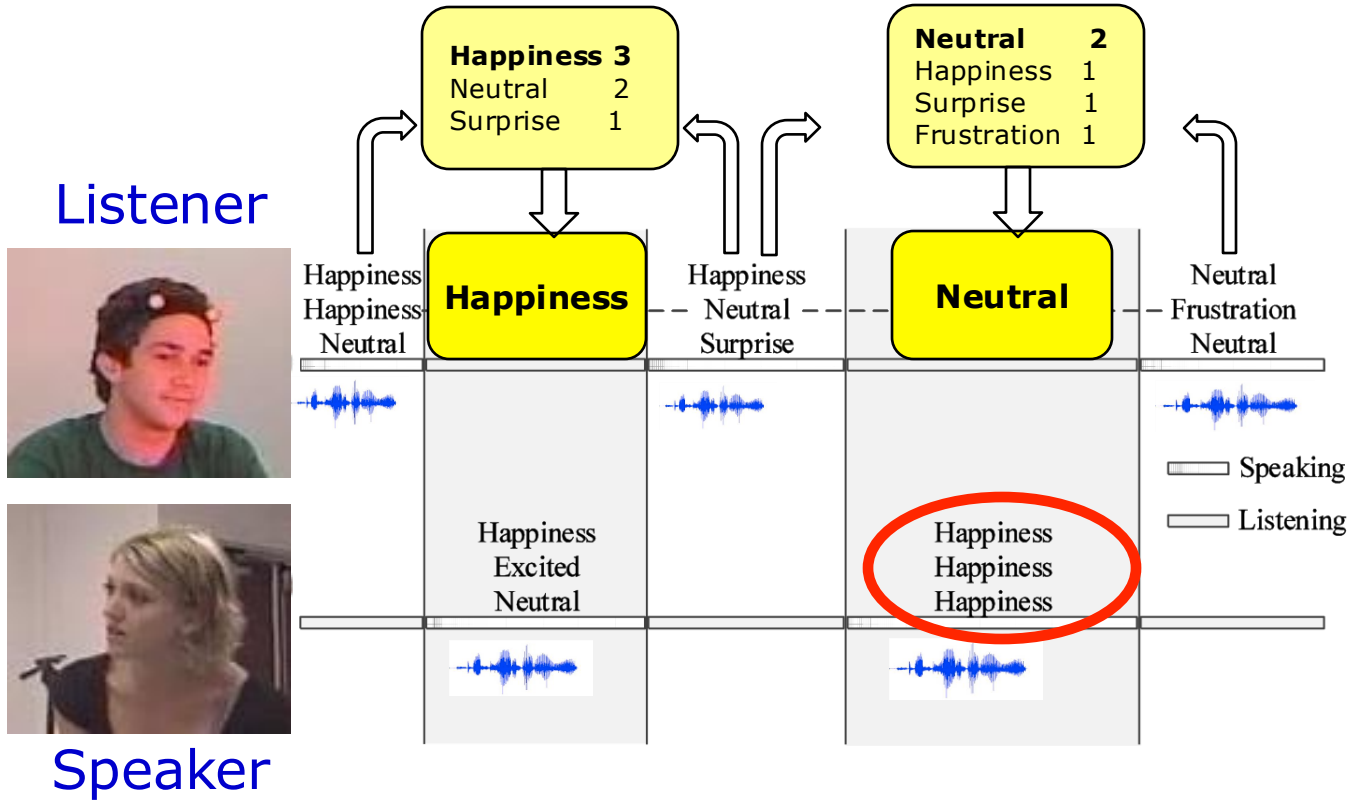
Implications on Emotional Entrainment

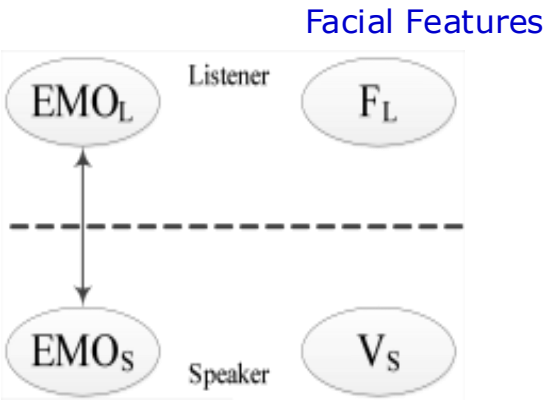
- **Adapting the verbal and non-verbal behaviors to the conversation partner (human or machine)**
  - Linguistic
    - Choice of words [Brennan 1996]
    - Pronunciation [Babel 2009]
    - Speaking rate [Levitan & Hirschberg 2011]
    - fundamental frequency [Coulston et al. 2002]
    - voice quality [Levitan et al. 2011]
  - Gestural
    - Body gesture [Krahmer et al. 2009]
    - Body posture [Chartrand and Bargh 1999]
- **How the emotions of one subjects influence the emotions of the dialog partner? Can we make use of this dependency?**

S. Marioryad and C. Busso, "Exploring cross-modality affective reactions for audiovisual emotion recognition," IEEE Transactions on Affective Computing, vol. 4, no. 2, pp. 183-196, April-June 2013.

Exploring Cross-Modality Affective Reactions

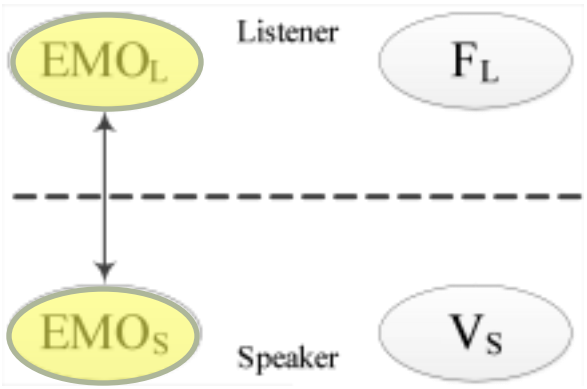
- **Analysis of dyadic interaction**
  - Listening versus speaking turns
  - Interpolation of emotional labels for listening turns





Entrainment Analysis

- **Emotion Entrainment**
  - Co-occurring emotions
    - Emotion listener
    - Emotion speaker

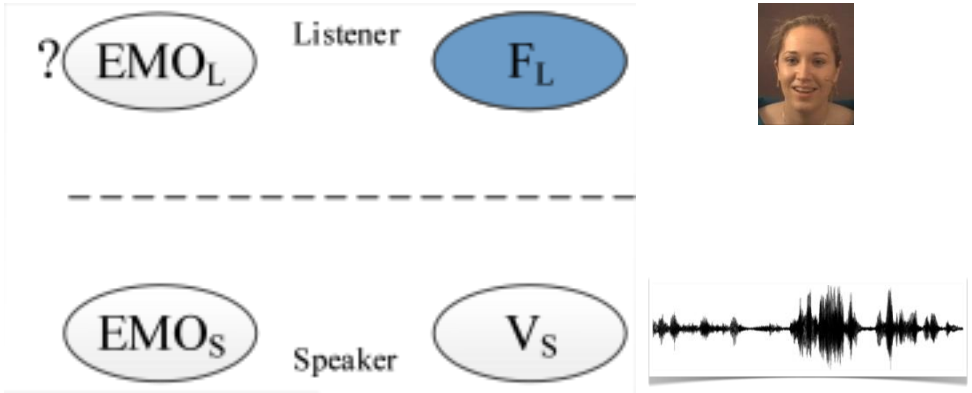


		Listener			
		Ang	Hap	Sad	Neu
Speaker	Ang	**488**	3	11	39
	Hap	4	**113**	13	29
	Sad	4	2	**192**	24
	Neu	81	83	58	**108**

same emotion in 72% of the time (random level is 30%)

Discriminant Analysis (cont.)

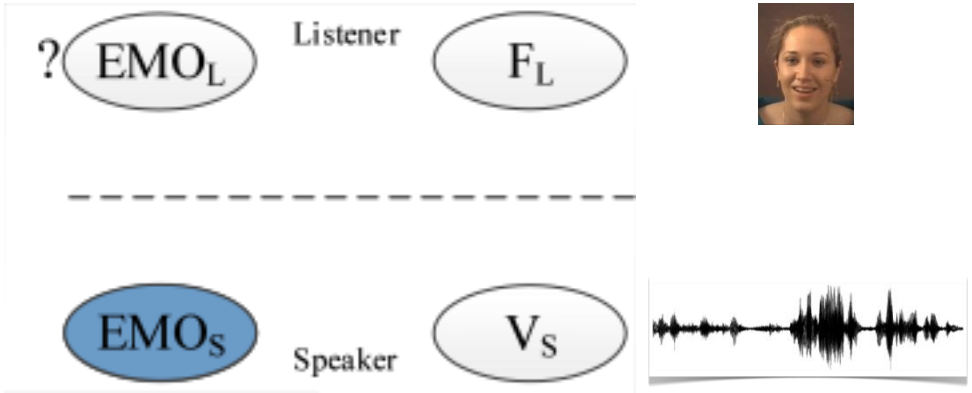
- **Listener**



Method	A	F
SVM_L (F_L) [Baseline]	62.3	52.0
SVM_L (EMO_S)	70.2	64.9
SVM_L (F_L, EMO_S)	72.3	64.9
SVM_L (V_S)	55.0	45.5
Cascade SVM_L (F_L, V_S)	66.2	57.3
SVM_L (F_L, V_S)	67.3	59.5

Discriminant Analysis (cont.)

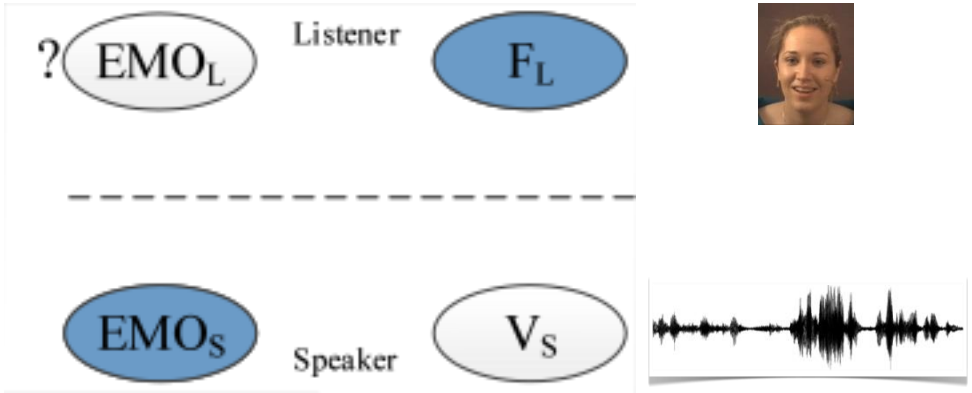
- **Listener**



Method	A	F
SVM_L (F_L) [Baseline]	62.3	52.0
SVM_L (EMO_S)	70.2	64.9
SVM_L (F_L, EMO_S)	72.3	64.9
SVM_L (V_S)	55.0	45.5
Cascade SVM_L (F_L, V_S)	66.2	57.3
SVM_L (F_L, V_S)	67.3	59.5

Discriminant Analysis (cont.)

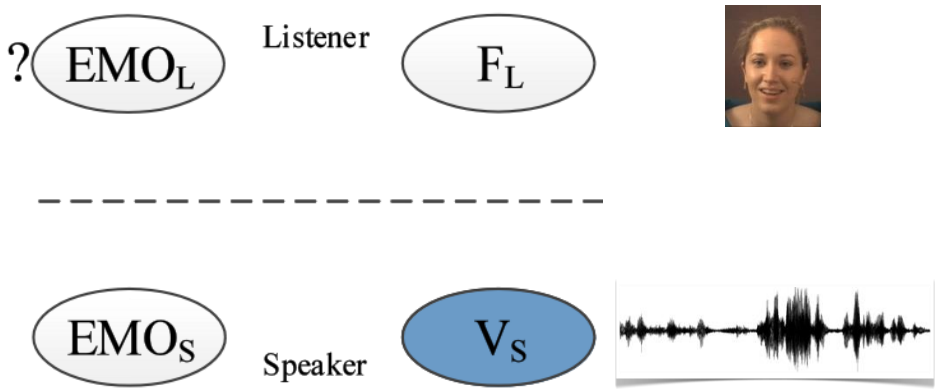
- **Listener**



Method	A	F
SVM_L (F_L) [Baseline]	62.3	52.0
SVM_L (EMO_S)	70.2	64.9
SVM_L (F_L, EMO_S)	72.3	64.9
SVM_L (V_S)	55.0	45.5
Cascade SVM_L (F_L, V_S)	66.2	57.3
SVM_L (F_L, V_S)	67.3	59.5

Discriminant Analysis (cont.)

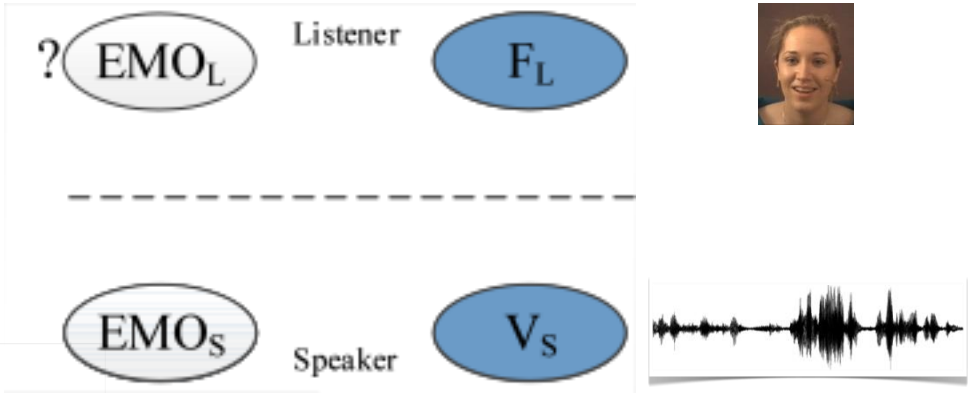
- **Listener**

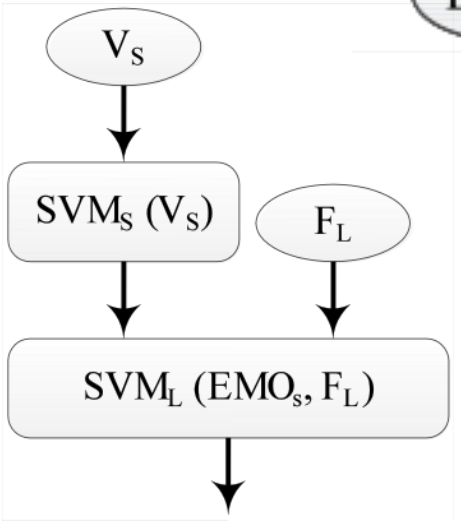


Method	A	F
SVM_L (F_L) [Baseline]	62.3	52.0
SVM_L (EMO_S)	70.2	64.9
SVM_L (F_L, EMO_S)	72.3	64.9
SVM_L (V_S)	55.0	45.5
Cascade SVM_L (F_L, V_S)	66.2	57.3
SVM_L (F_L, V_S)	67.3	59.5

Discriminant Analysis (cont.)

- **Listener**

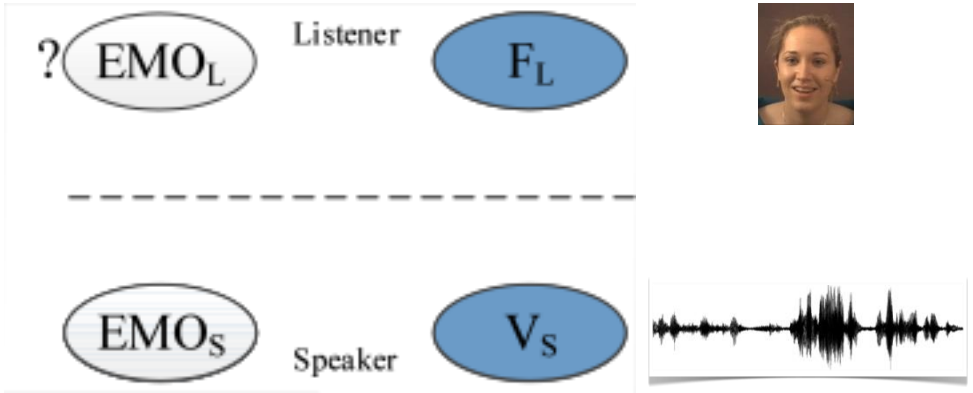




Method	A	F
SVM_L (F_L) [Baseline]	62.3	52.0
SVM_L (EMO_S)	70.2	64.9
SVM_L (F_L, EMO_S)	72.3	64.9
SVM_L (V_S)	55.0	45.5
Cascade SVM_L (F_L, V_S)	66.2	57.3
SVM_L (F_L, V_S)	67.3	59.5

Discriminant Analysis (cont.)

- **Listener**

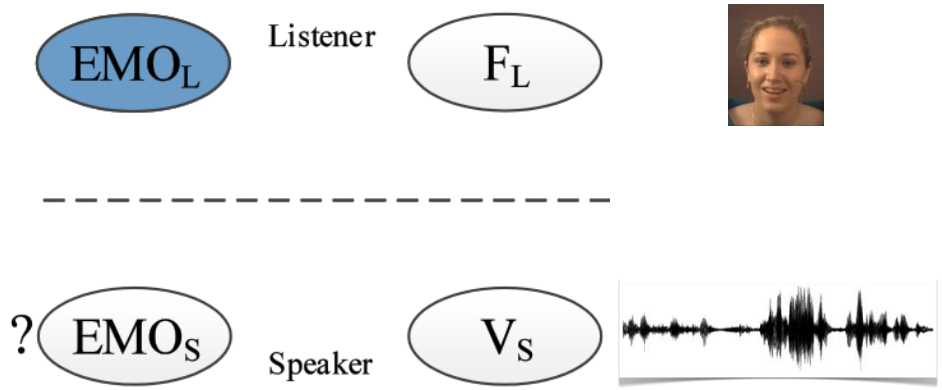


Method	A	F
SVM_L (F_L) [Baseline]	62.3	52.0
SVM_L (EMO_S)	70.2	64.9
SVM_L (F_L, EMO_S)	72.3	64.9
SVM_L (V_S)	55.0	45.5
Cascade SVM_L (F_L, V_S)	66.2	57.3
SVM_L (F_L, V_S)	67.3	59.5

Discriminant Analysis (cont.)

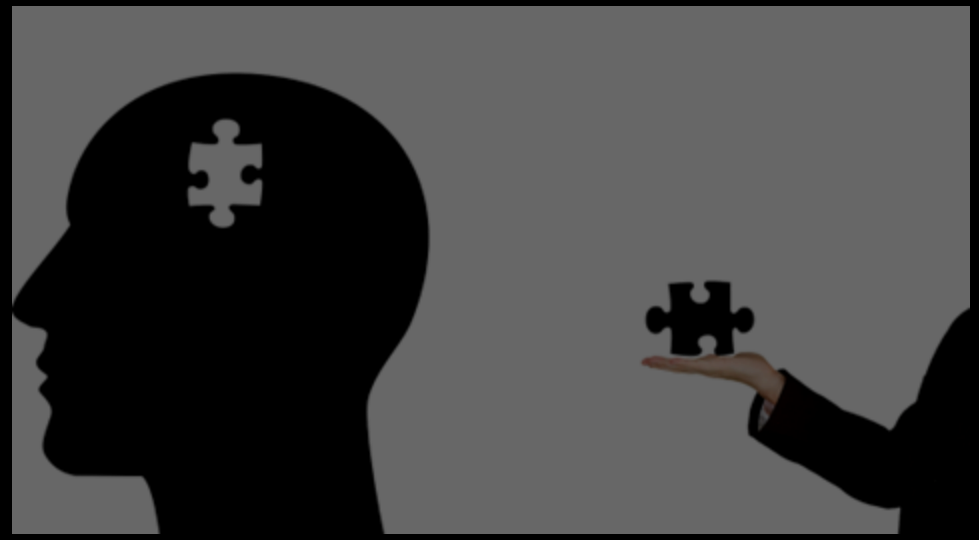
- **Speaker**

Method	A	F
SVM_S (V_S) [Baseline]	54.0	51.8
SVM_S (EMO_L)	72.0	68.1
SVM_S (V_S, EMO_L)	74.0	70.6
SVM_S (F_L)	54.6	46.1
Cascade SVM_S (F_L, V_S)	62.5	59.7
SVM_S (F_L, V_S)	62.3	60.4



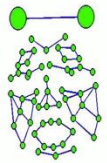
Similar results when we use the listening behaviors to recognize the speaker's emotions

Synthesis of Human Behaviors



Motivation

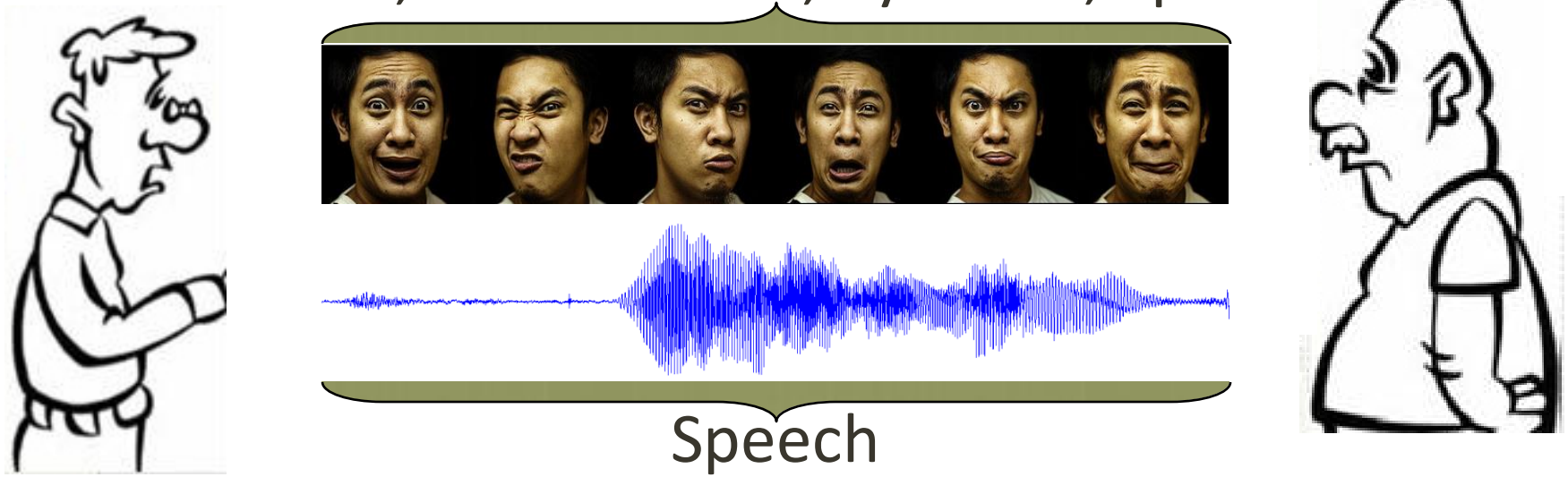
- **Nonverbal behaviors for conversational agents (CAs)**
  - Increase speech intelligibility [Munhall et al., 2004]
  - Increase the level of warmth [van Welbergen et al., 2015]
  - Increase naturalness [Mariooryad and Busso, 2013]
- **Body movements and speech patterns are strongly coupled**
  - More than 90% of human gestures occur while speaking [McNeil, 1992]
- **Believable conversational agents should capture this relationship**
  - Head motion
  - Lip motion
  - Hand gestures
  - Facial expressions



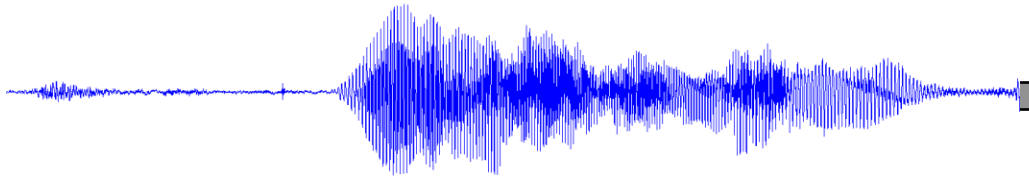
The gesture should match the intended message

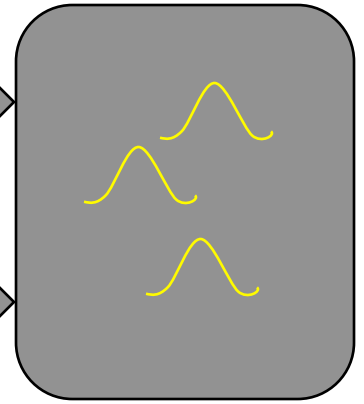
Gesture/Speech Mapping

- **Mapping between gestures and speech can be learned**
- **Gestures can be synthesized from acoustic features**



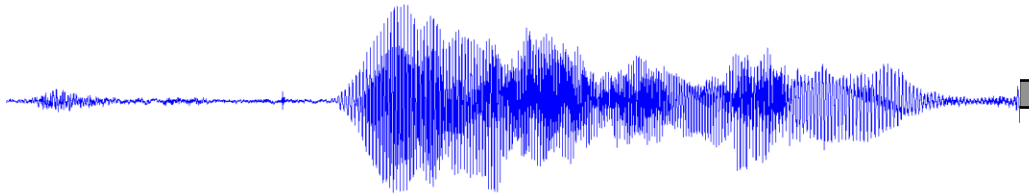
Speech-Driven Facial Animation

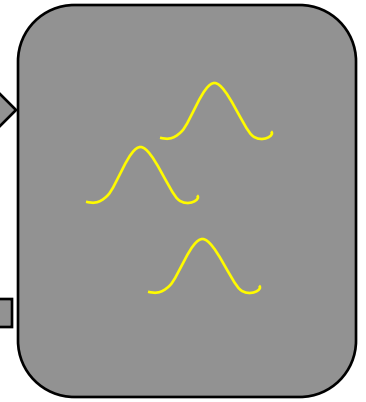




Training

Multimodal Model





Synthesis

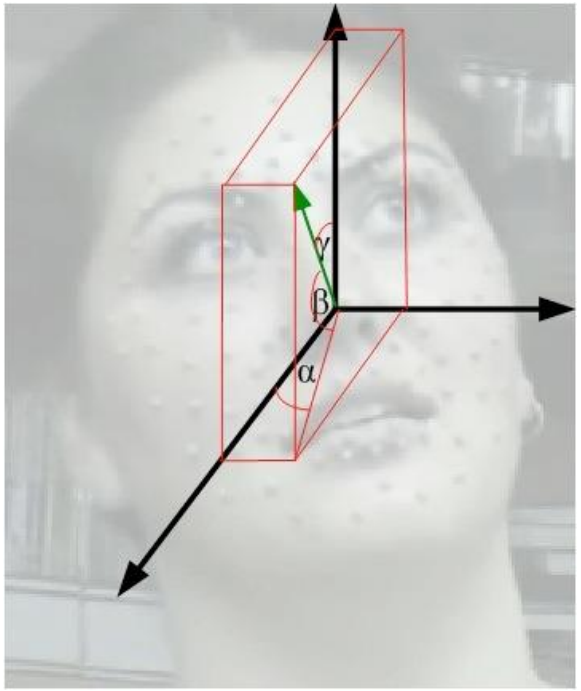
Multimodal Model

Synthesis of Human Behaviors

Speech/data driven animations

Motivation

- **A look at our first attempt: head motion**
- **Why head motion?**
  - Important to acknowledge active listening
  - Improves acoustic perception [Munhall, 2004]
  - Distinguish interrogative and declarative statements [Munhall, 2004]
  - Recognize speaker identity [Hill, 2001]
  - Segment spoken content [Graf, 2002]
  - Important for believable facial animations [Busso, 2007]
  - This data-driven framework is extensible to other gestures



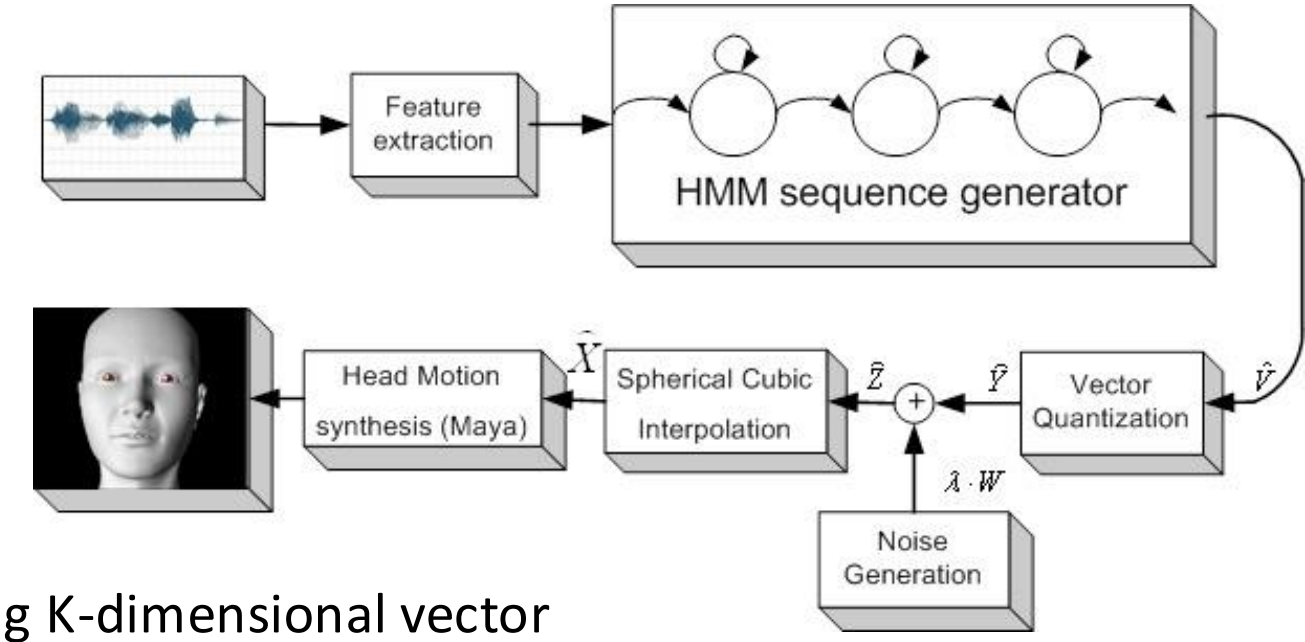
Carlos Busso, Zhigang Deng, Michael Grimm, Ulrich Neumann, and Shrikanth S. Narayanan, "Rigid head motion in expressive speech animation: Analysis and synthesis," IEEE Transactions on Audio, Speech and Language Processing, vol. 15, no. 3, pp. 1075-1086, March 2007.

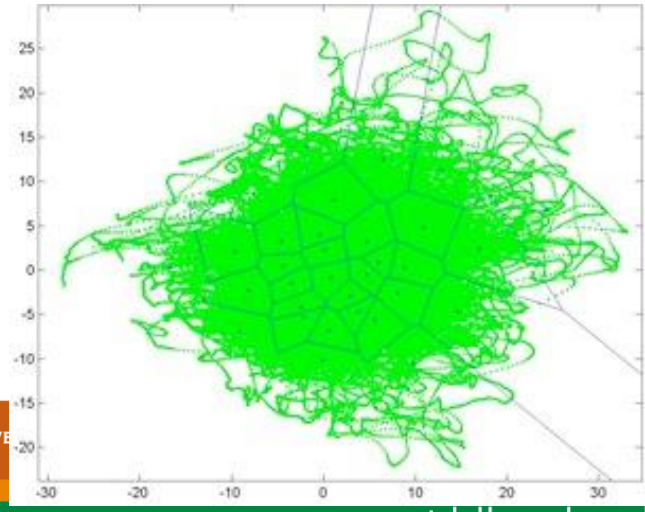
Audio Visual Representation

- **Acoustic Features ~ Prosody (6D)**
  - Pitch
  - RMS energy
  - First and second derivatives
- **Head motion ~ head rotation (3DOF)**
- **Discrete head pose representation**
  - The 3D head motion data is quantized using K-dimensional vector quantization

$$HeadPose = (\alpha, \beta, \gamma) \approx V_i \qquad i \in \{1..K\}$$

  - Each cluster is characterized by its mean, U_i and covariance, Σ_i





Subjective Naturalness Assessment

- **Database:**
  - Facial Motion Capture Database (FMCD) [Busso et al., 2004]
- **17 subjects (1~robot-like 5~human-like)**
  - Head movements increase naturalness
  - Synthesized head movements were perceived as natural as original sequences

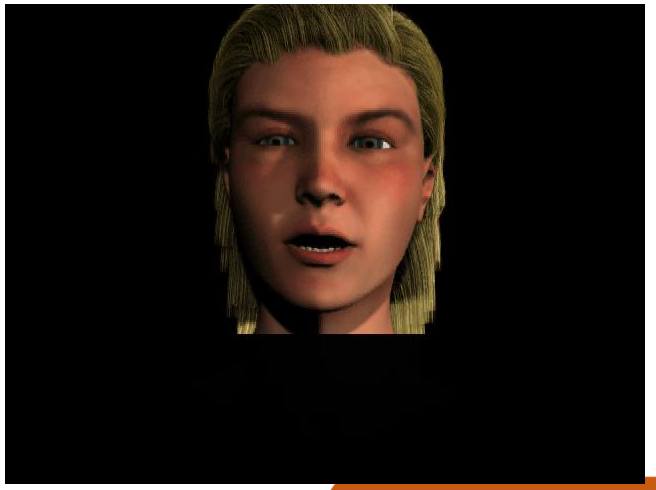
Head Motion	Neutral		Sadness		Happiness		Anger	
Data	Mean	Std	Mean	Std	Mean	Std	Mean	Std
Original	3.76	0.90	3.76	0.83	3.71	0.99	3.00	1.00
Synthesized	4.00	0.79	3.12	1.17	3.82	1.13	3.71	1.05
Fixed Head	3.00	1.06	2.76	1.25	3.35	0.93	3.29	1.45

Results

Neutral

Happiness

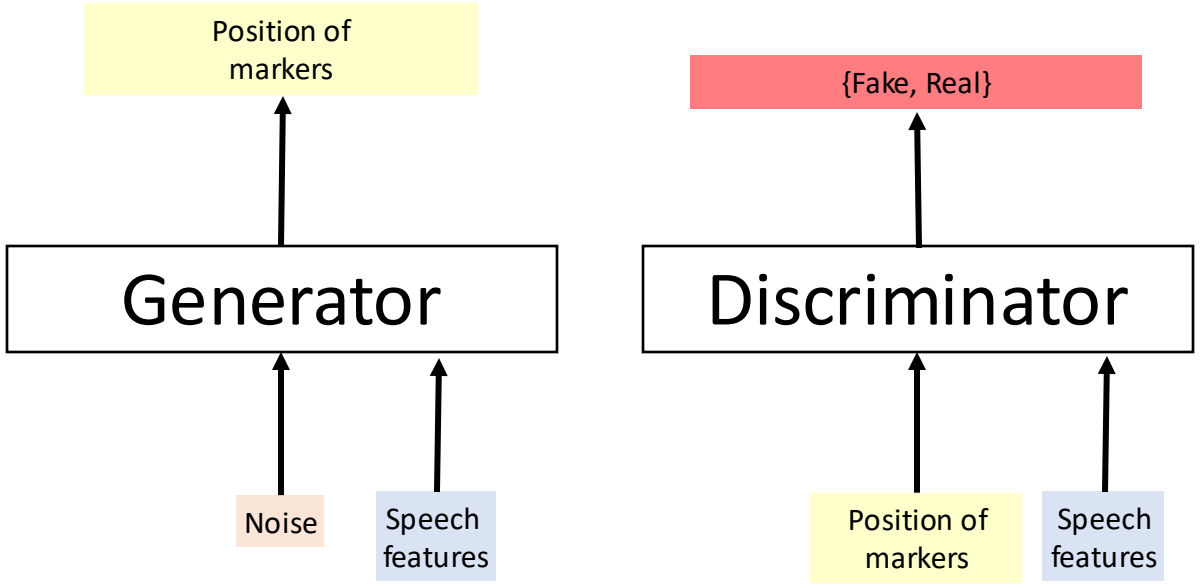
Sadness

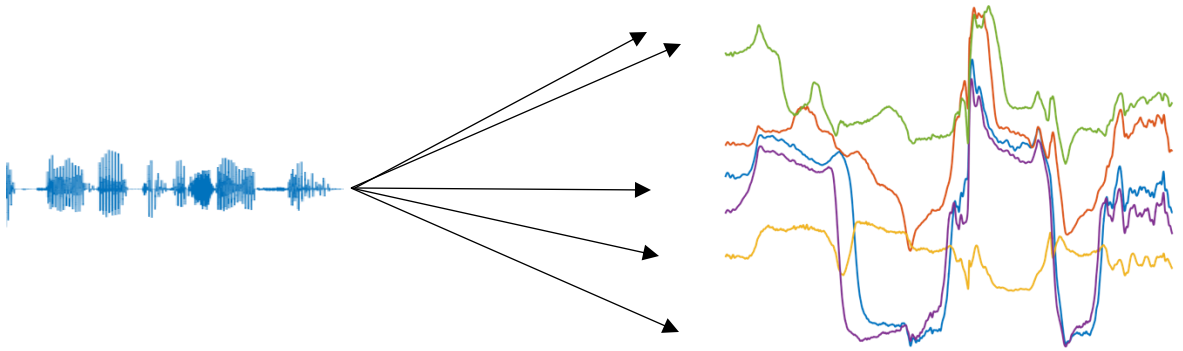


Anger

Head Motion Generation Using Conditional Generative Adversarial Network

- **Advances in generative approaches opened new opportunities**
- **Example: Conditional *Generative Adversarial Networks* (GANs)**



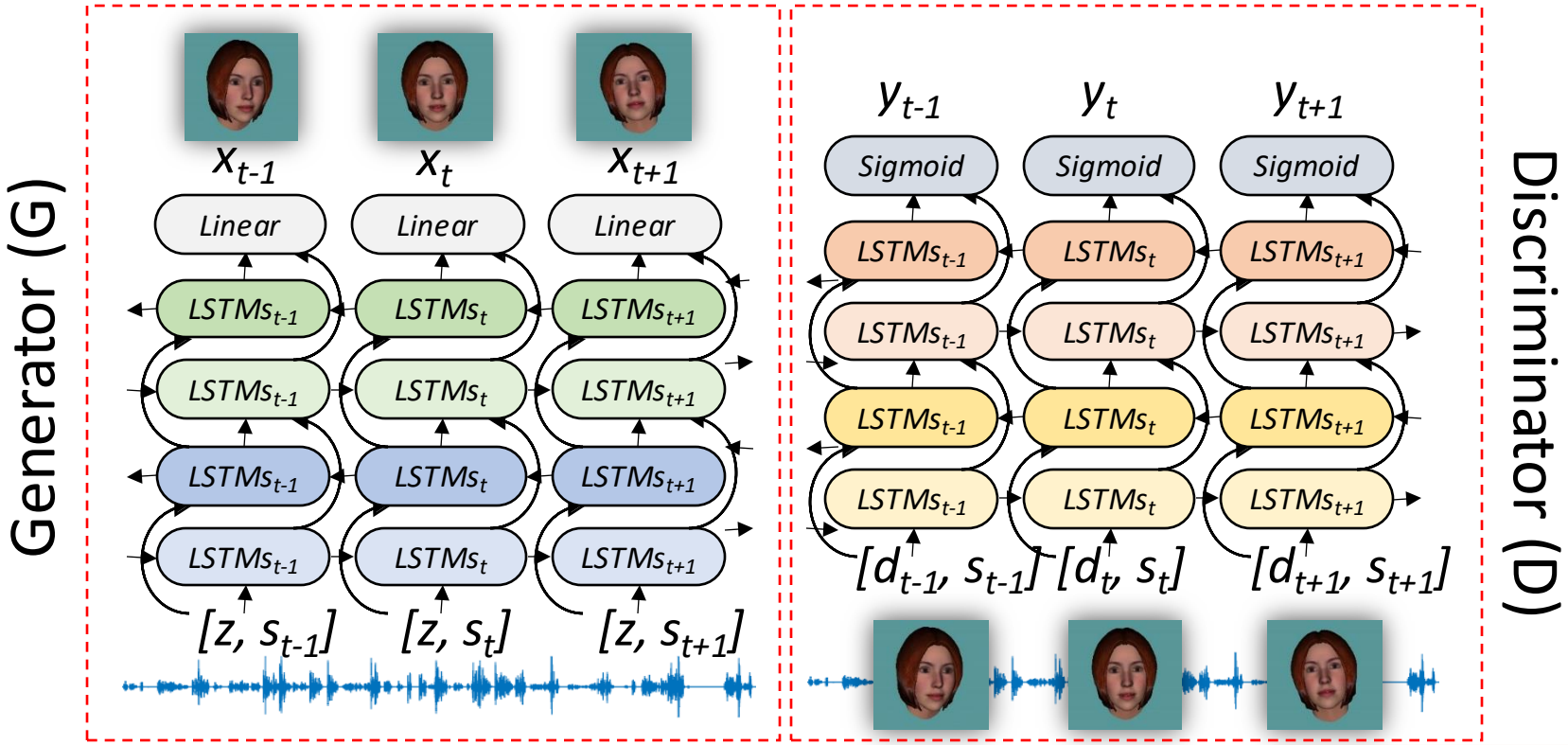


One to many problem

Najmeh Sadoughi and Carlos Busso, "Novel realizations of speech-driven head movements with generative adversarial networks," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018), Calgary, AB, Canada, April 2018, pp. 6169-6173.

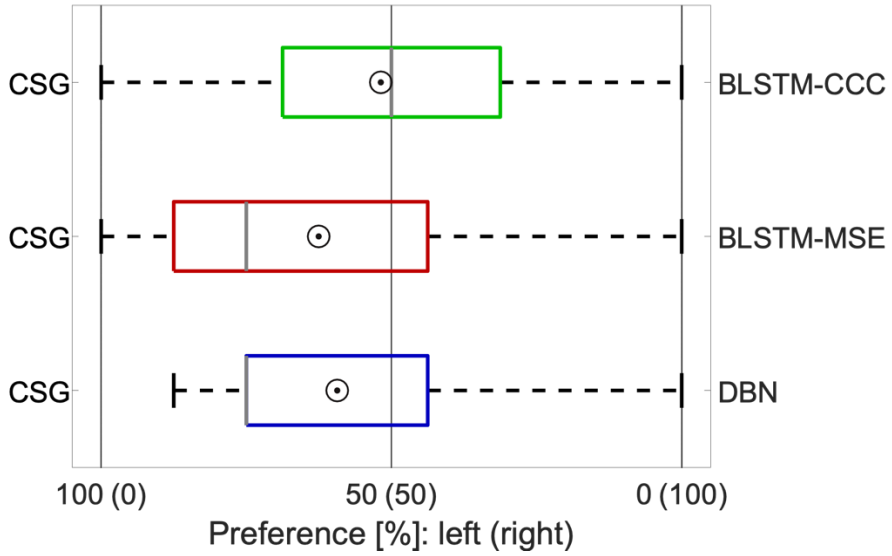
Conditional GANs (CSG) Model

- **Audio:**
  - Speech prosodic features
  - Speaking / listening
- **Motion capture recordings:**
  - Three head angles



Subjective Evaluation

- **12 evaluators**
  - Each comparing 5 pairs of videos, with average length of 39.2s
- **CSG model is the most competitive framework**



Cronbach's alpha:

0.7177

0.6401

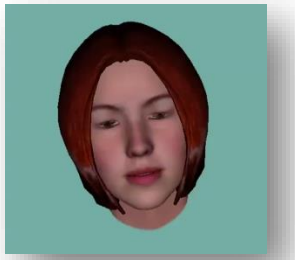
0.4590

Results – Multiple Generated Sequences

z_1

z_2

z_3



z_4

Proposed

DBN

BLSTM-MSE

BLSTM-CCC

Lip Movements Synthesis

- **Can we apply these models to synthesize lip movements?**

Animated movies

Video games

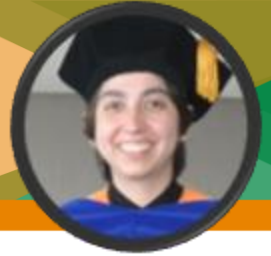
Social Interactive Agents

Animations generated by:

- Actors
- Manually

→ Time consuming

Lip Movements Synthesis

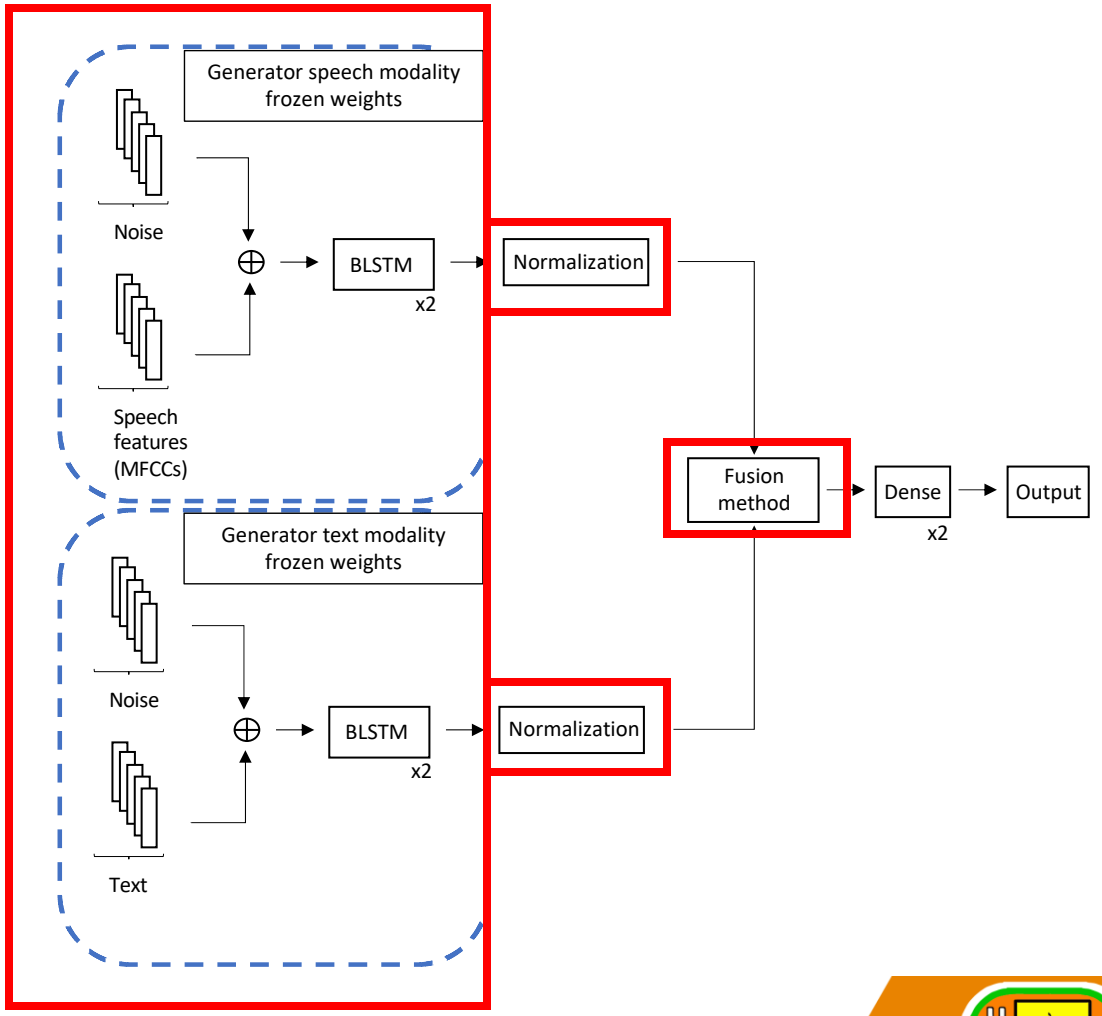


- **We propose a multimodal model that fuses speech and text features**
  - Model:
    - Conditional generative adversarial networks (CGANs)
      - Generator
      - Discriminator
  - Attention mechanism: fuses de modality by determining the relevant temporal information
    - Self-attention
    - Cross-attention
  - Target: Lip landmarks
  - Animation: blendshapes

Andrea Vidal, and Carlos Busso, "Multimodal attention for lip synthesis using conditional generative adversarial networks," Speech communication, vol. 153, pp. 102959, September 2023.

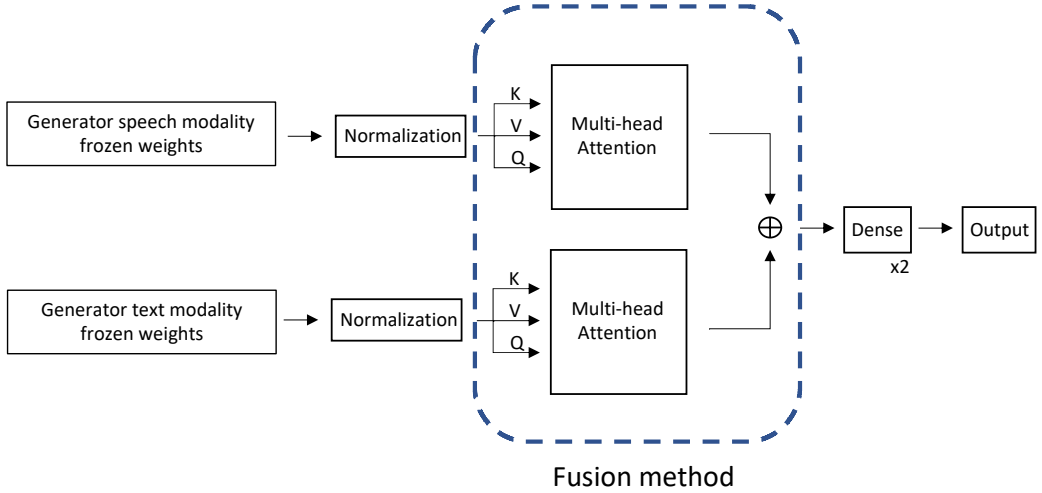
Multimodal CGANs

- **Generator model**
  - Pre-trained generator model for the speech modality
  - Pre-trained generator model for the text modality (ASR outputs)
- **Normalization**
  - Feature normalization by the 98% percentile
- **Fusion method**



Fusion methods

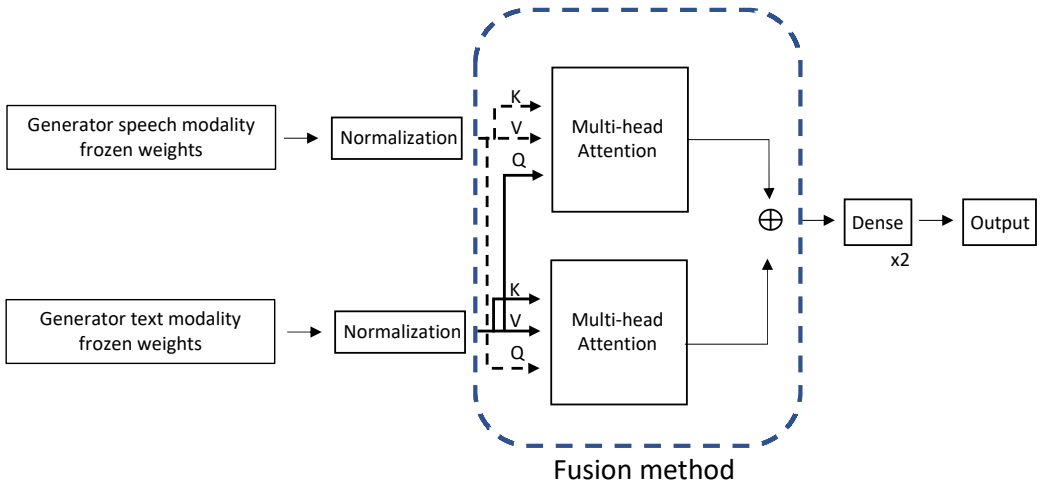
Self-attention method



$$\text{MultiHeadAttention}\left(Q_{\text{text}}, K_{\text{text}}, V_{\text{text}}\right)$$

$$\text{MultiHeadAttention}\left(Q_{\text{speech}}, K_{\text{speech}}, V_{\text{speech}}\right)$$

Cross-attention method



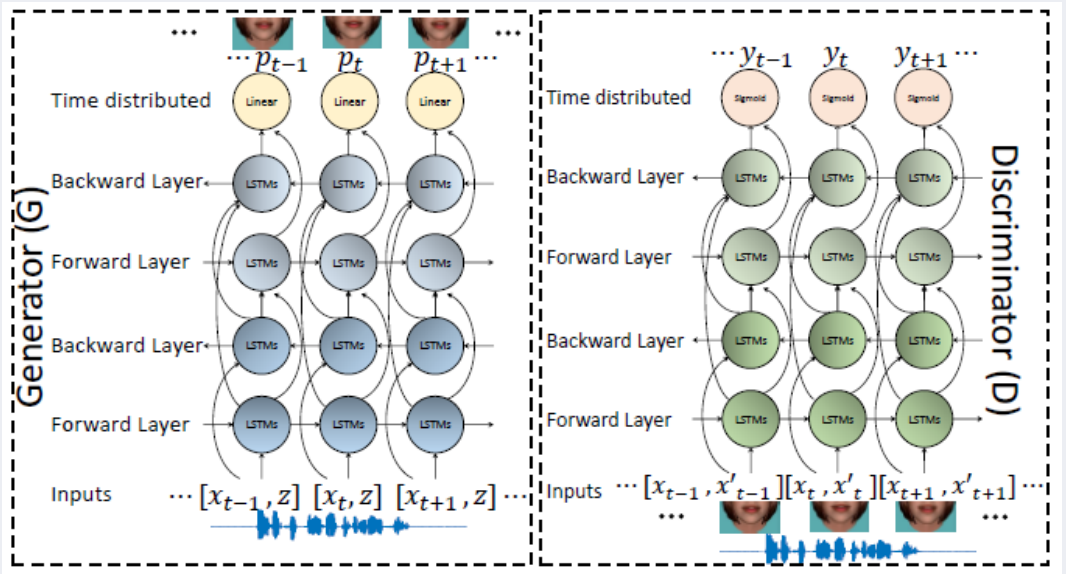
$$\text{MultiHeadAttention}\left(Q_{\text{speech}}, K_{\text{text}}, V_{\text{text}}\right)$$

$$\text{MultiHeadAttention}\left(Q_{\text{text}}, K_{\text{speech}}, V_{\text{speech}}\right)$$

Experimental Settings

- **Baselines**
  - BLSTM text
  - BLSTM speech
  - BLSTM speech+text
  - CGAN text
  - CGAN speech
  - CGAN speech+text

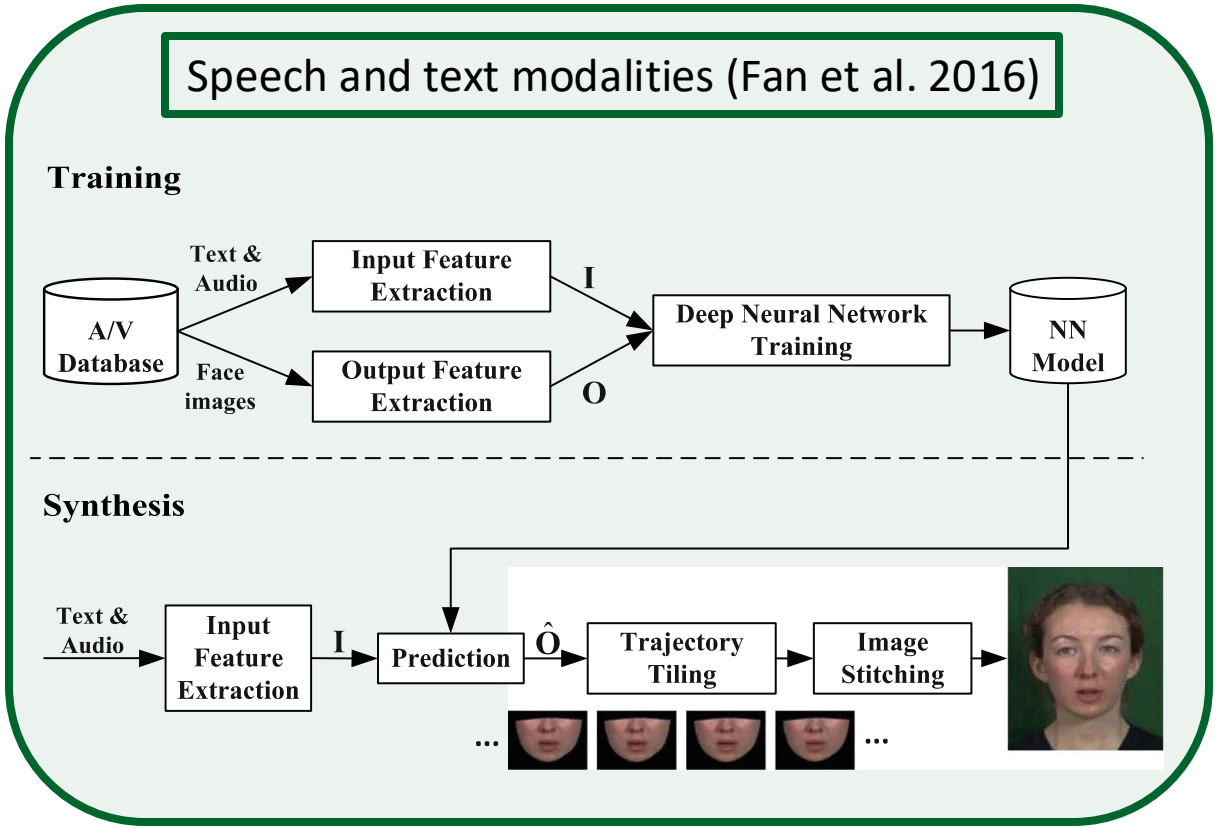
CGAN speech modality (Sadoughi and Busso 2021)



Sadoughi, Najmeh, and Carlos Busso. "Speech-driven expressive talking lips with conditional sequential generative adversarial networks." IEEE Transactions on Affective Computing 12, no. 4 (2019): 1031-1044.

Experimental settings

- **Baselines**
  - BLSTM text
  - BLSTM speech
  - BLSTM speech+text
  - CGAN text
  - CGAN speech
  - CGAN speech+text



Fan, Bo, Lei Xie, Shan Yang, Lijuan Wang, and Frank K. Soong. "A deep bidirectional LSTM approach for video-realistic talking head." Multimedia Tools and Applications 75 (2016): 5287-5309.

Experimental settings

- **Loss functions of the model**
  - Concordance correlation coefficient

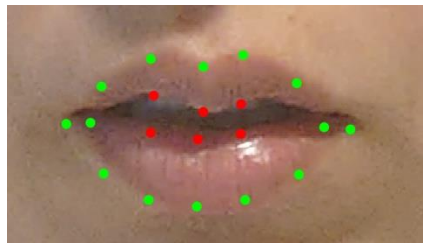
$$CCC = \frac{2 \cdot \sigma_{\hat{y}} \sigma_y}{\sigma_{\hat{y}}^2 + \sigma_y^2 + (\mu_{\hat{y}} - \mu_y)^2}$$

- Correlation between predicted and true landmarks
- Distance between trajectories

  - Mean squared error (only inner landmarks)

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

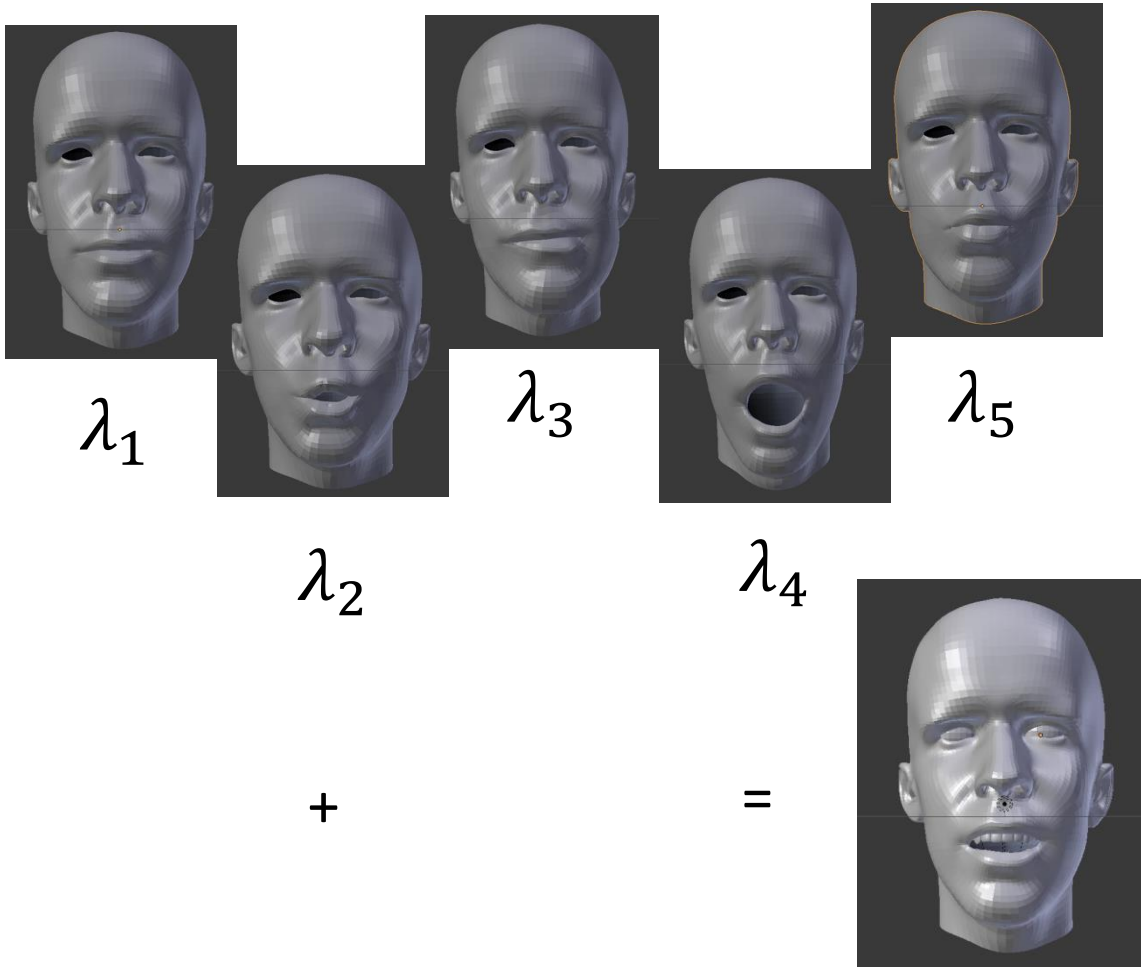
- Capture open and close mouth e.g., /m/, /æ/



$$L = (1 - CCC) + MSE$$

Animation

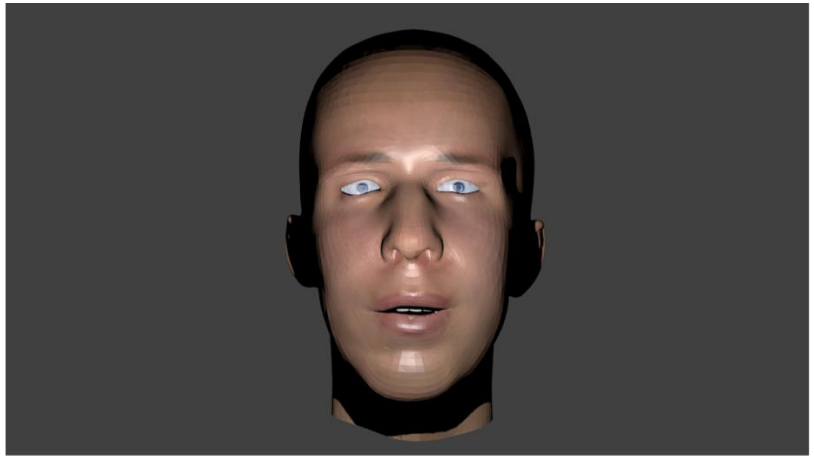
- **Blendshapes**
  - A blendshape model is a combination of target faces
  - 47 blendshapes (FaceWarehouse dataset)
  - Find a linear combination such that the weights for each blendshape ($\lambda_i \in [0,1]$ with $i \in \{1,2,3,4,5\}$)
  - $\sum_i \lambda_i = 1$



FaceWarehouse dataset: Cao, C., Weng, Y., Zhou, S., Tong, Y., & Zhou, K. (2013). Facewarehouse: A 3d facial expression database for visual computing. IEEE Transactions on Visualization and Computer Graphics, 20(3), 413-425.

Subjective Evaluation 1

- **Naturalness of the animation**



How natural was the animation on the video? 1 (not natural) to 10 (natural)

○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9 ○ 10

Method	Mean	Standard deviation
Original	5.4313	2.0851
BLSTM text	4.0750	1.8878
BLSTM speech	3.9875	2.2133
BLSTM speech+text	4.1625	2.0466
CGAN text	5.6750	2.0424
CGAN speech	4.7875	2.2764
CGAN speech+text	5.0000	2.0189
CGAN speech+text (Self-attention)	**5.7875**	2.1447
CGAN speech+text (Cross-attention)	**5.7375**	2.1093

Subjective evaluation 2

- **Video preference**

Method	Mean	Standard deviation
Original	5.4313	2.0851
BLSTM text	4.0750	1.8878
BLSTM speech	3.9875	2.2133
BLSTM speech+text	4.1625	2.0466
CGAN text	5.6750	2.0424
CGAN speech	4.7875	2.2764
CGAN speech+text	5.0000	2.0189
CGAN speech+text (Self-attention)	**5.7875**	2.1447
CGAN speech+text (Cross-attention)	**5.7375**	2.1093

Video 1

Video 2

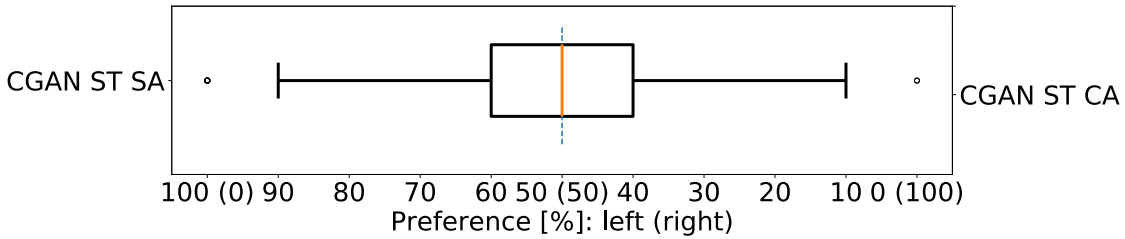
Which animation looks more natural?

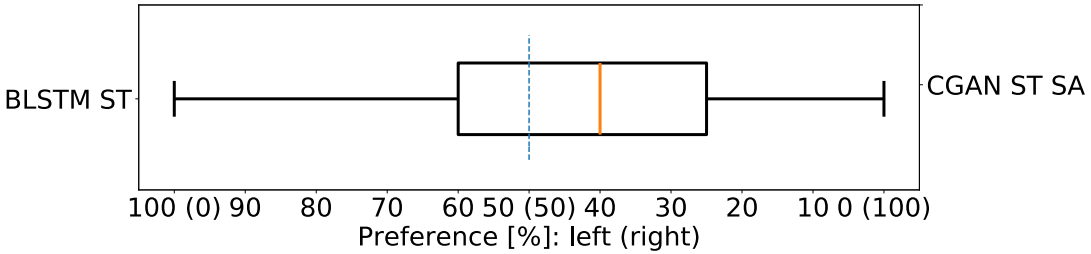
- Definitely Video 1
- Video 1
- Moderately Video 1
- Slightly Video 1
- Both look similar
- Slightly Video 2
- Moderately Video 2
- Video 2
- Definitely Video 2

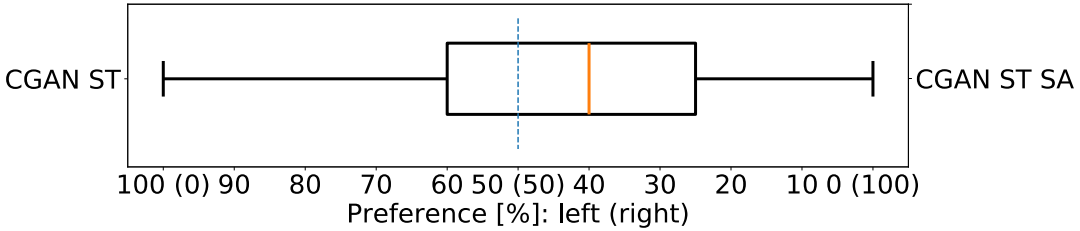
Subjective Evaluation 2

- **Notation**
  - **BLSTM ST**: BLSTM speech-text
  - **CGAN ST**: CGAN speech-text
  - **CGAN ST SA**: CGAN speech-text(self-att)
  - **CGAN ST CA**: CGAN speech-text(cross-att)

Method	Mean	Standard deviation
BLSTM speech+text	4.1625	2.0466
CGAN speech+text	5.0000	2.0189
CGAN speech+text (Self-attention)	**5.7875**	2.1447
CGAN speech+text (Cross-attention)	**5.7375**	2.1093

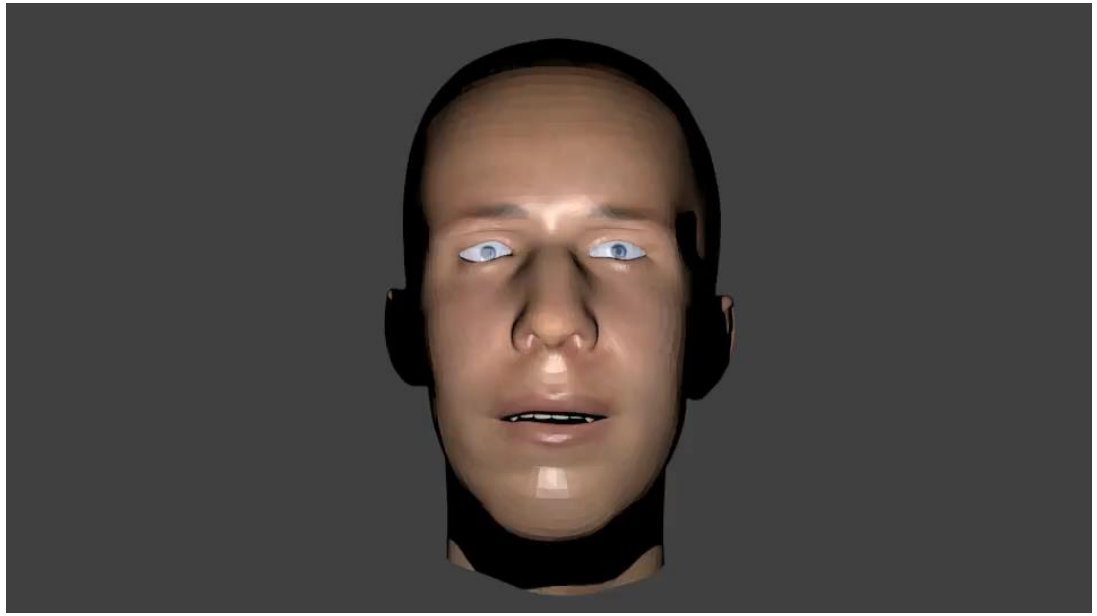






Examples

CGAN ST self-attention



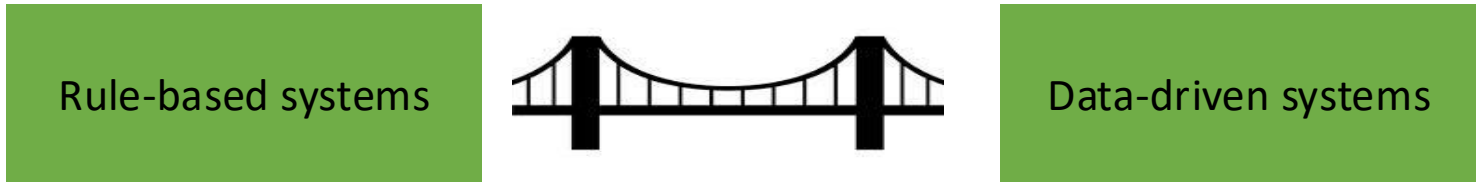
CGAN ST cross-attention

Synthesis of Human Behaviors

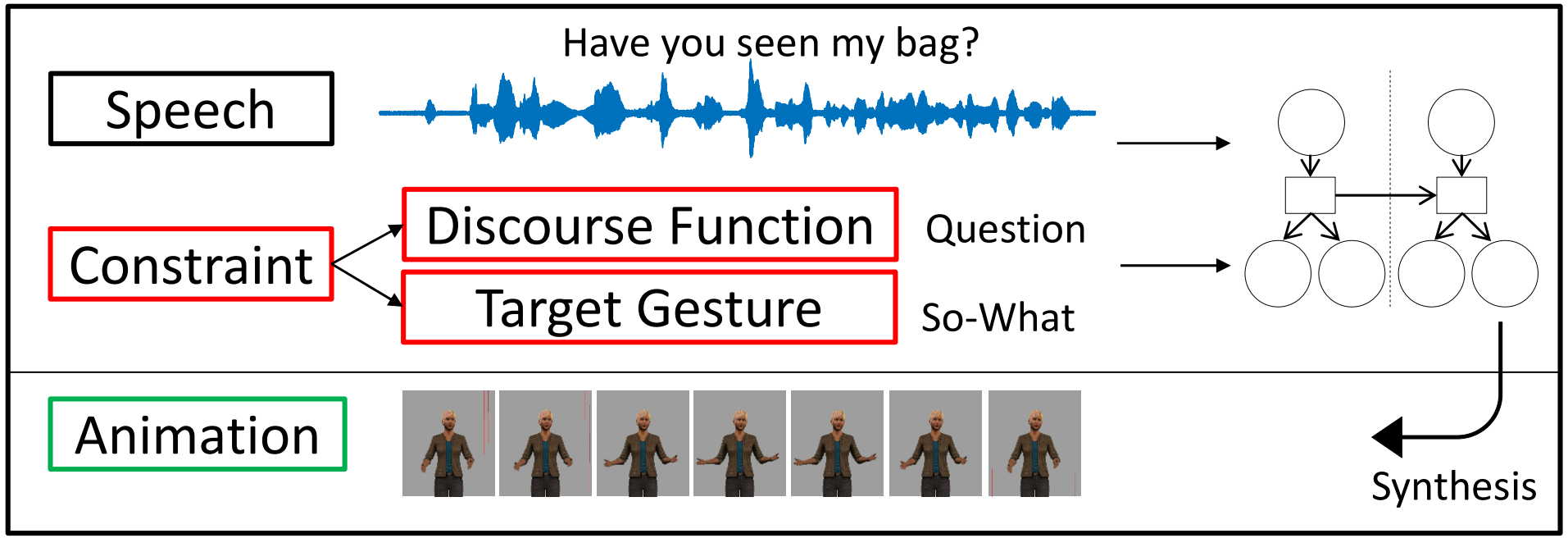
Synthesizing behaviors with meaning

Observations

- **We can capture the *beat* nature of movements**
  - Can we create gestures that capture the meaning of the message?
- **Goal: creating a bridge between rule-based systems and data-driven framework**
  - Synthesizing behaviors that are timely aligned and coordinated with speech
  - Synthesizing behaviors that convey the right meaning
- **Our contributions**
  - Constraining a speech-driven system by discourse function
  - Constraining a speech-driven system by prototypical gestures

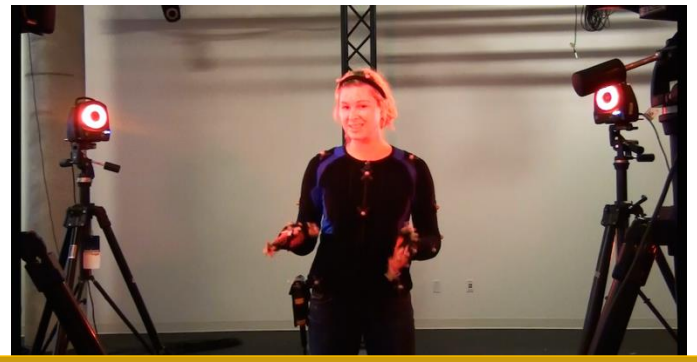


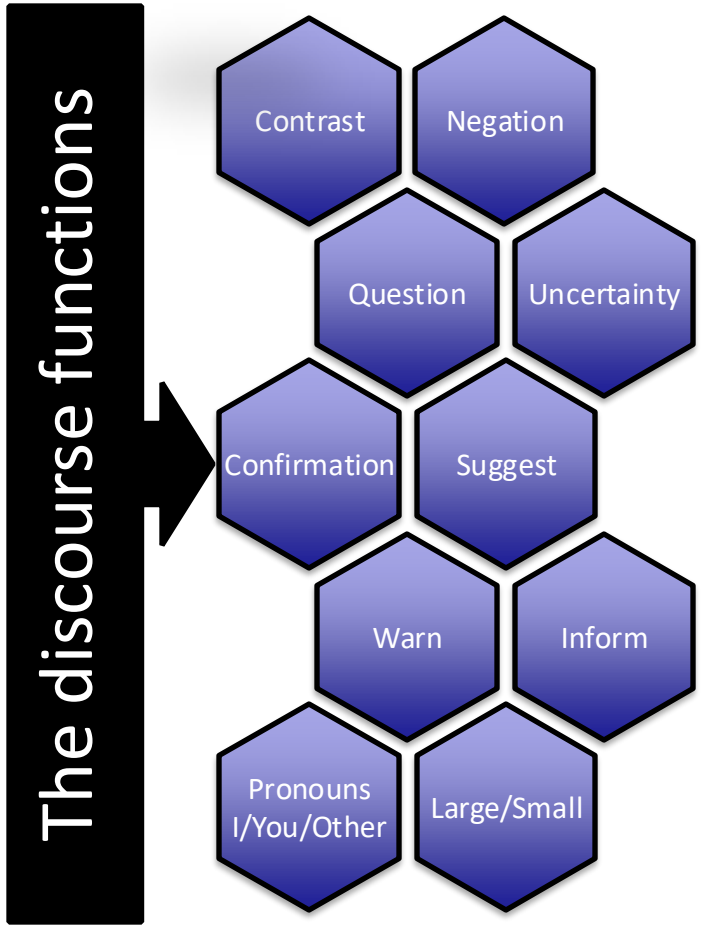
Vision


Have you seen my bag?
Speech
Constraint
Discourse Function
Target Gesture
Question
So-What
Animation
Synthesis

Identifying Relevant Discourse Functions

- **Discourse functions that elicit specific gestural behaviors**
- **Selection guided by previous studies**
  - Poggi et al [2005]
  - Marsella et al. [2013]
- **MSP-Avatar corpus**
  - 2-5 scenarios per discourse function
  - Four dyadic interactions between actors

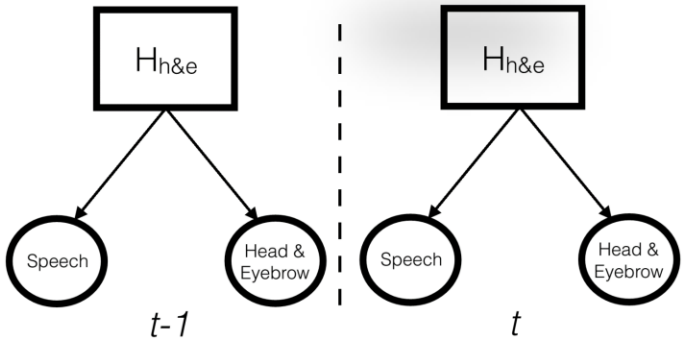




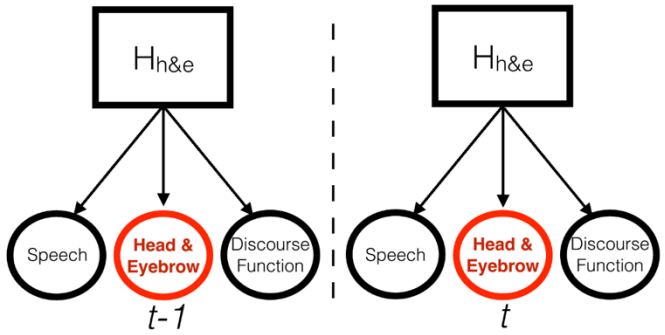
Najmeh Sadoughi, Yang Liu, and Carlos Busso, "MSP-AVATAR corpus: Motion capture recordings to study the role of discourse functions in the design of intelligent virtual agents," in 1st International Workshop on Understanding Human Activities through 3D Sensors (UHA3DS 2015), Ljubljana, Slovenia, May 2015, pp. 1-6.

Speech-driven Model with Dynamic Bayesian Network (DBN)

- **Modeling the joint representation of speech and head/eyebrow movements [Marioryad and Busso, 2013]**
  - Full observation during training
  - Partial observation during testing
- **Adding discourse function as an extra node (child)**
  - A binary variable



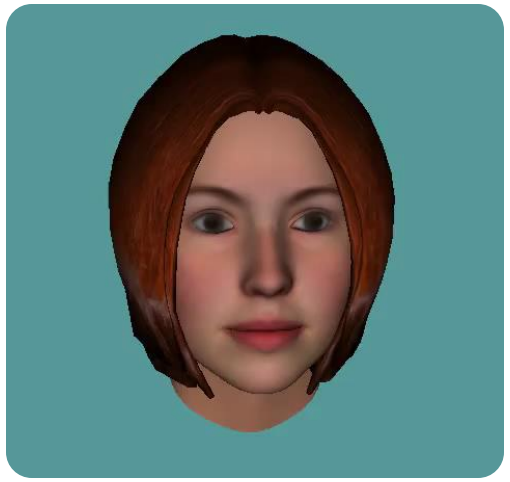
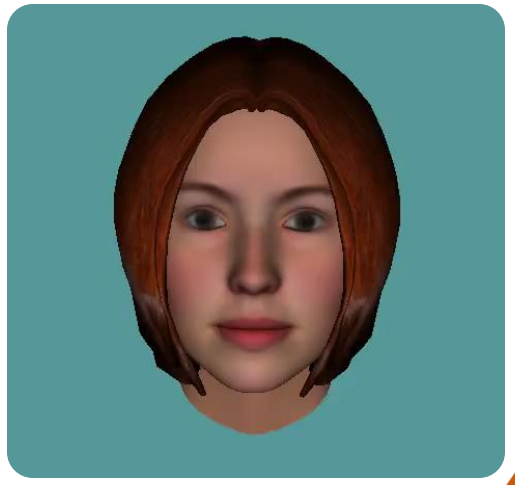
DBN



CDBN

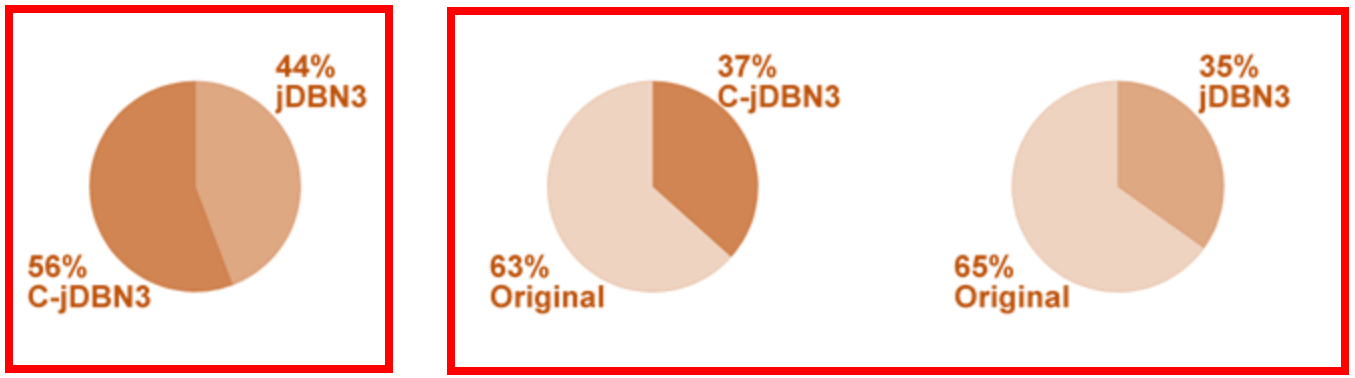
Najmeh Sadoughi and Carlos Busso, "Speech-driven animation with meaningful behaviors," Speech Communication, vol. 110, pp. 90-100, July 2019.

Results

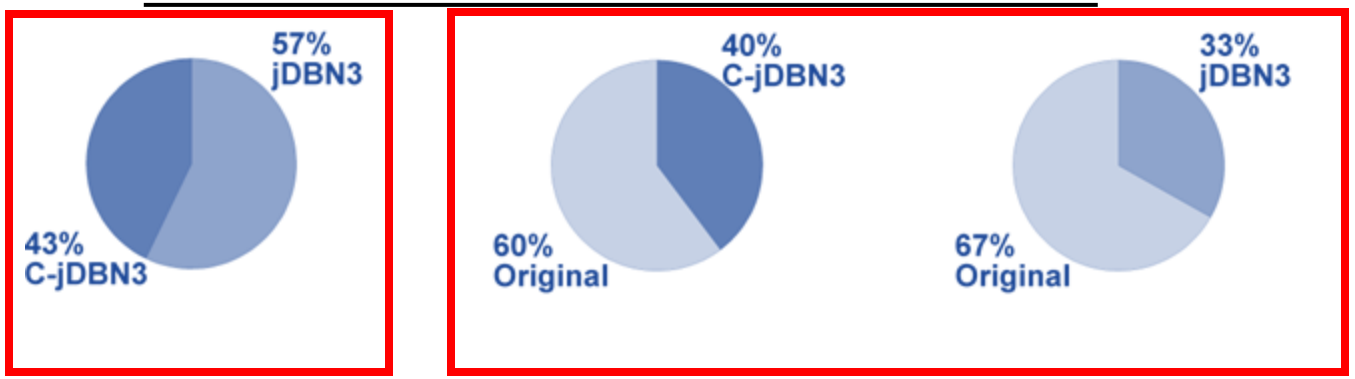
	Unconstrained	Constrained
Affirmation		
Question		

Subjective Results

Constraint is "Question"



Constraint is "Affirmation"



- **Three settings**
  - Original movements
  - Model without constraint (jDBN3)
  - Model with constraint (C-jDBN3)
- **Which video do you prefer?**
  - Direct comparisons
  - Indirect comparisons

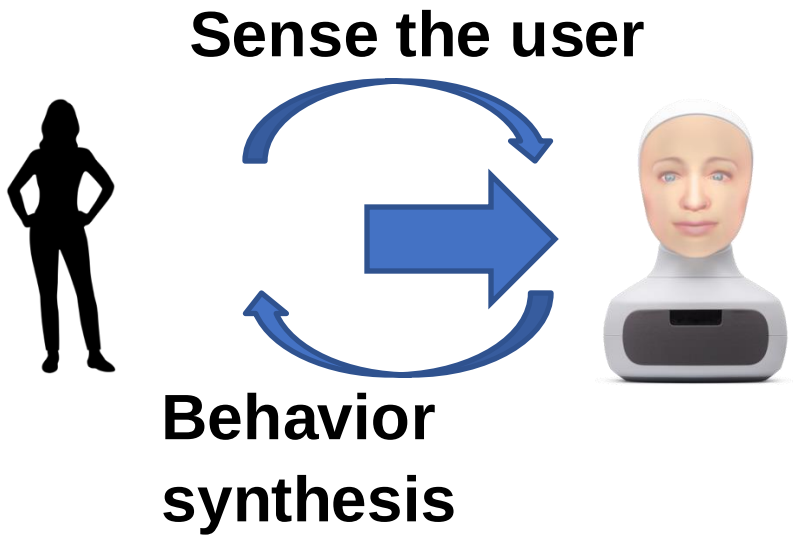
Discourse function constraints improve the generated sequences

Synthesis of Human Behaviors

Modeling of Entrainment

Need for Closing-the-Loop

- **The introduction of *social interactive agents (SIAs)* to daily activities that requires understanding human interaction**
  - The SIAs' coarticulation is important but we need to consider other factors
  - SIAs have to behave as natural as possible during an interaction and keep the interest in a conversation



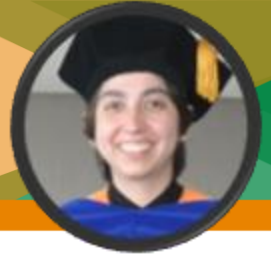
- This can be achieved by entraining the interlocutor in a conversation

Study of Entrainment

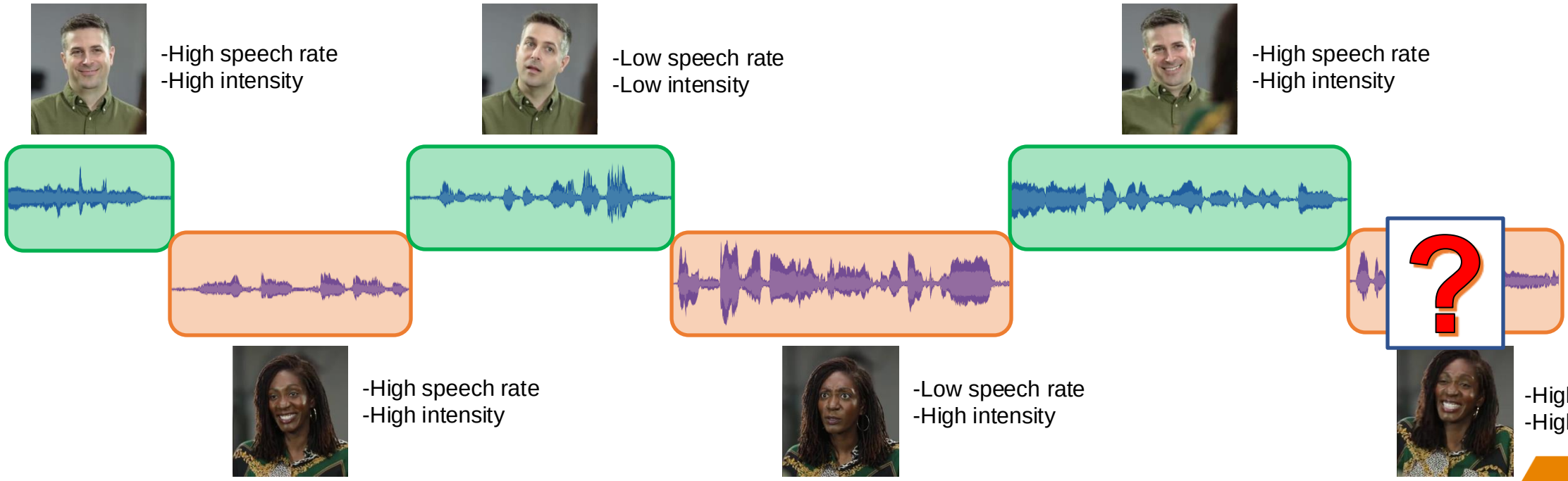
- **Entrainment:** "Phenomenon where the interlocutors adapt their behavior, which through the interaction becomes similar"
- **People tend to adapt their verbal and non-verbal behaviors**
  - spoken style, gestures, and expressions
- **Entrainment has been observed in human-computer interaction**
  - Mirror behaviors
  - Create entrainment behaviors
- **Entrainment can provide a mechanism to increase engagement by generating intentional behaviors for SIAs**

Key challenge:
Find relevant events that trigger entrainment behaviors

Predicting Events



- **Each turn is represented by behaviors related to:**
  - Speech
  - Facial gestures

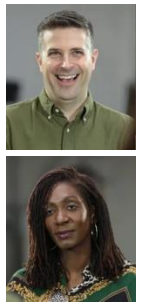


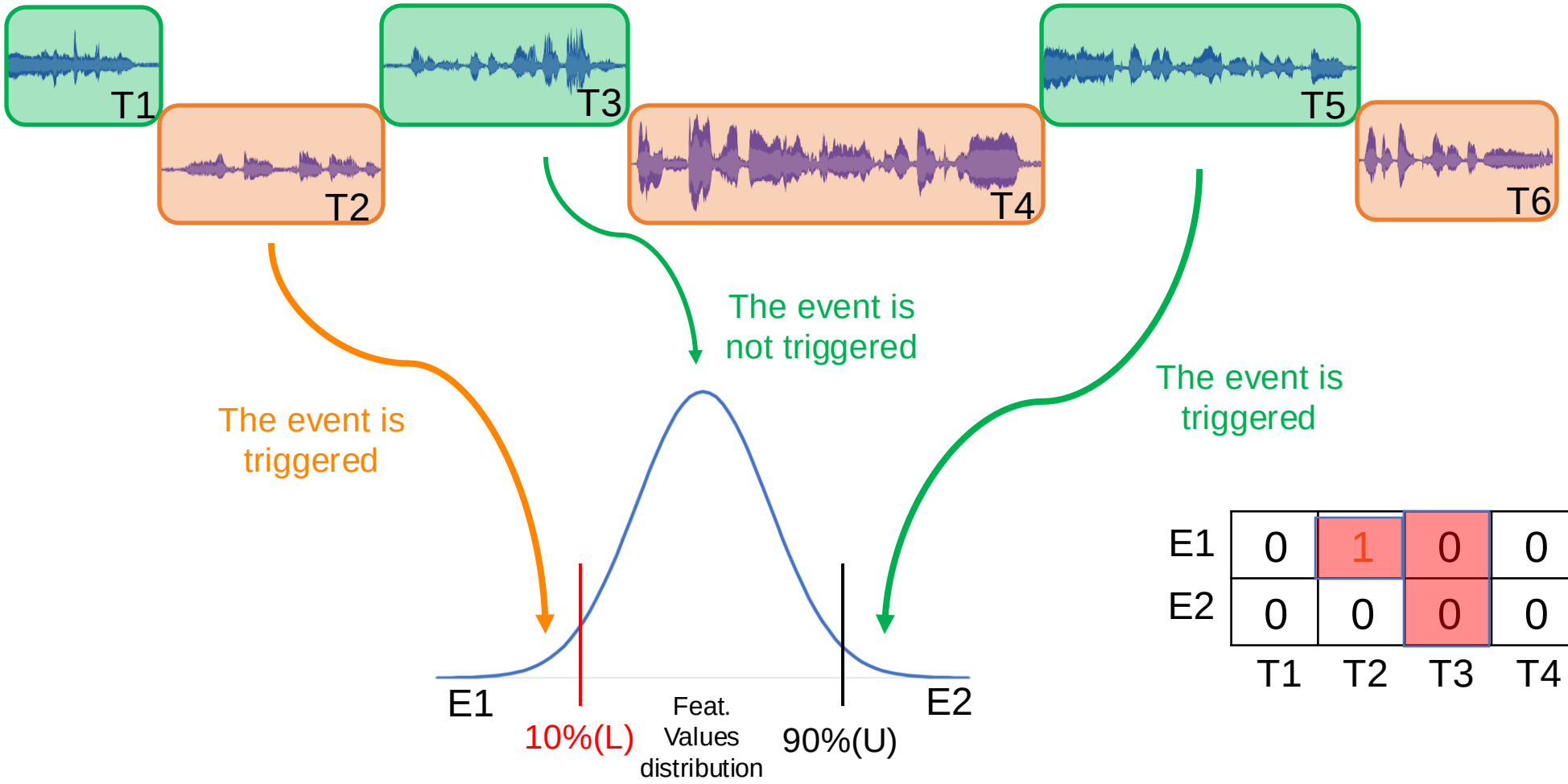
Andrea Vidal and Carlos Busso. Predicting behaviors in dyadic interactions using autoregressive transformer model. In preparation

MSP-Entrainment Corpus

- **Dyadic interaction database**
  - 273 videos
  - Duration 45.96 hrs.
- **Diversity of speakers**
  - Number of speakers 460
    - 268 females
    - 181 males
    - 11 non-binary
- **Characteristics**
  - Clear frontal face
  - Clear audio

Creation of *communicative events*





	T1	T2	T3	T4	T5	T6
E1	0	1	0	0	0	0
E2	0	0	0	0	1	0

Communicative Events – Audio Events (1-10)

- **Speech features (Gregory et al. 1993, Levitan and Hirschberg 2011 , Levitan et al. 2016, Beňuš et al. 2018)**
  - Fundamental frequency
  - Intensity
  - Speech rate

Events	Name
1	High speech rate
2	Low speech rate
3	High F0
4	Low F0
5	High intensity
6	Low intensity
7	High variance F0
8	Low variance F0
9	High variance intensity
10	Low variance intensity

Speech Events

Communicative Events – Visual Events (11-28)

- **We define 28 *communicative events***

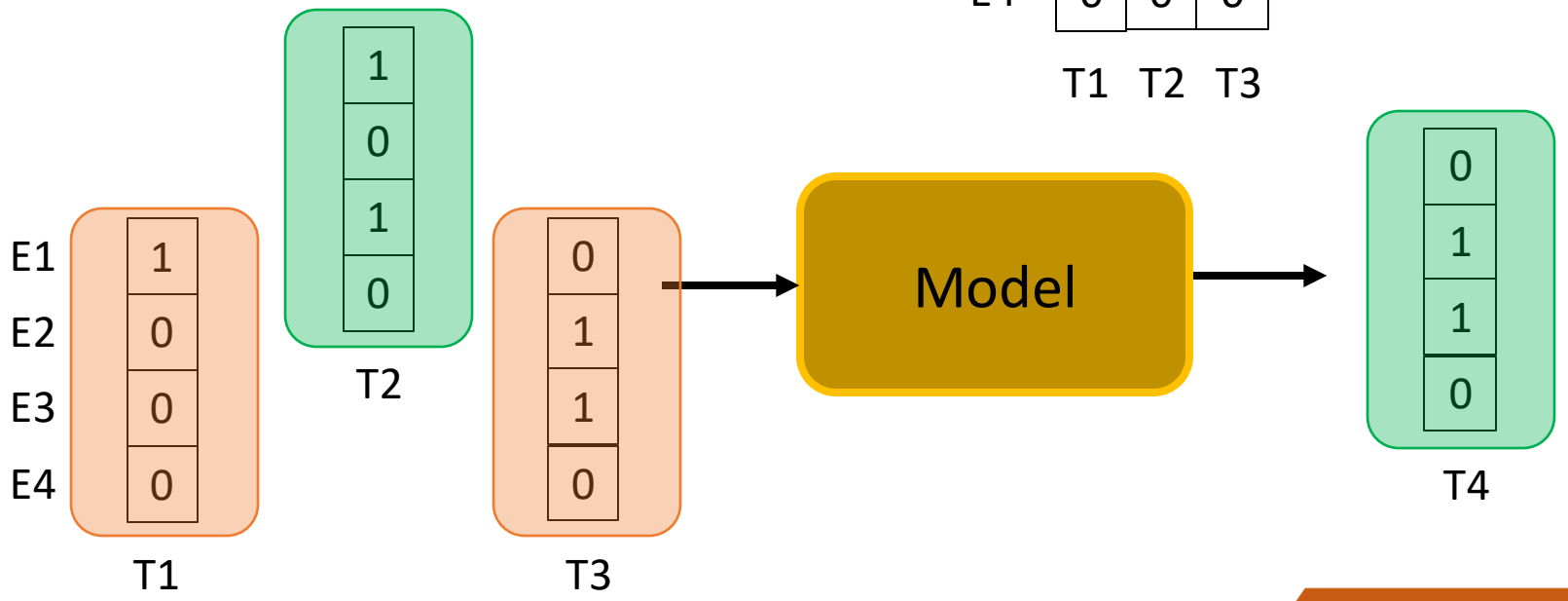
Events	Name	Events	Name
11	Smile	20	AU 12
12	AU 1	21	AU 14
13	AU 2	22	AU 15
14	AU 4	23	AU 17
15	AU 5	24	AU 20
16	AU 6	25	AU 23
17	AU 7	26	AU 25
18	AU 9	27	AU 26
19	AU 10	28	AU 45

Visual *events*

Predicting Behaviors

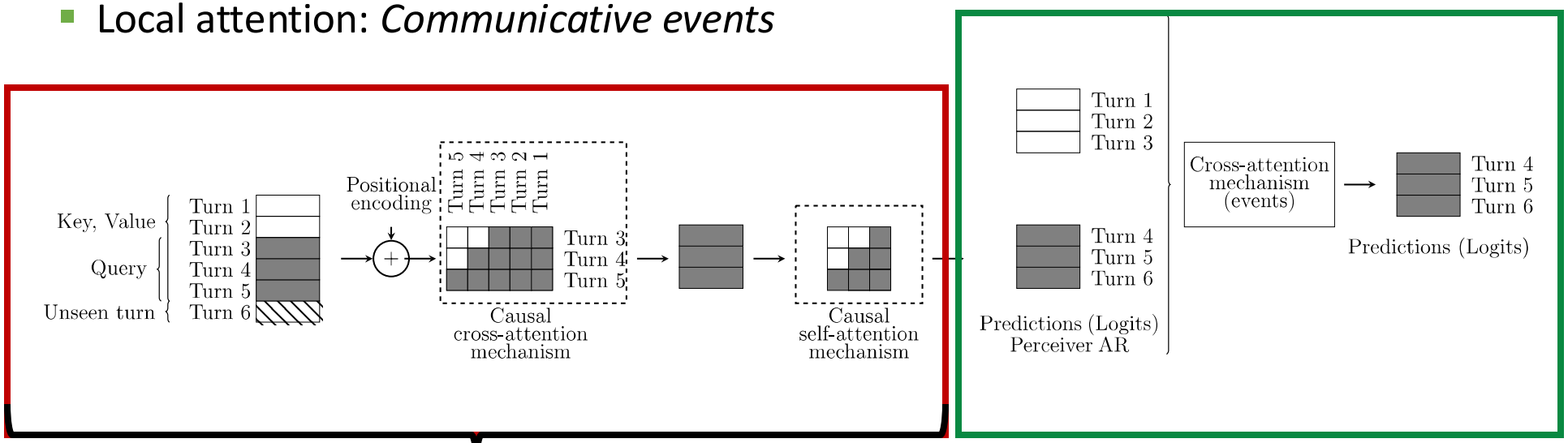
- **Can we predict the communicative events for the next turn?**
  - Interaction as a binary representation

	T1	T2	T3
E1	1	1	0
E2	0	0	1
E3	0	1	1
E4	0	0	0



Autoregressive Transformer for Modeling Interaction

- **Proposed model**
  - Global attention: Turns -> Causal attention mask
    - Based on Perceiver AR architecture
  - We propose the addition of a new block for refining the prediction of the model
    - Local attention: *Communicative events*



Follows the Perceiver AR architecture (Hawthorne et al., 2022)

Global attention

Local attention

Evaluation Results

- **Evaluation**
  - We measure the performance by using precision, recall, and f1-score
- **Findings:**
  - Our proposed model outperformed the baselines
  - The "local attention" block contributes to the model performance
  - t-test -> our proposed model results are statistically significant with p-value < 0.01

Methods	Precision	Recall	F1-Score
LSTM	0.6137±0.0179	0.4800±0.0195	0.5333±0.0091
Perceiver AR architecture	0.5436±0.0209	0.4199±0.0148	0.4705±0.0066
Proposed model	**0.6532±0.0191**	**0.5352±0.0130**	**0.5862±0.0059**

What is next?

What is Next?

- **Close the loop in interactions with social interactive agents**
  - Increase entrainment by synthesizing communication events
  - Evaluate the effect of entrainment models

- **Constrain generative models by social signals inferred from the user**

- **Better use of text using *large language models* (LLMs)**
  - Improving discourse constraints in the generated behaviors
  - Capture better contextual information

Multimodal Speech Processing (MSP)

Are you Interested in joining CMU? I am hiring at all levels

Tel: (972) 883-4351
Email: busso@cmu.edu, busso@utdallas.edu
Web: http://utdallas.edu/~busso/

ALI SALMAN
Ph.D. Student
Face Emotion Recognition

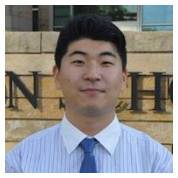
ABINAY REDDY NAINI
Ph.D. Student
Speech Emotion Recognition

KAYLA CAUGHLIN
Ph.D. Student
Oral cancer detection

LUZ MARTINEZ-LUCAS
Ph.D. Student
Affective computing

LUCAS GONÇALVES
Ph.D. Student
Multimodal emotion recognition

PRAVIN MOTE
Ph.D. Student
Multimodal emotion recognition



SEONG-GYUN LEEM
Ph.D. Student
Noisy Speech Processing

KAREN ROSERO
Ph.D. Student
Multimodal processing for Healthcare

Toward Trustworthy AI for Mental Healthcare: Exploring Socio-Demographic Bias, Privacy Risks, and Collaborative Decision-Making

Theodora Chaspari, Ph.D.
HUman Bio-Behavioral Signals (HUBBS) Lab
Institute of Cognitive Science | Computer Science | University of Colorado Boulder

Second Workshop on Social Artificial Intelligence
University of Glasgow, Glasgow, UK
September 20th, 2024

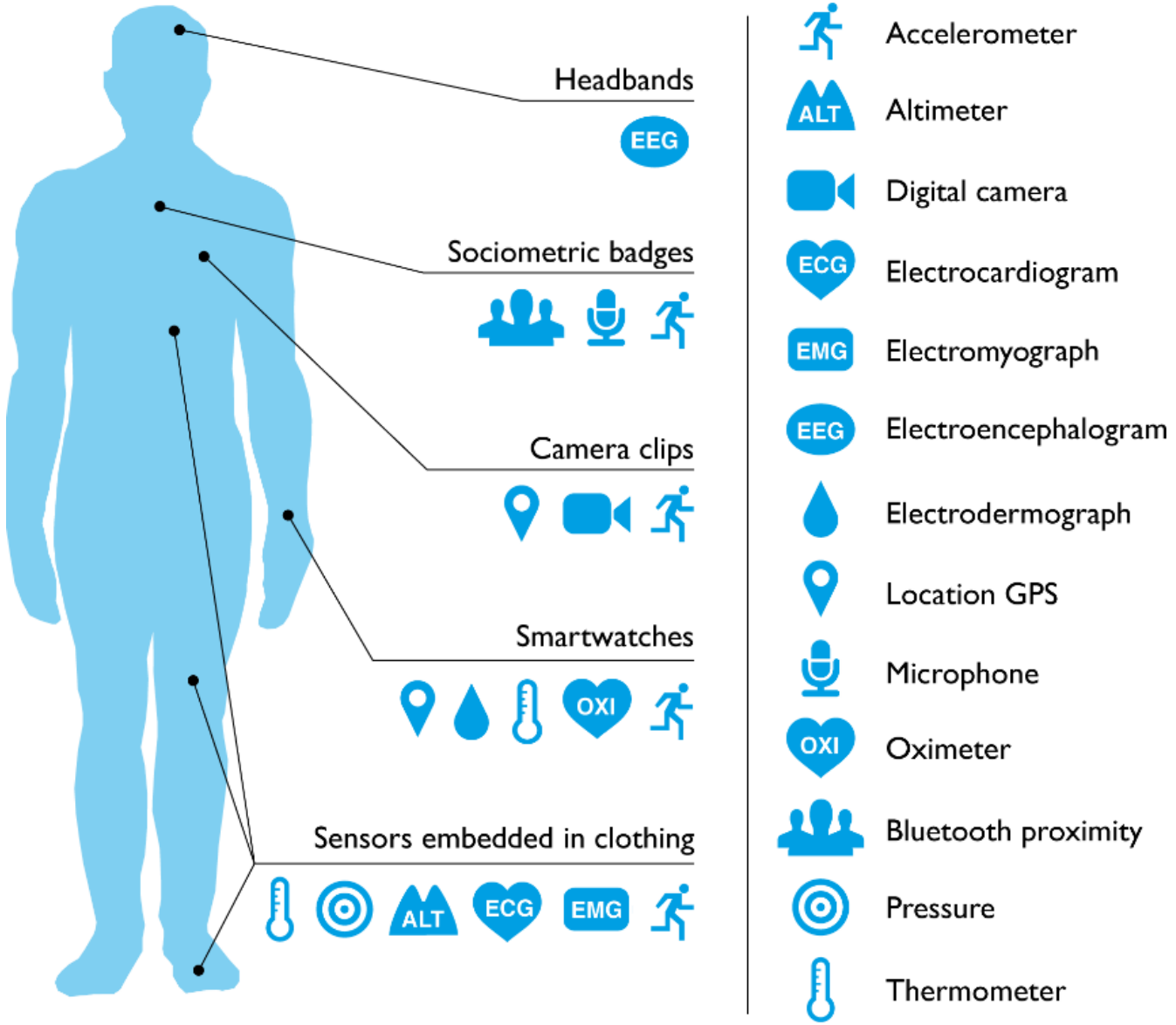
University of Colorado **Boulder**

HUBBS
HUman Bio-Behavioral Signals Lab

Ambulatory monitoring

Ecological monitoring of momentary experience via wearable devices and smartphones

Unique opportunity to investigate individual behaviors and social processes occurring in daily life

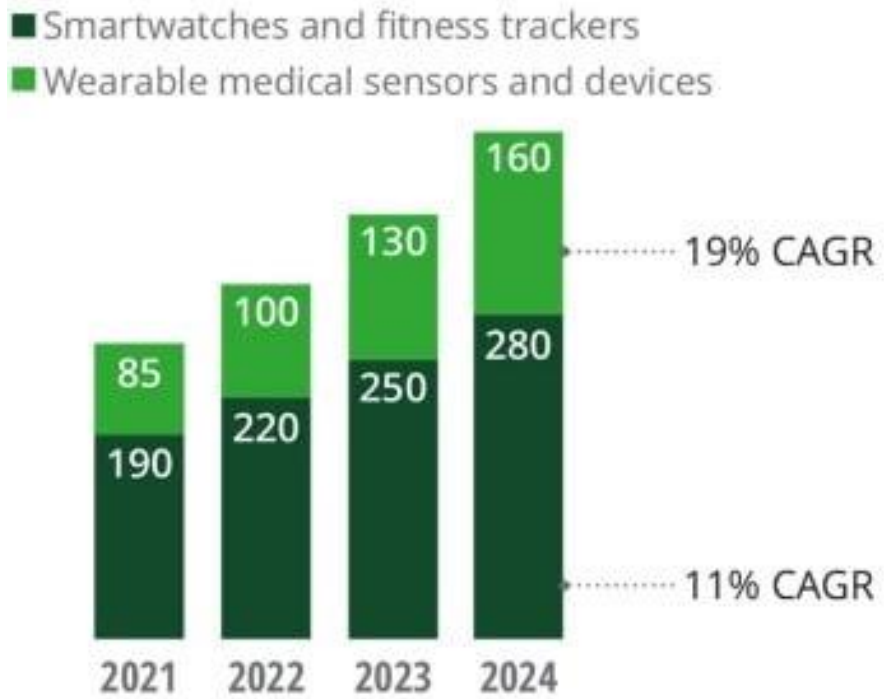


Source: https://blog.stormid.com/data-challenge-for-digital-healthcare/

FIGURE 1

The global health wearables market is already big and expanding fast

Number of units shipped globally (millions), 2021–2024



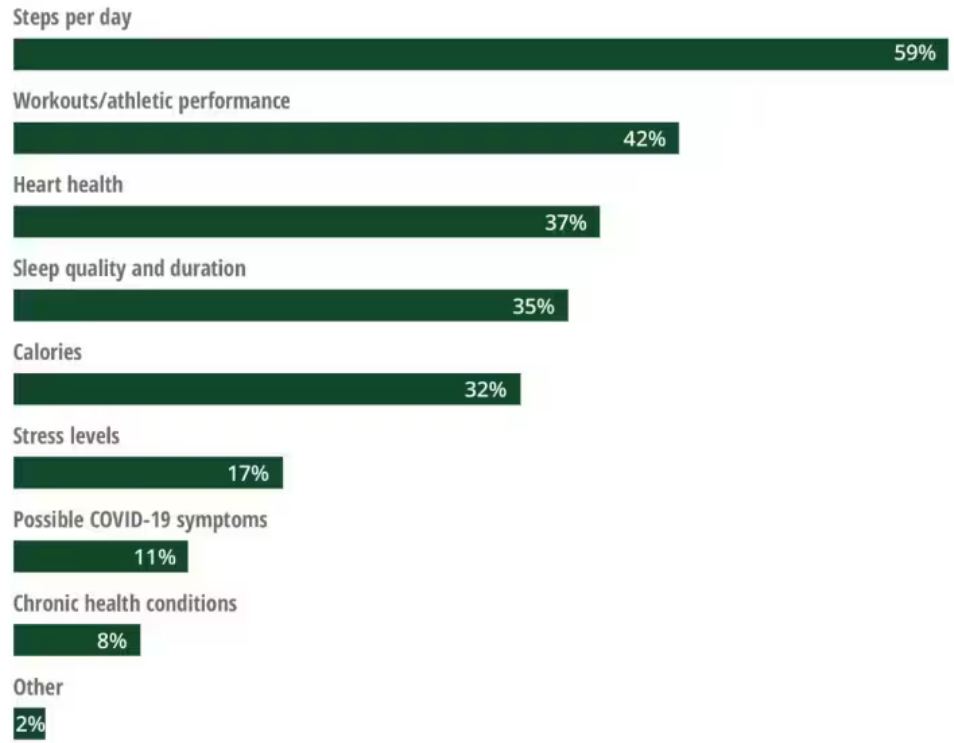
Source: Deloitte analysis of industry market sizing data.

Deloitte Insights | deloitte.com/insights

FIGURE 2

People use smartwatches to monitor heart health, sleep quality, and chronic conditions

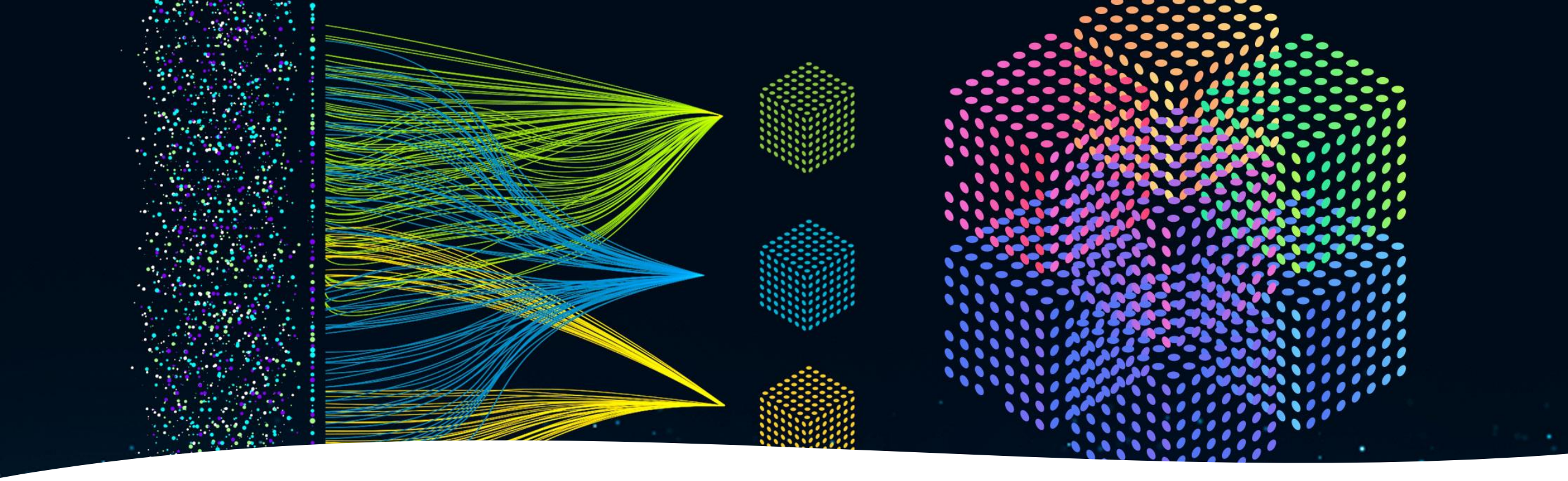
Which of the following do you use your smartwatch to measure? Select all that apply.



Notes: Respondents to this question both owned a fitness tracker or smartwatch personally *and* used these devices. The data reflects responses from US consumers to a survey conducted in June 2021.

Source: Deloitte 2021 Connectivity and Mobile Trends Survey.

Deloitte Insights | deloitte.com/insights



Human-centered machine learning

Machine learning algorithms can recognize complex patterns in data samples, identify subtle changes in social behavior and mental health, and predict early warnings allowing timely interventions in a personalized manner

Human-centered machine learning

social communication

aging-in-place

mental healthcare

human-AI teaming

team science

University of Colorado **Boulder**

HUBBS
HUman Bio-Behavioral Signals Lab

Human-centered machine learning

Computational challenges with ambulatory data

Data volume and scalability

Noise and data quality

Real-time processing and resource constraints

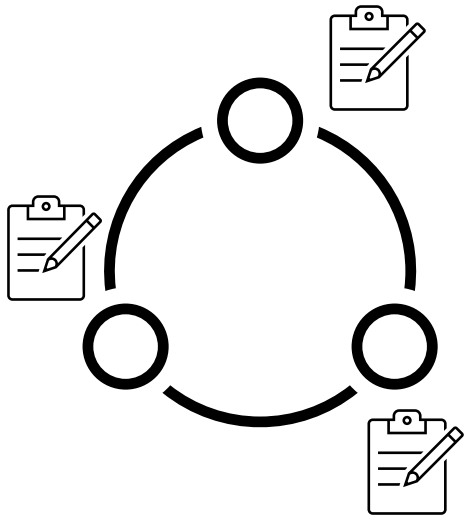
University of Colorado **Boulder**

HUBBS
HUman Bio-Behavioral Signals Lab

Human-centered machine learning

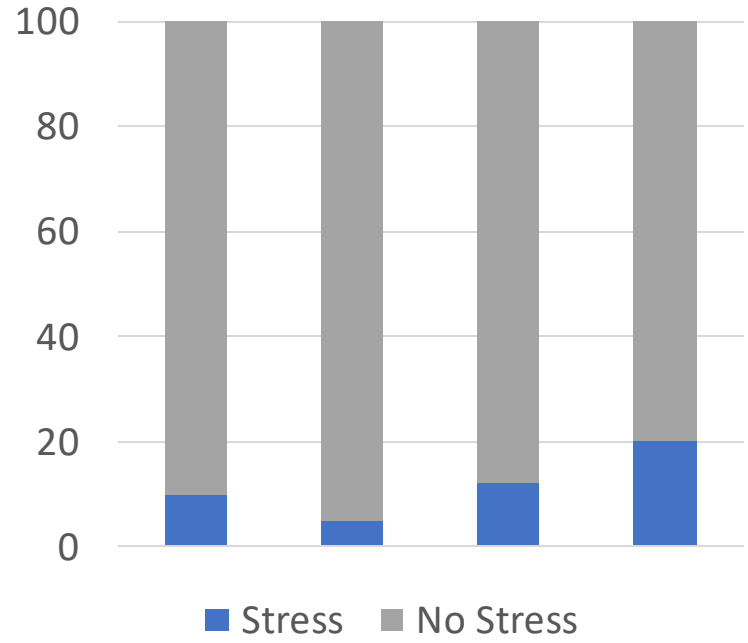
Computational challenges with ambulatory data

Temporal sparsity in labels



Inter-individual variability

Variable base rate of focal behaviors



Lee, C. C., Chaspari, T., Provost, E. M., & Narayanan, S. S. (2023). An Engineering View on Emotions and Speech: From Analysis and Predictive Models to Responsible Human-Centered Applications. Proceedings of the IEEE.

Automatically detecting behavioral states using ambulatory data

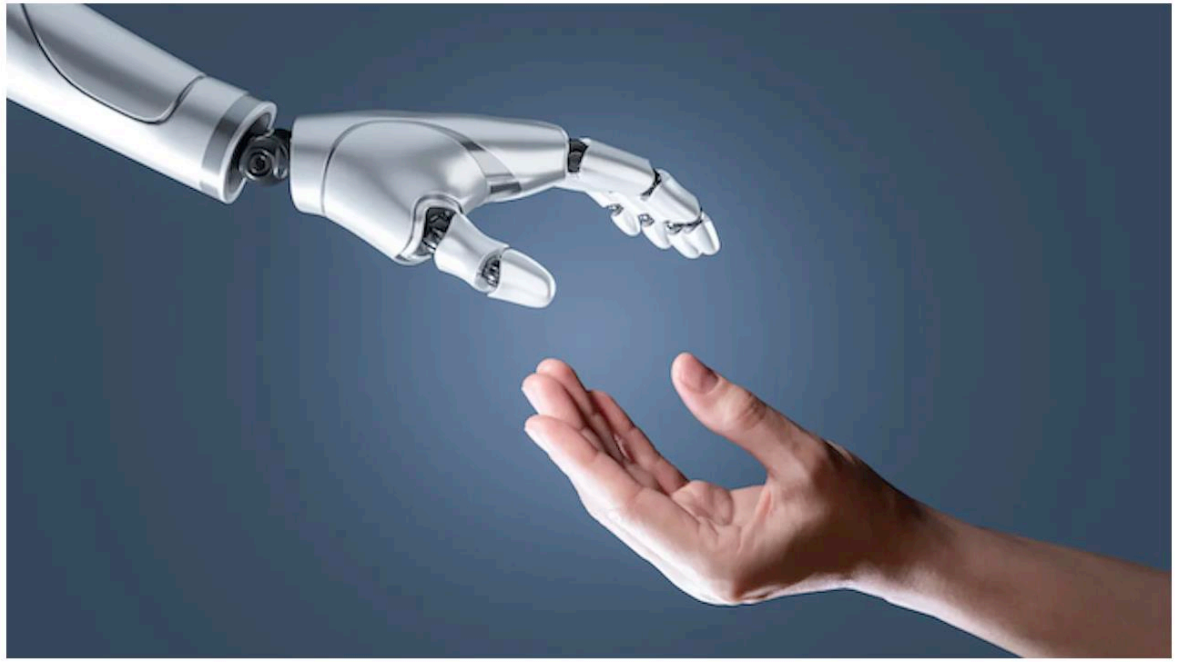
Proof-of-concept	Detecting interpersonal conflict based on ambulatory data via machine learning (Timmons et al., IEEE Computer 2017)
Group-specific machine learning models	Stratifying participants in theoretically-driven groups, learning multimodal conflict patterns separately per group (Gupta et al., ACM TIOT 2020)
Personalized machine learning models	Learning relative differences in multimodal data between stress and no-stress conditions for each participant (Paromita et al., Frontiers Health 2023)
Generalizable learning	Leveraging labelled data from in-lab conditions to effectively generalize ML models in real-life conditions (Feng & Chaspari, IEEE TAFFC 2021)

REPORT | JUNE 21, 2023

As AI Spreads, Experts Predict the Best and Worst Changes in Digital Life by 2035

They have deep concerns about people's and society's overall well-being. But they also expect great benefits in health care, scientific advances and education

BY JANNA ANDERSON AND LEE RAINIE



Anderson, J., & Rainie, L. (2023). As AI Spreads, Experts Predict the Best and Worst Changes in Digital Life by 2035.

JAMA Internal Medicine | Original Investigation

Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum

John W. Ayers, PhD, MA; Adam Poliak, PhD; Mark Dredze, PhD; Eric C. Leas, PhD, MPH; Zechariah Zhu, BS; Jessica B. Kelley, MSN; Dennis J. Faix, MD; Aaron M. Goodman, MD; Christopher A. Longhurst, MD, MS; Michael Hogarth, MD; Davey M. Smith, MD, MAS

IMPORTANCE The rapid expansion of virtual health care has caused a surge in patient messages concomitant with more work and burnout among health care professionals. Artificial intelligence (AI) assistants could potentially aid in creating answers to patient questions by drafting responses that could be reviewed by clinicians.

OBJECTIVE To evaluate the ability of an AI chatbot assistant (ChatGPT), released in November 2022, to provide quality and empathetic responses to patient questions.

DESIGN, SETTING, AND PARTICIPANTS In this cross-sectional study, a public and nonidentifiable database of questions from a public social media forum (Reddit's r/AskDocs) was used to randomly draw 195 exchanges from October 2022 where a verified physician responded to a public question. Chatbot responses were generated by entering the original question into a fresh session (without prior questions having been asked in the session) on December 22 and 23, 2022. The original question along with anonymized and randomly ordered physician and chatbot responses were evaluated in triplicate by a team of licensed health care professionals. Evaluators chose "which response was better" and judged both "the quality of information provided" (*very poor*, *poor*, *acceptable*, *good*, or *very good*) and "the empathy or bedside manner provided" (*not empathetic*, *slightly empathetic*, *moderately empathetic*, *empathetic*, and *very empathetic*). Mean outcomes were ordered on a 1 to 5 scale and compared between chatbot and physicians.

Invited Commentary page 596

Related article page 507

Supplemental content

Emergence of intelligent conversational agents and their future in healthcare

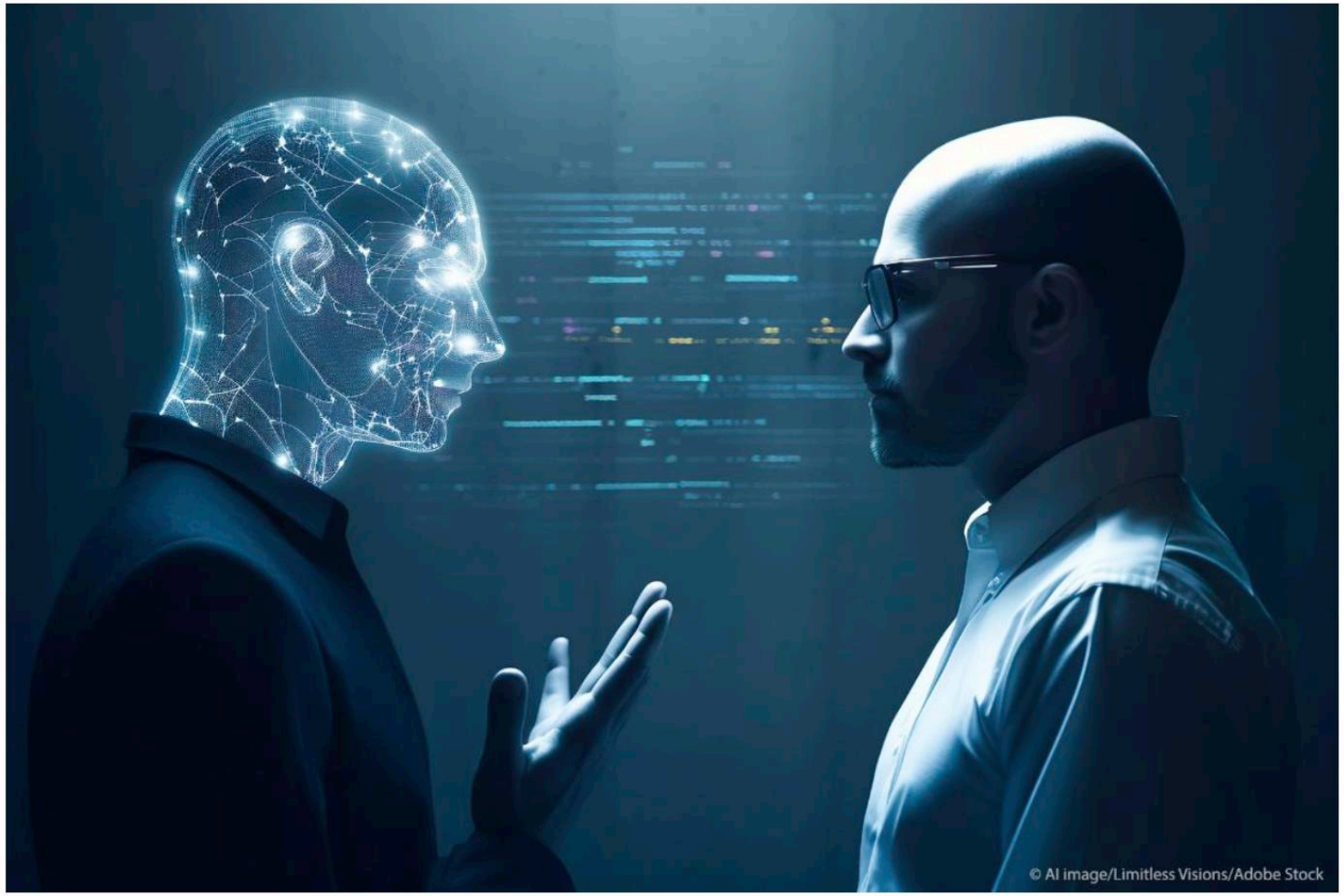
Chatbot responses were rated of significantly higher quality and more empathetic compared to physician responses.

Ayers, J. W., Poliak, A., Dredze, M., Leas, E. C., Zhu, Z., Kelley, J. B., Faix, D. J., Goodman, A. M., Longhurst C. A., & Smith, D. M. (2023). Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA internal medicine.

EU AI Act: first regulation on artificial intelligence

Society Updated: 14-06-2023 - 14:06
Created: 08-06-2023 - 11:40

The use of artificial intelligence in the EU will be regulated by the AI Act, the world's first comprehensive AI law. Find out how it will protect you.



This illustration of artificial intelligence has in fact been generated by AI

EU AI Act

A comprehensive legal framework for AI seeking to ensure that AI systems used in the European Union are safe, transparent, traceable, non-discriminatory and environmentally friendly.

Source:
https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

Blueprint for an AI Bill of Rights

MAKING AUTOMATED SYSTEMS WORK FOR THE AMERICAN PEOPLE

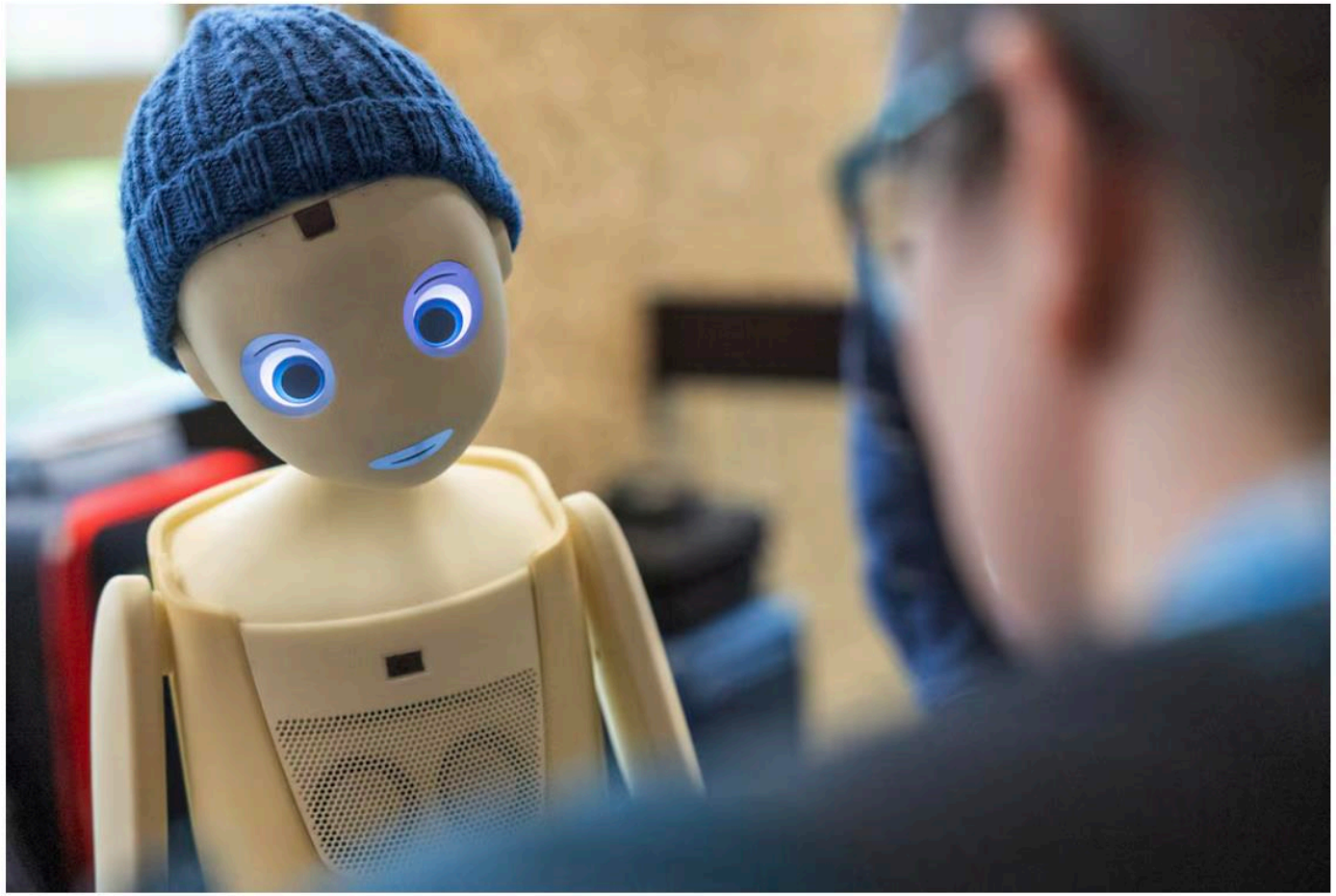
OSTP

UN advisory body makes seven recommendations for governing AI

Reuters

By **Supantha Mukherjee**

September 19, 2024 6:44 AM GMT+1 · Updated 6 hours ago



A visitor talks with the Navel robot, by Navel Robotics, during the AI for Good Global summit on artificial intelligence, organised by the International Telecommunication Union (ITU), in Geneva, Switzerland, May 30, 2024. REUTERS/Denis Balibouse/File photo Purchase Licensing Rights

- Global governance
- Global cooperation on establishing common understanding for pursuing common benefits
- Global AI fund to address gaps in capacity and collaboration
- Global AI data framework

United Nations. (2024). Governing AI for Humanity. Report

University of Colorado **Boulder**

HUBBS
HUman Bio-Behavioral Signals Lab

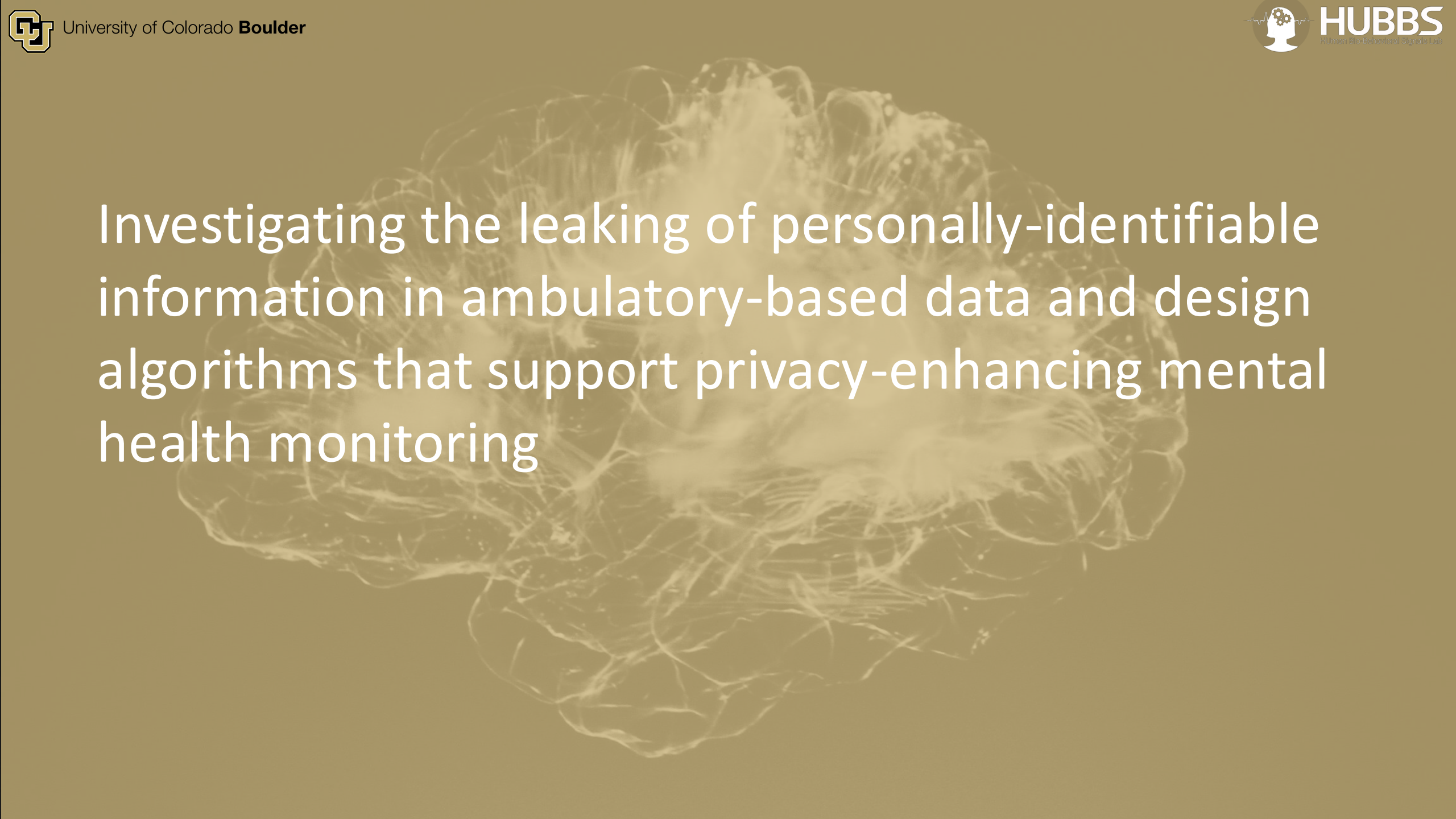
Trustworthy human-centered machine intelligence for augmenting healthcare

To address fundamental algorithmic challenges for achieving responsible human-centered AI and trustworthy human-AI partnership

Investigating the leaking of personally-identifiable information in ambulatory-based data and design algorithms that support privacy-enhancing mental health monitoring

Deconstructing socio-demographic bias and investigating methods for achieving equitable AI applications for mental health

Integrating domain expertise into the design of explainable AI for supporting clinician decision-making and training



Investigating the leaking of personally-identifiable information in ambulatory-based data and design algorithms that support privacy-enhancing mental health monitoring

University of Colorado **Boulder**

HUBBS
HUman Bio-Behavioral Signals Lab

Privacy-enhancing well-being and mental health monitoring

Research objectives

To investigate the presence of personally-identifiable information (PII), such as user identity, biological sex, and income

To examine the feasibility of feature selection and feature transformation algorithms to effectively suppress PII and preserve mental health information

Privacy-enhancing well-being monitoring

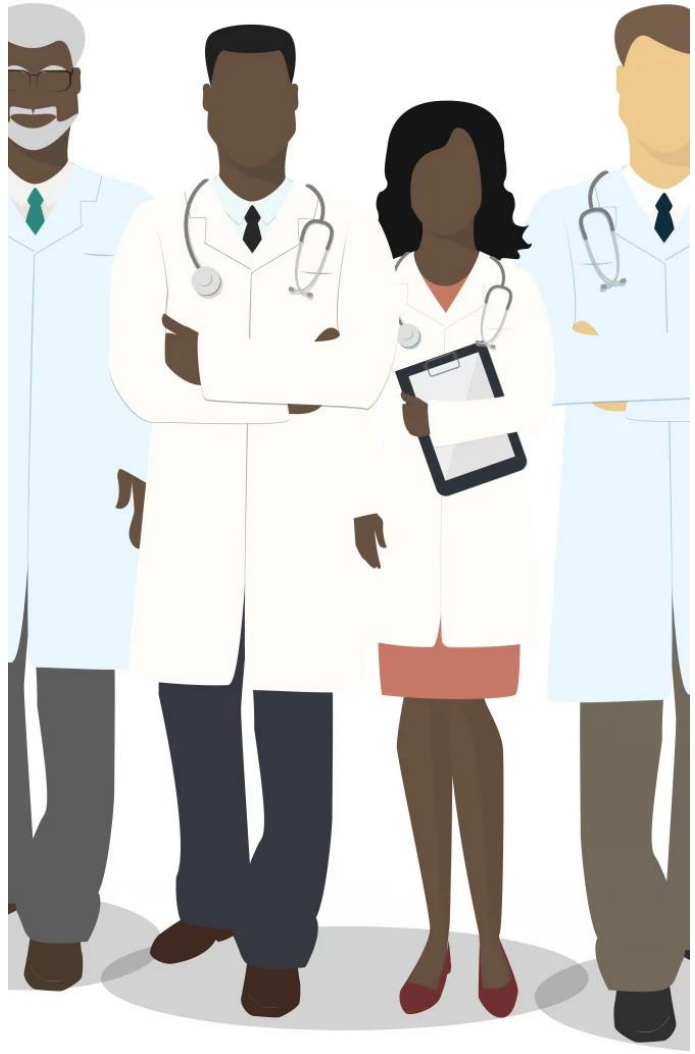
TILES-2018 Dataset (Mundnich et al., 2020)

212 Participants (21-65 years)

10 weeks

Wearable devices (Fitbit, Omsignal garment) and smartphones

Self-reports of anxiety and affect



Mundnich, K., Booth, B. M., l'Hommedieu, M., Feng, T., Girault, B., L'hommedieu, J., Wildman, M., Skaaden, S., Nadarajan, A., Villatte, J. L., Falk, T. H., Lerman, K., Ferrara, E., & Narayanan, S. (2020). TILES-2018, a longitudinal physiologic and behavioral data set of hospital workers. Scientific Data, 7(1), 354.

University of Colorado **Boulder**

HUBBS HUman Bio-Behavioral Signals Lab

Privacy-enhancing well-being monitoring

Task 1: Characterizing user re-identification risk

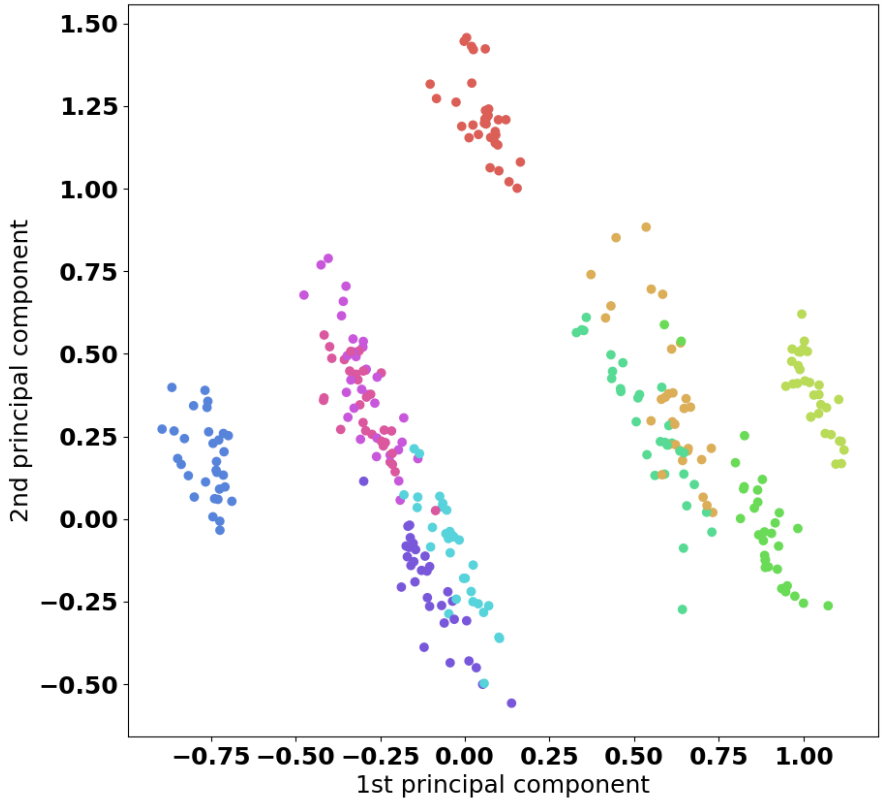
Methods

- 69 sensor features (25 daily activity, 29 acoustic, 15 physiology)
- User re-identification risk approximated as the user classification accuracy of a logistic regression model
- Large accuracy → High user re-identification risk
- Low accuracy → Low user re-identification risk

Results

- The original features bear considerable user re-identification risk: 81%-93% accuracy in the 212-class task
- Visually differentiable clusters of users based on the original feature

The first two principal components of features color-coded for 10 users



Pranjal, R., Seshadri, R., Kadaba, R. K. S. K., Feng, T., Narayanan, S. S., & Chaspari, T. (2023, June). Toward Privacy-Enhancing Ambulatory-Based Well-Being Monitoring: Investigating User Re-Identification Risk in Multimodal Data. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE. DOI: 10.1109/ICASSP49357.2023.10096235.

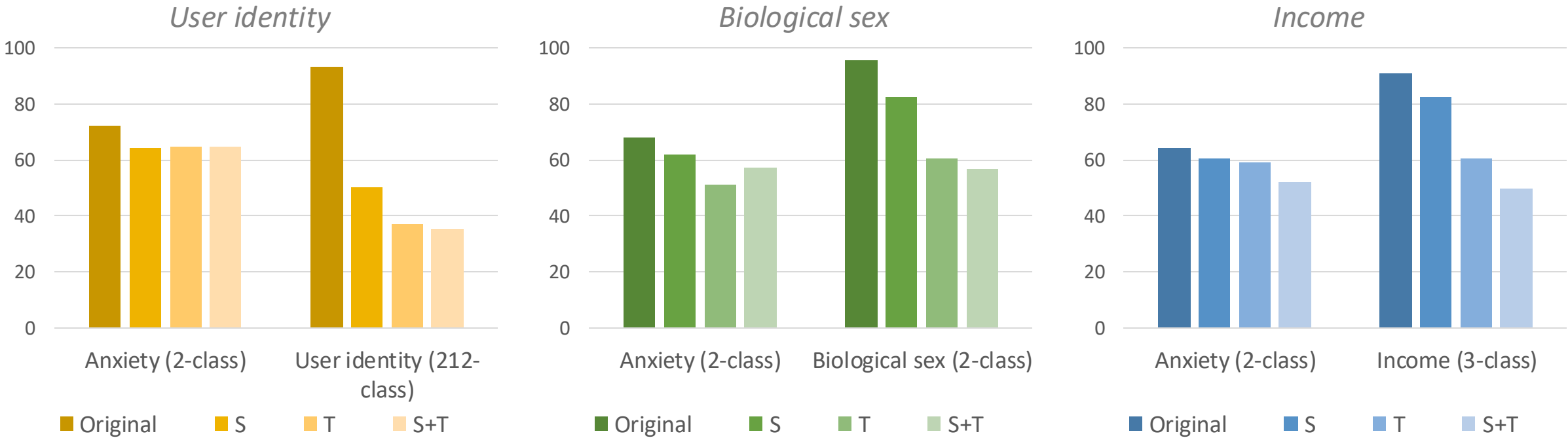
University of Colorado **Boulder**

HUBBS HUman Bio-Behavioral Signals Lab

Privacy-enhancing well-being monitoring

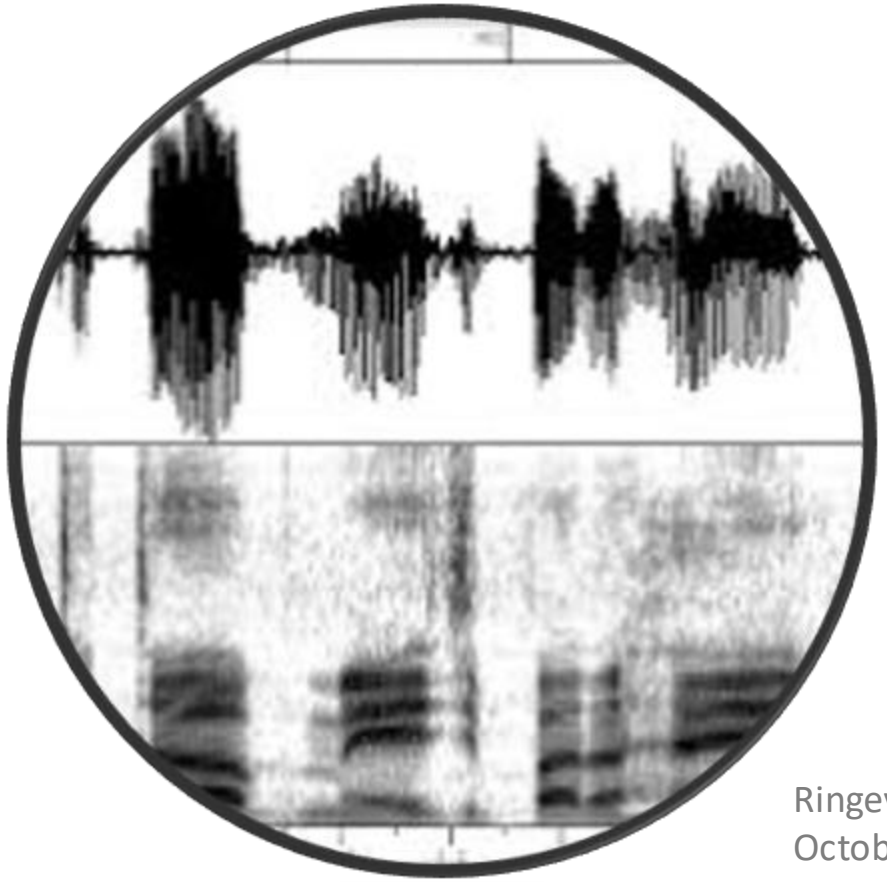
Task 2: Conducting privacy-preserving well-being estimation

- **Personally-identifiable information (PII):** user identity, biological sex, income
- **Well-being outcome:** Self-reported anxiety
- **Feature selection (S):** Selection of features that are the least discriminative of the target PII
- **Feature transformation (T):** Adversarial autoencoder preserving anxiety information and reducing PII



Pranjal, R., Seshadri, R., Kadaba, R. K. S. K., Feng, T., Narayanan, S. S., & Chaspari, T. (2023, June). Toward Privacy-Enhancing Ambulatory-Based Well-Being Monitoring: Investigating User Re-Identification Risk in Multimodal Data. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE. DOI: 10.1109/ICASSP49357.2023.10096235.

Speech as a biomarker for mental health



- Psychomotor symptoms associated with mental health can be reflected in **prosody**, **spectrotemporal** characteristics, and **vocal fold excitation**
- Automated assessment systems, **combining speech measures with machine learning (ML) algorithms**, can potentially augment depression diagnosis and screening via offering **data-driven insights** that can complement existing investigative techniques

Ringeval, F., Schuller, B., Valstar, M., Gratch, J., Cowie, R., Scherer, S., & Pantic, M. (2017, October). Avec 2017: Real-life depression and affect recognition workshop and challenge. In Proceedings of the 7th annual workshop on audio/visual emotion challenge (pp. 3-9).

Chen, W., Xing, X., Xu, X., Pang, J., & Du, L. (2022). Speechformer: A hierarchical efficient framework incorporating the characteristics of speech. In Interspeech. ISCA.

Feng, K., & Chaspari, T. (2023). A knowledge-driven vowel-based approach of depression classification from speech using data augmentation. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1-5). IEEE.

Behavior-aware speech anonymization

- Privacy is an impeding factor to trust and acceptance of these technologies
- Privacy-preserving algorithms for mental health would need to retain mental health and linguistic information as while anonymizing speech

Research objectives

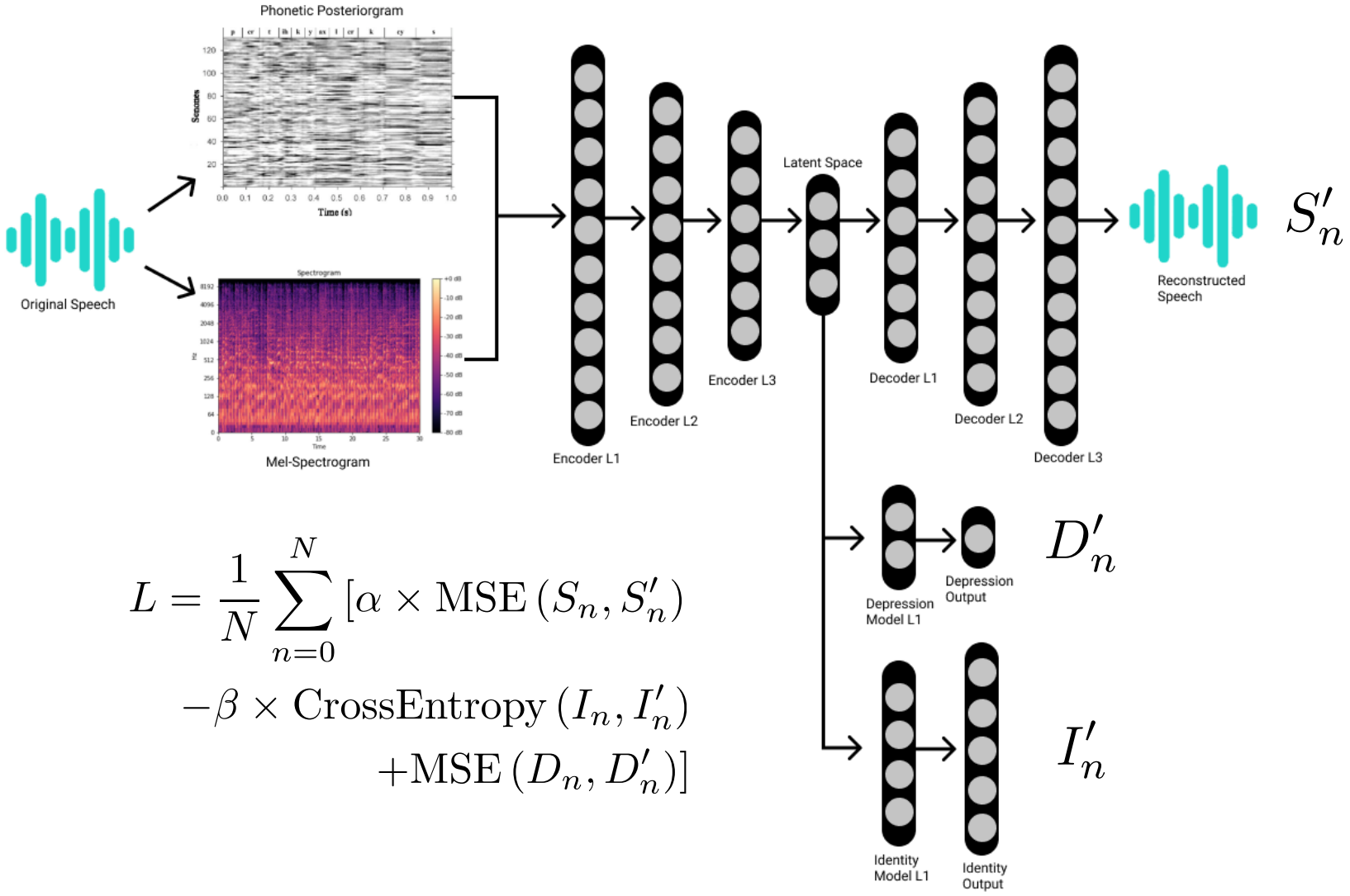
- To investigate the feasibility of an anonymization algorithm to effectively suppress information related to the speaker ID and preserve mental health information.
- To examine an auto-encoder architecture that can generate anonymized speech with additional identity and mental health constraints.

University of Colorado **Boulder**

HUBBS HUman Bio-Behavioral Signals Lab

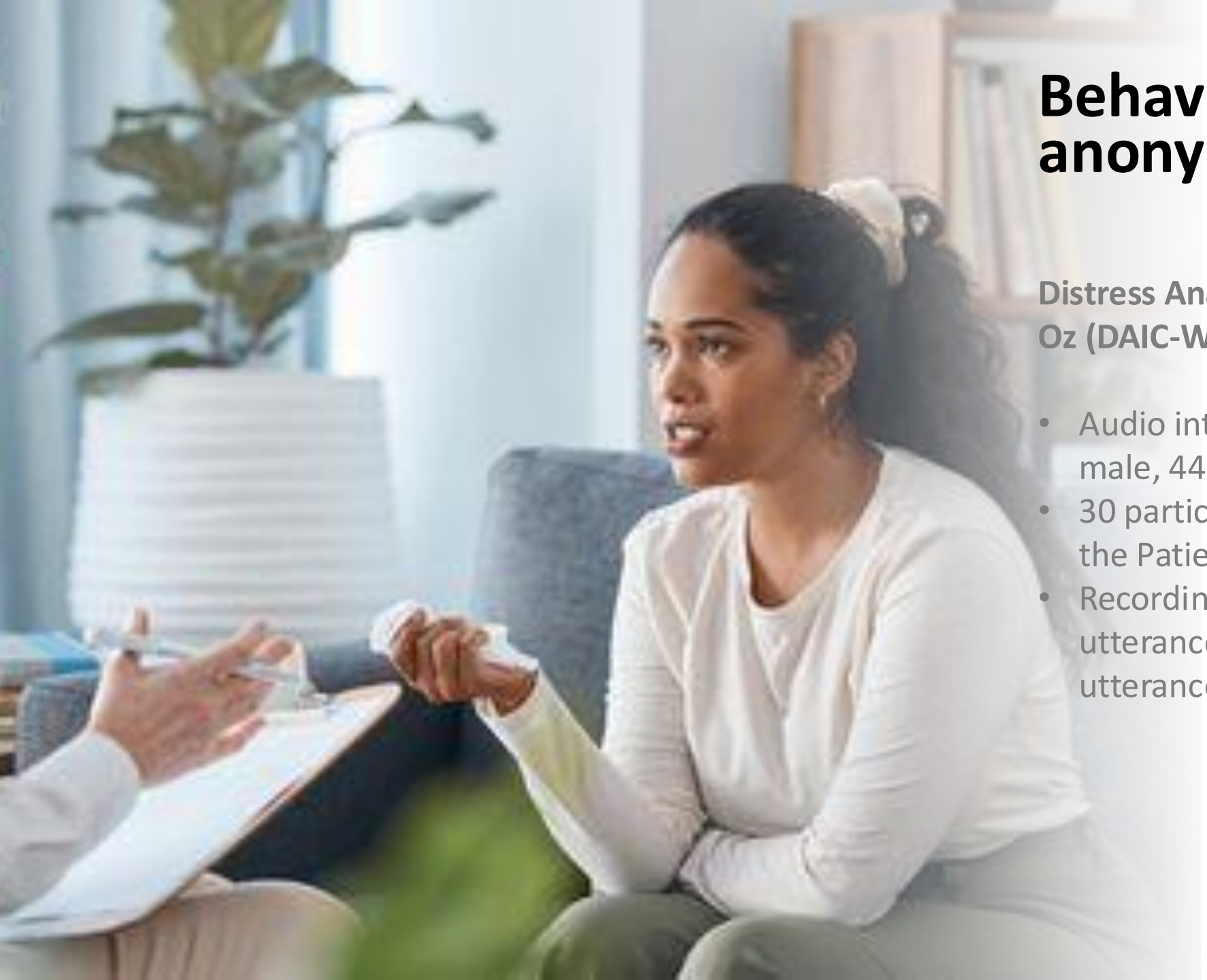
Behavior-aware speech anonymization

- Adversarial learning that minimizes loss of mental health information and maximizes user ID classification loss
- Preserving phonetic information via minimizing reconstruction error of phonetic posteriogram
- Preserving acoustic information via minimizing reconstruction error of speech spectrogram



$$L = \frac{1}{N}\sum_{n=0}^{N}[\alpha \times \mathrm{MSE}\,(S_n, S'_n) - \beta \times \mathrm{CrossEntropy}\,(I_n, I'_n) + \mathrm{MSE}\,(D_n, D'_n)]$$

Ravuri, V., Gutierrez-Osuna, R., & Chaspari, T. (2022, October). Preserving Mental Health Information in Speech Anonymization. In 2022 10th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW) (pp. 1-8). IEEE.



Behavior-aware speech anonymization

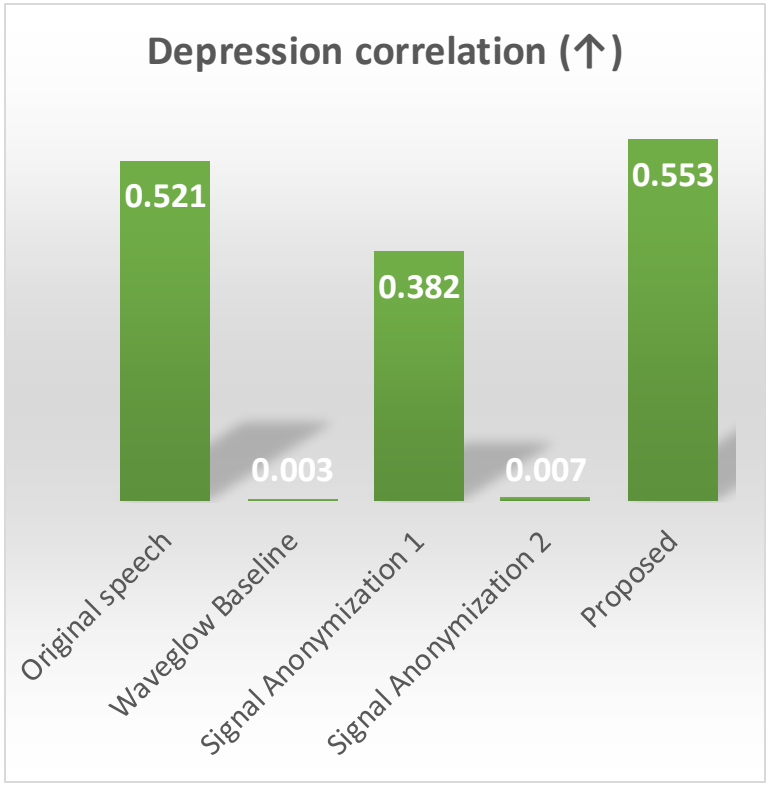
Distress Analysis Interview Corpus Wizard of Oz (DAIC-WoZ) dataset (Gratch et al., 2014)

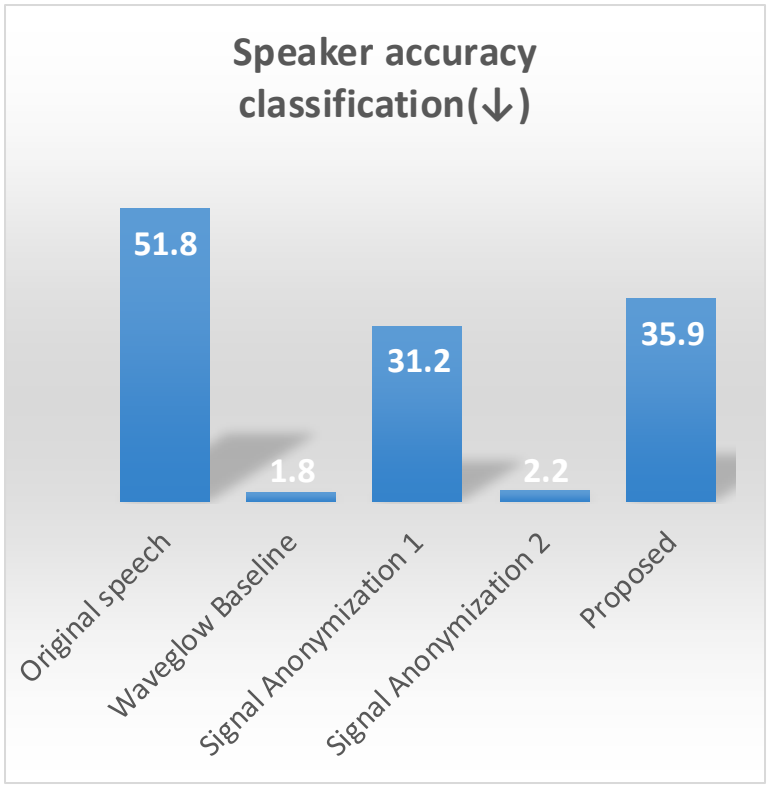
- Audio interviews of 107 participants (63 male, 44 female)
- 30 participants with depression based on the Patient Health Questionnaire (PHQ-8)
- Recordings were segmented into individual utterances (1s - 30s duration) → 11,993 utterances

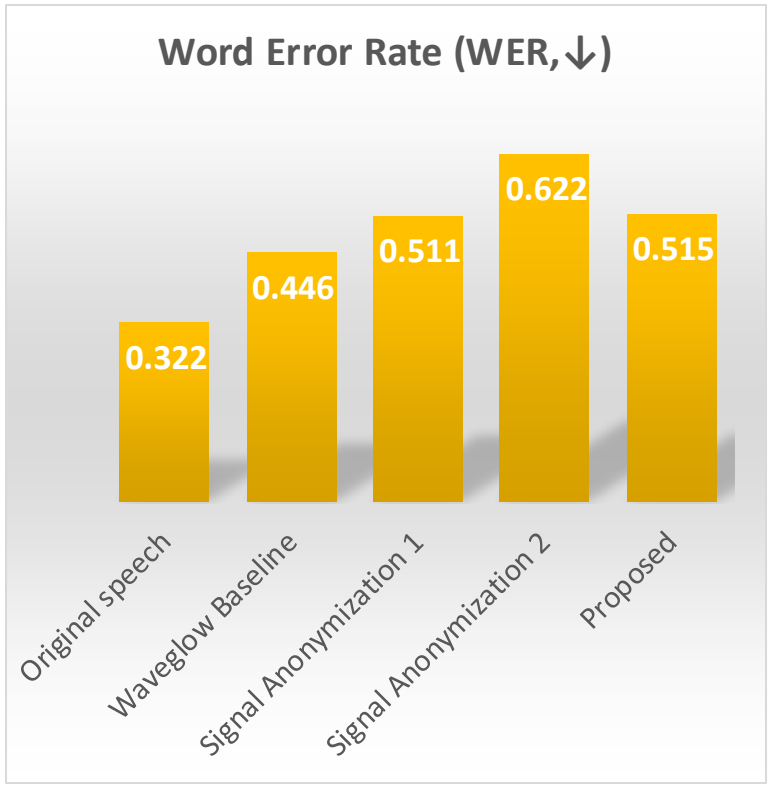
University of Colorado **Boulder**

HUBBS
HUman Bio-Behavioral Signals Lab

Behavior-aware speech anonymization







The proposed speech anonymization method preserves the depression detection ability of the original speech while resulting in lower speaker classification accuracy.

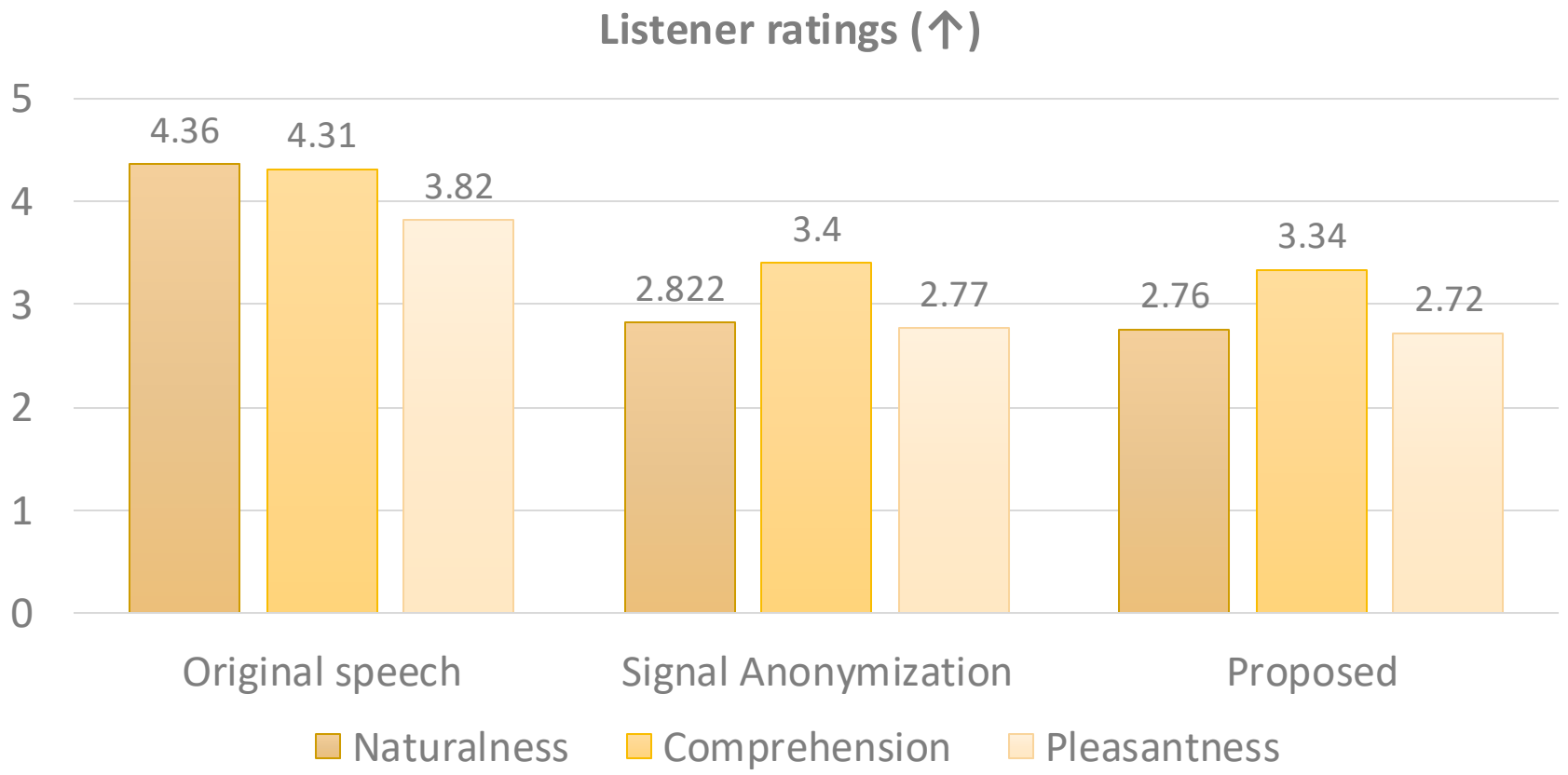
Waveglow: Original spectrogram serving as input to Waveglow speech synthesizer; **Speech Anonymization 1:** Auto-encoder that contains only the speaker classification module, but not the depression estimation; **Speech Anonymization 2:** Cascade of signal processing methods that suppresses speaker information while maintaining speech intelligibility

Ravuri, V., Gutierrez-Osuna, R., & Chaspari, T. (2022, October). Preserving Mental Health Information in Speech Anonymization. In 2022 10th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW) (pp. 1-8). IEEE.

University of Colorado **Boulder**

HUBBS
HUman Bio-Behavioral Signals Lab

Behavior-aware speech anonymization



The proposed speech anonymization method depicts reduced ratings in terms of perceived speech naturalness, comprehension, and pleasantness.

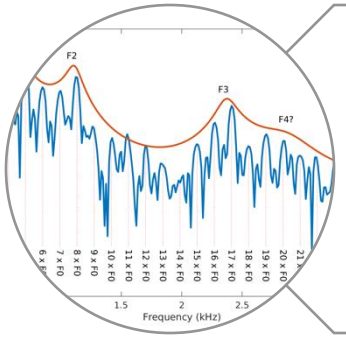
Ravuri, V., Gutierrez-Osuna, R., & Chaspari, T. (2022, October). Preserving Mental Health Information in Speech Anonymization. In 2022 10th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW) (pp. 1-8). IEEE.

University of Colorado Boulder

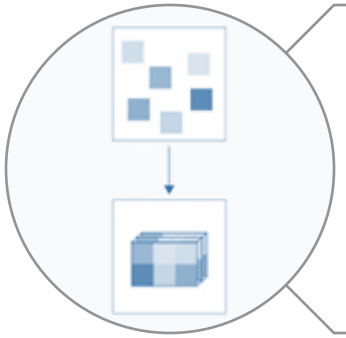
Investigating the leaking of personally identifiable information

HUBBS
HUman Bio-Behavioral Signals Lab

Summary of findings



Multimodal, including acoustic and physiological, features are indicative of socio-demographic information, posing the risk of personally identifiable information (PII) leakage

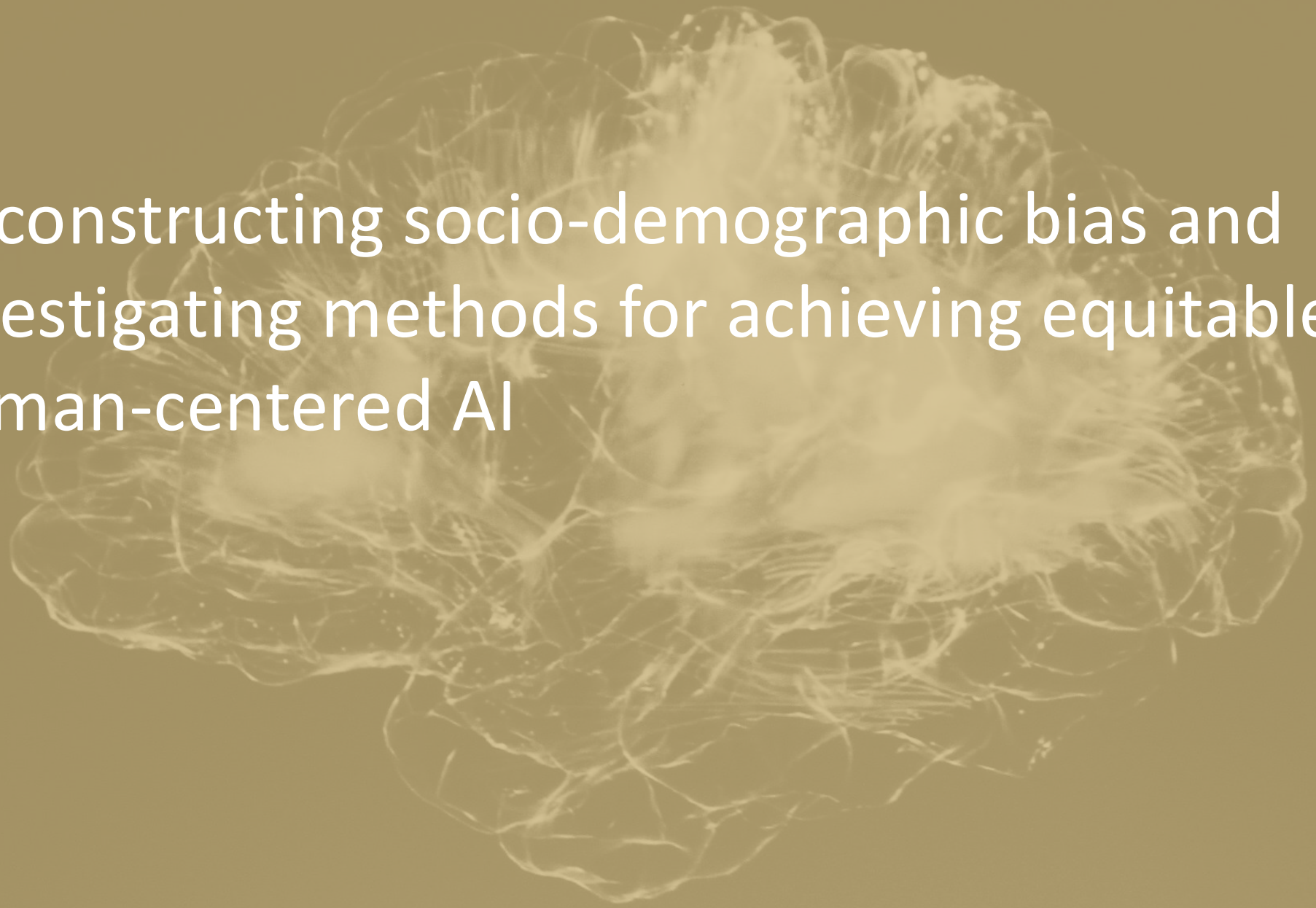


Feature transformation via adversarial learning and preserving features relevant to the focal mental health outcome only, improves ML performance, and can result in some cases in reduced PII leakage

Additional research on privacy enhancing techniques is necessary to effectively decouple inherent associations between PII and mental health information, especially in speech

University of Colorado **Boulder**

HUBBS
Human Bio-Behavioral Signals Lab



Deconstructing socio-demographic bias and investigating methods for achieving equitable human-centered AI

A Call to Action on Assessing and Mitigating Bias in Artificial Intelligence Applications for Mental Health

Perspectives on Psychological Science
1–35

DOI: 10.1177/17456916221134490
www.psychologicalscience.org/PPS
SAGE

Adela C. Timmons[1,2], Jacqueline B. Duong[1], Natalia Simo Fiallo[3], Theodore Lee[3], Huong Phuc Quynh Vo[4], Matthew W. Ahle[2], Jonathan S. Comer[3], LaPrincess C. Brewer[5,6], Stacy L. Frazier[3], and Theodora Chaspari[4]

[1]Department of Psychology, University of Texas at Austin Institute for Mental Health Research; [2]Colliga Apps Corporation, Austin, Texas; [3]Department of Psychology, Florida International University; [4]Department of Computer Science & Engineering, Texas A&M University; [5]Department of Cardiovascular Medicine, Mayo Clinic College of Medicine; and [6]Center for Health Equity and Community Engagement Research, Mayo Clinic, Rochester, Minnesota

Abstract

Advances in computer science and data-analytic methods are driving a new era in mental health research and application. Artificial intelligence (AI) technologies hold the potential to enhance the assessment, diagnosis, and treatment of people experiencing mental health problems and to increase the reach and impact of mental health care. However, AI applications will not mitigate mental health disparities if they are built from historical data that reflect underlying social biases and inequities. AI models biased against sensitive classes could reinforce and even perpetuate existing inequities if these models create legacies that differentially impact who is diagnosed and treated, and how effectively. The current article reviews the health-equity implications of applying AI to mental health problems, outlines state-of-the-art methods for assessing and mitigating algorithmic bias, and presents a call to action to guide the development of *fair-aware AI* in psychological science.

Keywords

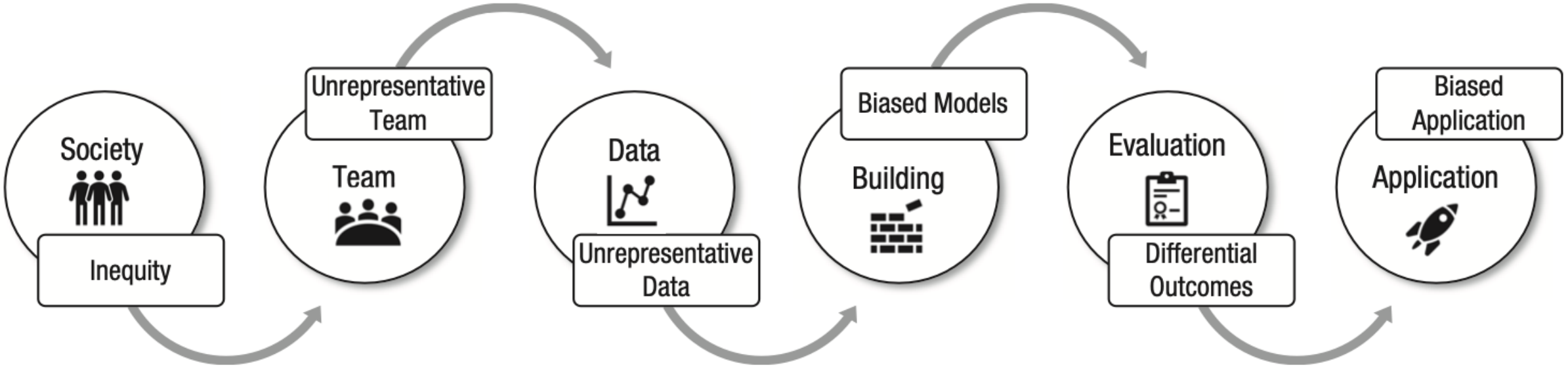
artificial intelligence, fair aware, bias, mental health equity

University of Colorado **Boulder**

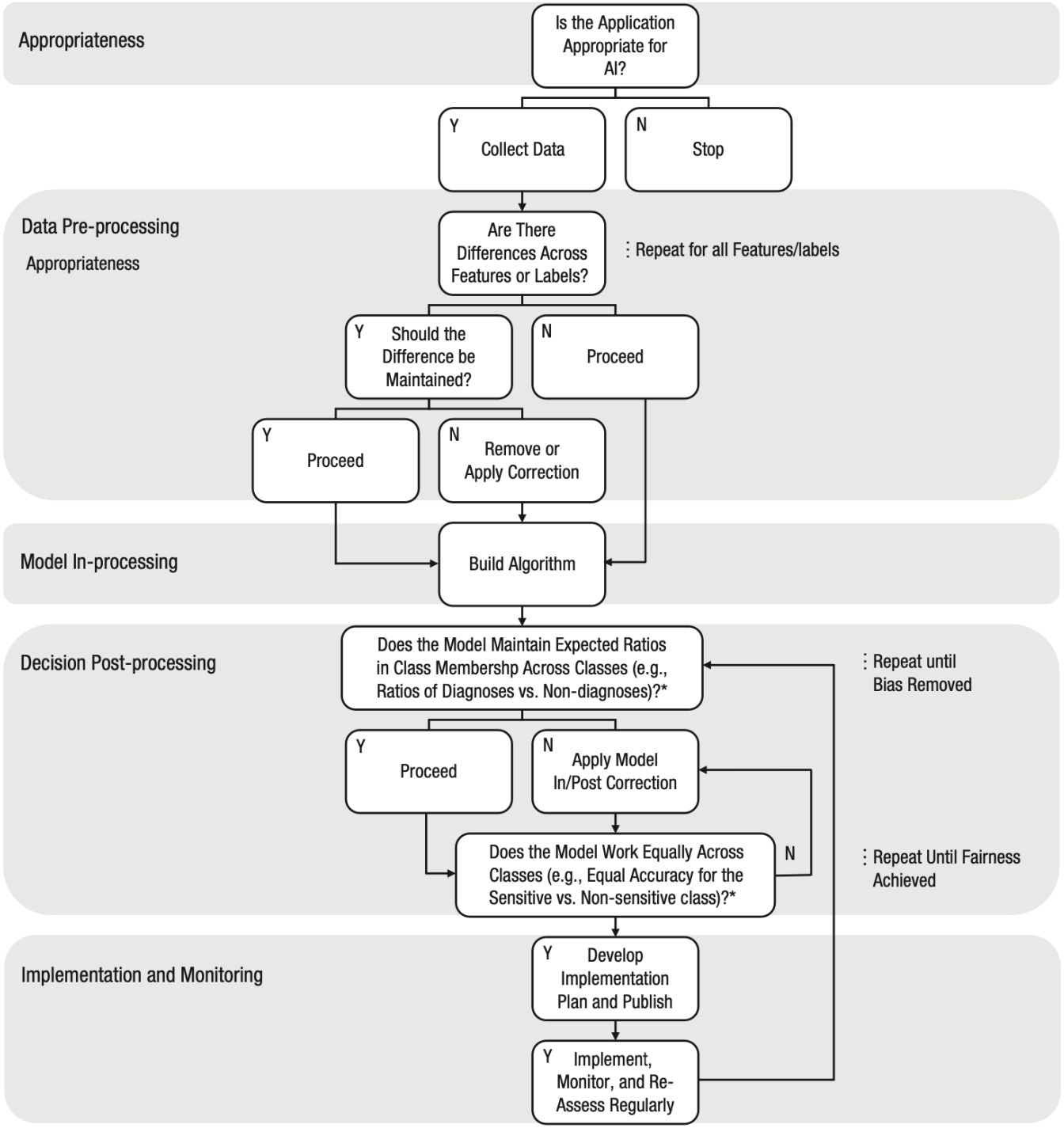
The University of Texas at Austin
Department of Psychology
College of Liberal Arts

Assessing and Mitigating Bias in Artificial Intelligence Applications for Mental Health

Overview of the model-building process and points at which bias can be introduced to models



Timmons, A. C., Duong, J. B., Simo Fiallo, N., Lee, T., Vo, H. P. Q., Ahle, M. W., Comer, J. S., Brewer, L. C., Frazier, S. L., & Chaspari, T. (2023). A call to action on assessing and mitigating bias in artificial intelligence applications for mental health. Perspectives on Psychological Science, 18(5), 1062-1096.



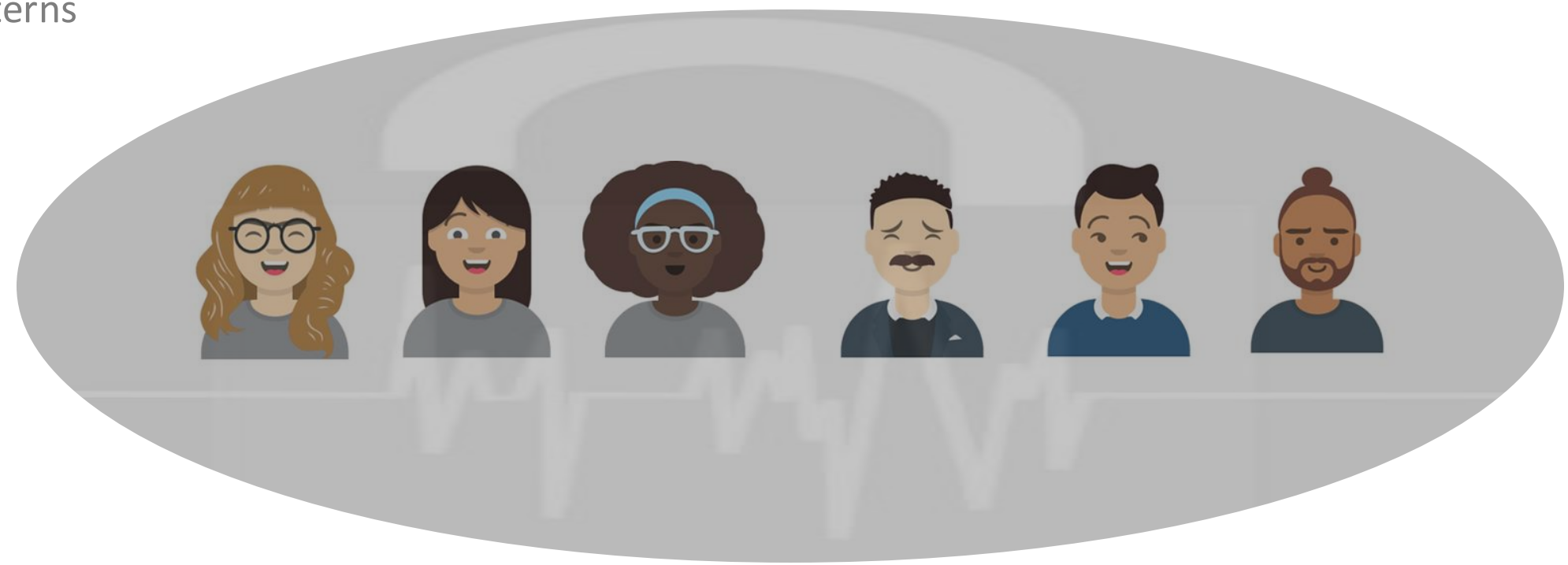
Assessing and Mitigating Bias in Artificial Intelligence Applications for Mental Health

Steps for incorporating bias assessment and mitigation in AI. Stakeholders and the target population should be interviewed at each stage and their feedback incorporated.

Timmons, A. C., Duong, J. B., Simo Fiallo, N., Lee, T., Vo, H. P. Q., Ahle, M. W., Comer, J. S., Brewer, L. C., Frazier, S. L., & Chaspari, T. (2023). A call to action on assessing and mitigating bias in artificial intelligence applications for mental health. Perspectives on Psychological Science, 18(5), 1062-1096.

Bias in speech-based machine learning

- Speech is inherently confounded by various demographic factors (e.g., biological sex, race)
- Differences in anatomical structure that depend on factors such as biological sex and age, influence the vocal fold size and vocal tract length, contributing to observed significant differences among socio-demographic groups in terms of acoustic measures
- Societal expectations, gender roles, cultural norms, and language background can influence speech patterns



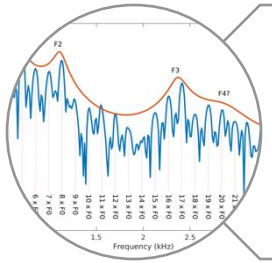
Yang, M., El-Attar, A. A., & Chaspari, T. (2024). Deconstructing demographic bias in speech-based machine learning models for digital health. Frontiers in Digital Health, 6, 1351637.

University of Colorado **Boulder**

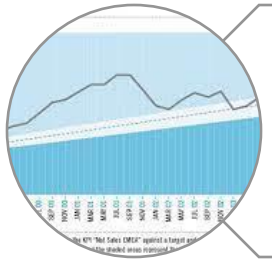
HUBBS
HUman Bio-Behavioral Signals Lab

Bias in speech-based machine learning

Research objectives



To examine acoustic feature differences among demographic groups as a potential source of algorithmic bias in speech-based ML for mental health



To underscore the significance of preserving potentially meaningful distinctions among populations

To examine performance differences among groups and explore the effectiveness of de-biasing methods in the domain of mental health analytics

Yang, M., El-Attar, A. A., & Chaspari, T. (2024). Deconstructing demographic bias in speech-based machine learning models for digital health. Frontiers in Digital Health, 6, 1351637.

Bias in speech-based machine learning

Datasets

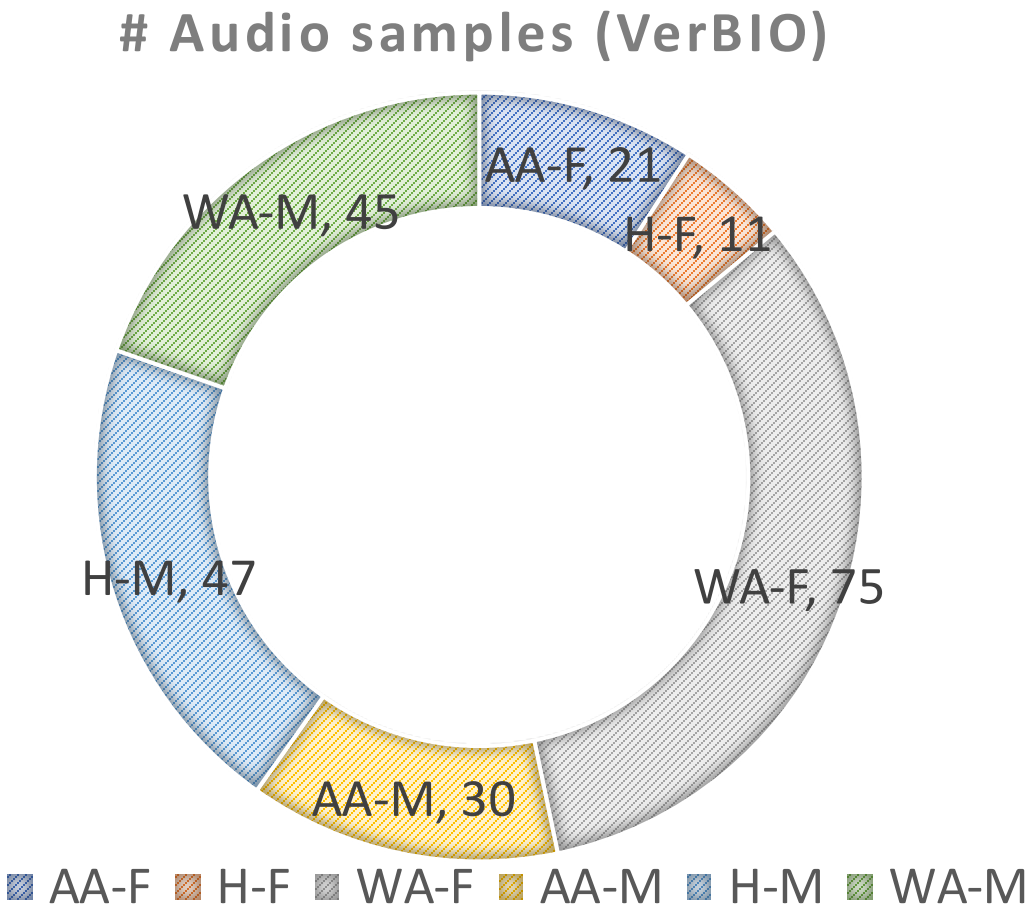
	VerBIO (Yadav et al., 2020)	DAIC-WOZ (Gratch et al., 2014)
Elicitation method	Real-life and virtual-reality public speaking presentations designed to elicit affective and stressor responses	Clinical interviews designed to support diagnosis of psychological conditions
Participants	30 Native English speakers (14 female, 16 male)	107 participants (44 female, 63 male)
Outcome	Self-reported anxiety	Patient Health Questionnaire (PHQ-8) Depression (PHQ-8>9)
Acoustic features	F0, loudness, jitter, shimmer	extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) indicating frequency, energy/amplitude, and spectral balance

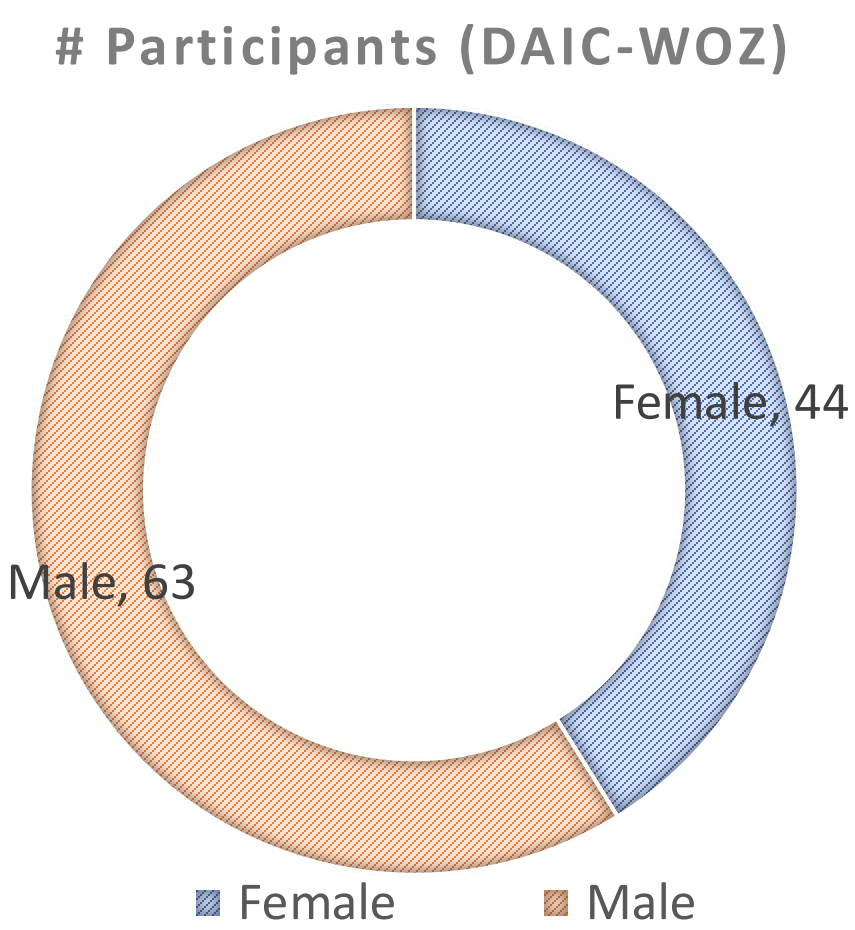
Yang, M., El-Attar, A. A., & Chaspari, T. (2024). Deconstructing demographic bias in speech-based machine learning models for digital health. Frontiers in Digital Health, 6, 1351637.

University of Colorado **Boulder**

HUBBS
HUman Bio-Behavioral Signals Lab

Bias in speech-based machine learning





AA-F: African American female, H-F: Hispanic female, WA-F: White American female, AA-M: African American male, H-F: Hispanic male, WA-M: White American male

Yang, M., El-Attar, A. A., & Chaspari, T. (2024). Deconstructing demographic bias in speech-based machine learning models for digital health. Frontiers in Digital Health, 6, 1351637.

University of Colorado **Boulder**

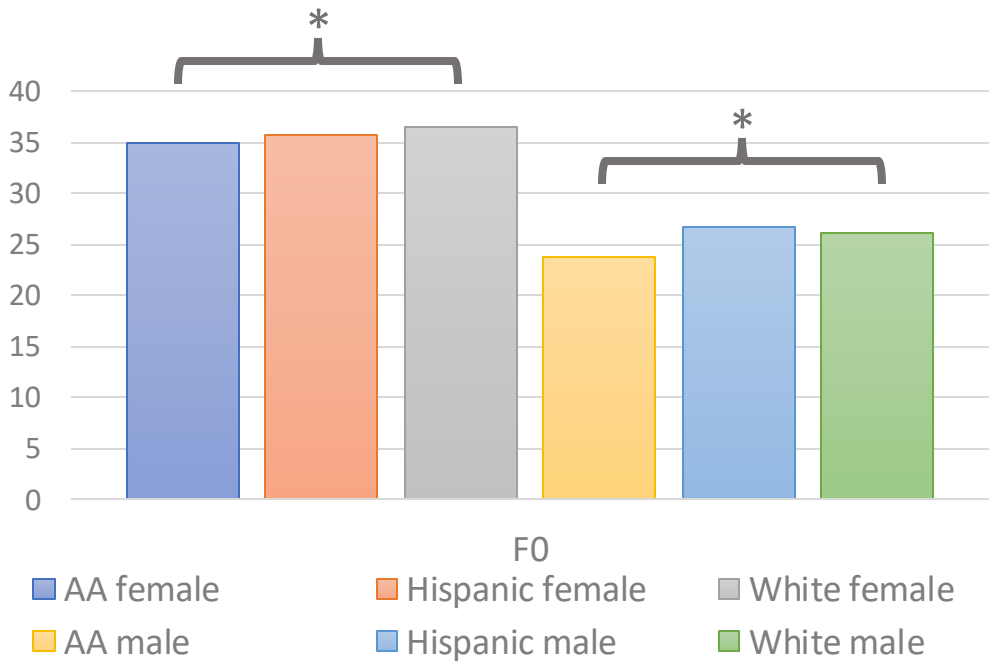
HUBBS HUman Bio-Behavioral Signals Lab

Bias in speech-based machine learning

Task 1: To examine acoustic feature differences among demographic groups as a potential source of algorithmic bias in speech-based ML decision-making (VerBIO dataset)

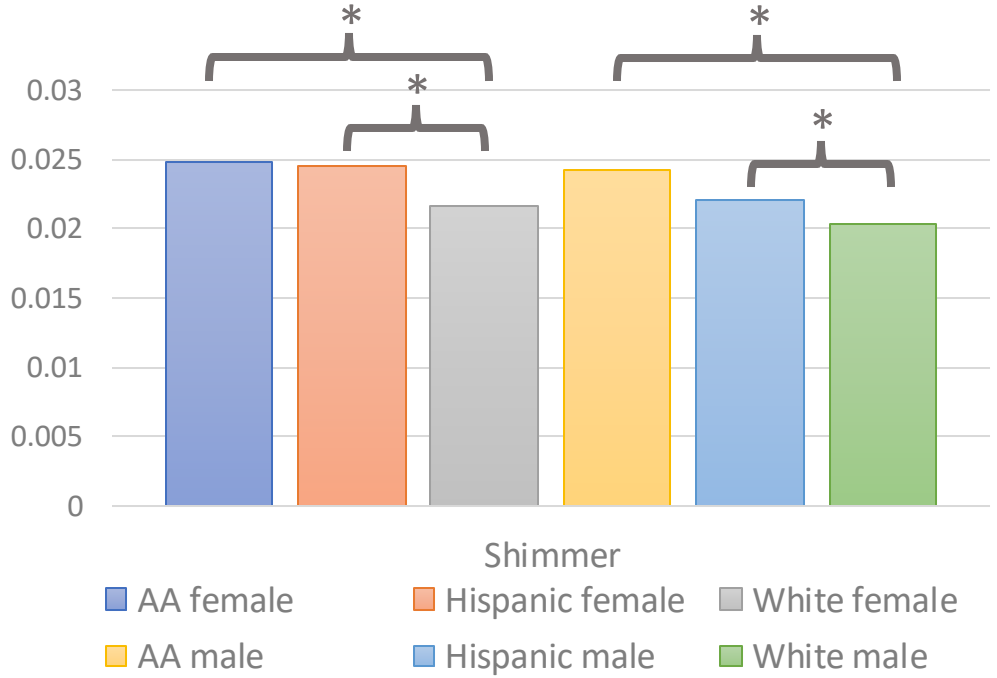
- Significant differences wrt acoustic measures in terms of biological sex and race/ethnicity

Fundamental frequency (F0)



Biological sex	F(2, 223)=19.96, p<0.01
Race/ethnicity	F(1, 223)=1083.710, p<0.01
Sex x Race/ethnicity	F(2, 223)=2.496, p<0.01

Shimmer



Biological sex	F(2, 223)=15.024, p<0.01
Race/ethnicity	F(1, 223)=5.868, p<0.01
Sex x Race/ethnicity	F(2, 223)=0.523, p<0.01

University of Colorado **Boulder**

HUBBS
HUman Bio-Behavioral Signals Lab

Bias in speech-based machine learning

Task 1: To examine acoustic feature differences among female and male speakers as a potential source of algorithmic bias in speech-based ML decision-making (DAIC-WOZ dataset)

Top 20 most discriminative features of depression

Feature	Group
F3bandwidth_sma3nz_amean	Frequency
F0semitoneFrom27.5Hz_sma3nz_percentile50.0	Frequency
F0semitoneFrom27.5Hz_sma3nz_percentile20.0	Frequency
F0semitoneFrom27.5Hz_sma3nz_amean	Frequency
F0semitoneFrom27.5Hz_sma3nz_percentile80.0	Frequency
logRelF0-H1-H2_sma3nz_amean	Spectral balance
equivalentSoundLevel_dBp	Energy/amplitude
mfcc2V_sma3nz_amean	Spectral balance
HNRdBACF_sma3nz_amean	Energy/amplitude
F1amplitudeLogRelF0_sma3nz_amean	Spectral balance
mfcc2_sma3_amean	Energy/amplitude
hammarbergIndexV_sma3nz_amean	Spectral balance
F3amplitudeLogRelF0_sma3nz_amean	Spectral balance
F2amplitudeLogRelF0_sma3nz_amean	Spectral balance
logRelF0-H1-A3_sma3nz_amean	Spectral balance
loudnessPeaksPerSec	Energy/amplitude
F1amplitudeLogRelF0_sma3nz_stddevNorm	Spectral balance
F2bandwidth_sma3nz_amean	Frequency
F3amplitudeLogRelF0_sma3nz_stddevNorm	Spectral balance
MeanUnvoicedSegmentLength	Temporal

Top 20 most discriminative features of gender

Feature	Group
F1Frequency_sma3nz_amean	Frequency
mfcc4_sma3_stddevNorm	Spectral balance
F0semitoneFrom27.5Hz_sma3nz_meanFallingSlope	Frequency
equivalentSoundLevel_dBp	Energy/amplitude
F2Frequency_sma3nz_amean	Frequency
F3Frequency_sma3nz_amean	Frequency
logRelF0-H1-H2_sma3nz_stddevNorm	Spectral balance
F0semitoneFrom27.5Hz_sma3nz_percentile80.0	Frequency
F0semitoneFrom27.5Hz_sma3nz_pctlrange0-2	Frequency
F1amplitudeLogRelF0_sma3nz_amean	Spectral balance
F3amplitudeLogRelF0_sma3nz_amean	Spectral balance
F2amplitudeLogRelF0_sma3nz_amean	Spectral balance
F0semitoneFrom27.5Hz_sma3nz_percentile50.0	Frequency
mfcc1_sma3_amean	Energy/amplitude
hammarbergIndexUV_sma3nz_amean	Spectral balance
F0semitoneFrom27.5Hz_sma3nz_amean	Frequency
mfcc1V_sma3nz_amean	Energy/amplitude
loudnessPeaksPerSec	Energy/amplitude
jitterLocal_sma3nz_stddevNorm	Frequency
mfcc4V_sma3nz_stddevNorm	Energy/amplitude

University of Colorado **Boulder**

HUBBS
HUman Bio-Behavioral Signals Lab

Bias in speech-based machine learning

Task 1: To examine acoustic feature differences among female and male speakers as a potential source of algorithmic bias in speech-based ML decision-making (DAIC-WOZ dataset)

Top 20 most discriminative features of depression

Feature	Group
F3bandwidth_sma3nz_amean	Frequency
F0semitoneFrom27.5Hz_sma3nz_percentile50.0	Frequency
F0semitoneFrom27.5Hz_sma3nz_percentile20.0	Frequency
F0semitoneFrom27.5Hz_sma3nz_amean	Frequency
F0semitoneFrom27.5Hz_sma3nz_percentile80.0	Frequency
logRelF0-H1-H2_sma3nz_amean	Spectral balance
equivalentSoundLevel_dBp	Energy/amplitude
mfcc2V_sma3nz_amean	Spectral balance
HNRdBACF_sma3nz_amean	Energy/amplitude
F1amplitudeLogRelF0_sma3nz_amean	Spectral balance
mfcc2_sma3_amean	Energy/amplitude
hammarbergIndexV_sma3nz_amean	Spectral balance
F3amplitudeLogRelF0_sma3nz_amean	Spectral balance
F2amplitudeLogRelF0_sma3nz_amean	Spectral balance
logRelF0-H1-A3_sma3nz_amean	Spectral balance
loudnessPeaksPerSec	Energy/amplitude
F1amplitudeLogRelF0_sma3nz_stddevNorm	Spectral balance
F2bandwidth_sma3nz_amean	Frequency
F3amplitudeLogRelF0_sma3nz_stddevNorm	Spectral balance
MeanUnvoicedSegmentLength	Temporal

Top 20 most discriminative features of gender

Feature	Group
F1Frequency_sma3nz_amean	Frequency
mfcc4_sma3_stddevNorm	Spectral balance
F0semitoneFrom27.5Hz_sma3nz_meanFallingSlope	Frequency
equivalentSoundLevel_dBp	Energy/amplitude
F2Frequency_sma3nz_amean	Frequency
F3Frequency_sma3nz_amean	Frequency
logRelF0-H1-H2_sma3nz_stddevNorm	Spectral balance
F0semitoneFrom27.5Hz_sma3nz_percentile80.0	Frequency
F0semitoneFrom27.5Hz_sma3nz_pctlrange0-2	Frequency
F1amplitudeLogRelF0_sma3nz_amean	Spectral balance
F3amplitudeLogRelF0_sma3nz_amean	Spectral balance
F2amplitudeLogRelF0_sma3nz_amean	Spectral balance
F0semitoneFrom27.5Hz_sma3nz_percentile50.0	Frequency
mfcc1_sma3_amean	Energy/amplitude
hammarbergIndexUV_sma3nz_amean	Spectral balance
F0semitoneFrom27.5Hz_sma3nz_amean	Frequency
mfcc1V_sma3nz_amean	Energy/amplitude
loudnessPeaksPerSec	Energy/amplitude
jitterLocal_sma3nz_stddevNorm	Frequency
mfcc4V_sma3nz_stddevNorm	Energy/amplitude

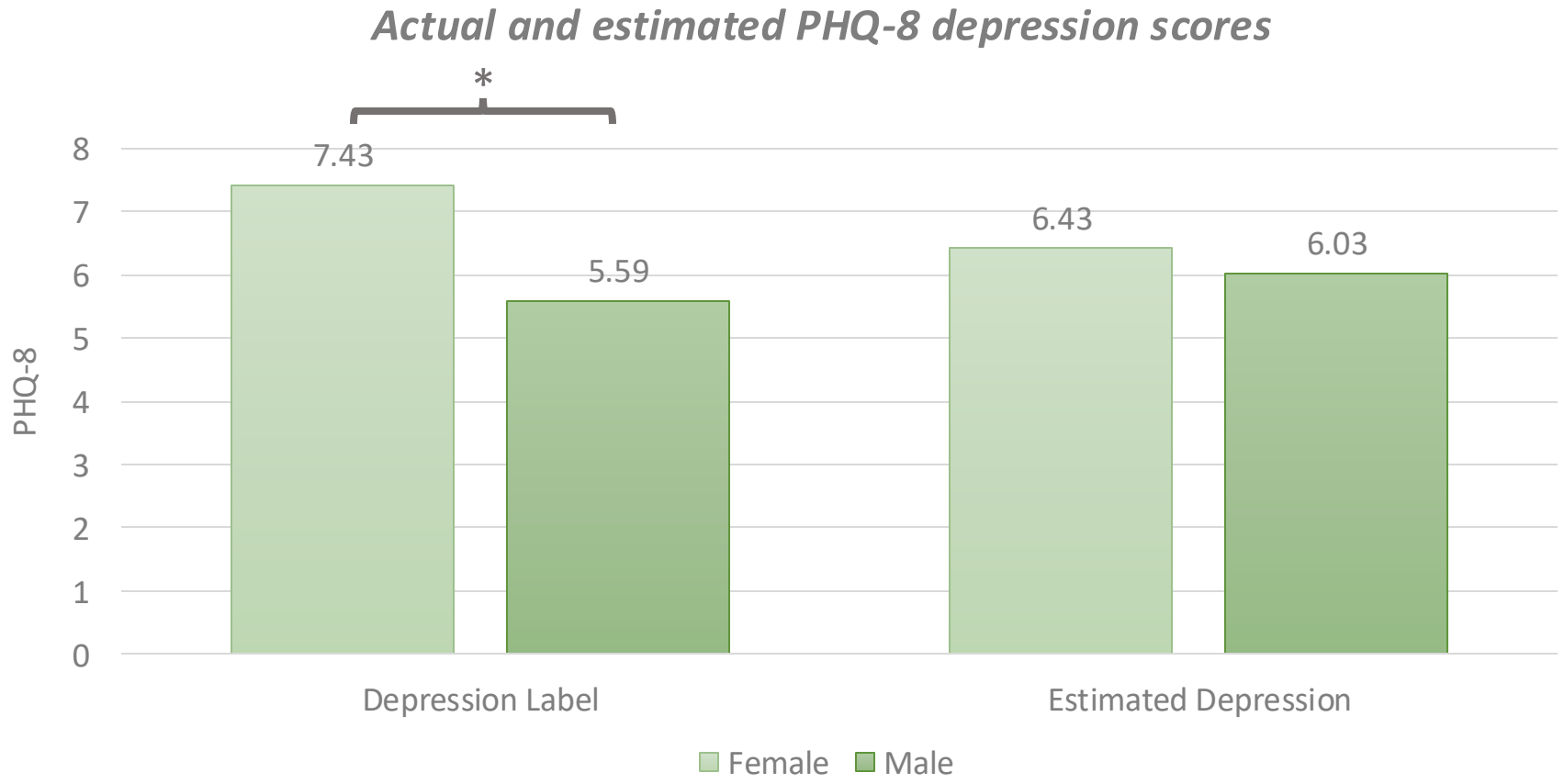
University of Colorado **Boulder**

HUBBS HUman Bio-Behavioral Signals Lab

Bias in speech-based machine learning

Task 2: To underscore the significance of preserving potentially meaningful distinctions among female and male speakers (DAIC-WOZ)

- Estimation of PHQ-8 from acoustic measures via linear regression
- Statistically significant differences in ground truth PHQ-8 not preserved in estimated score
- Depression prevalence in DAIC-WOZ: 38.6% for female participants, 22.2% for male participants



University of Colorado **Boulder**

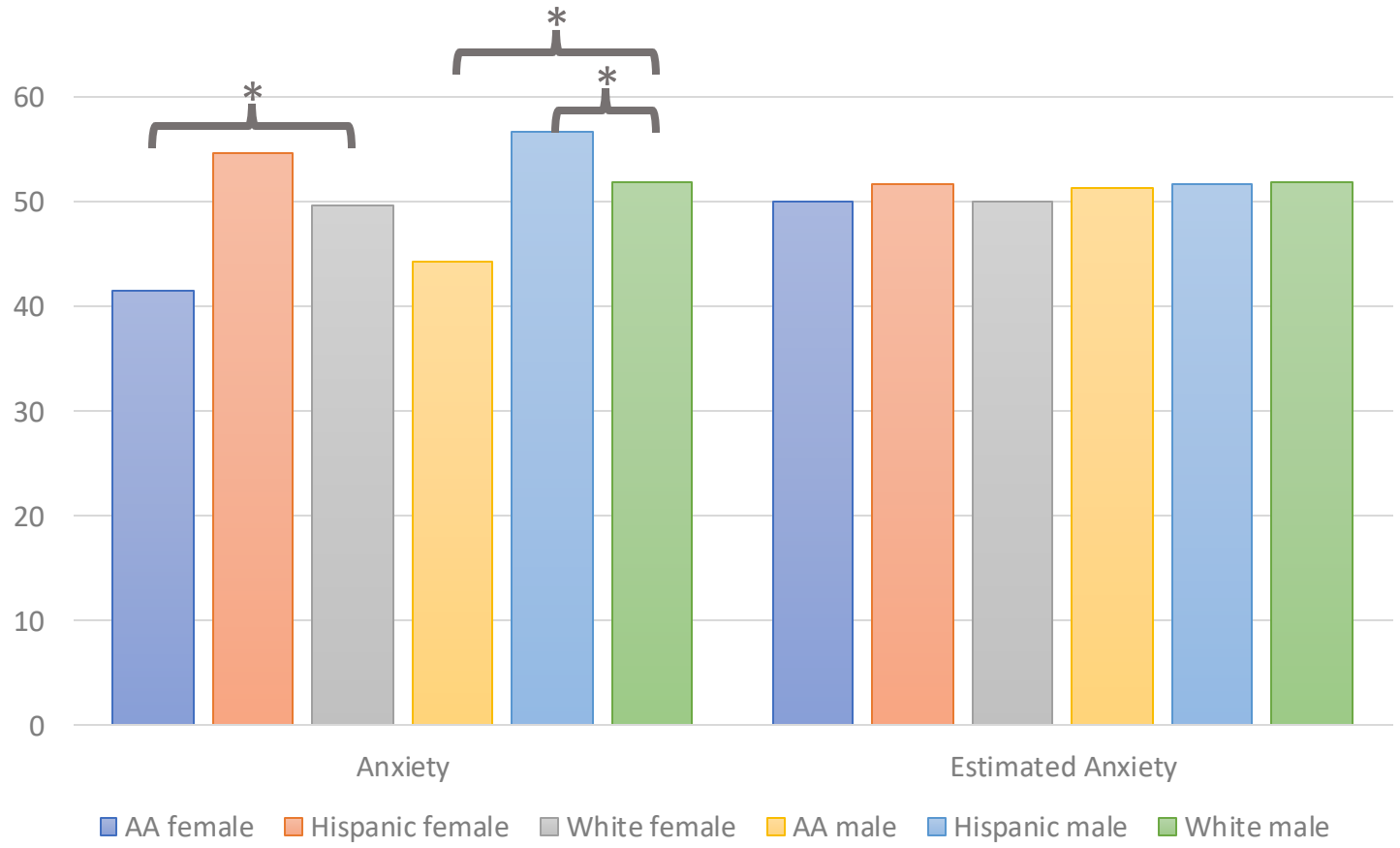
HUBBS HUman Bio-Behavioral Signals Lab

Bias in speech-based machine learning

Task 2: To underscore the significance of preserving potentially meaningful distinctions among populations (VerBIO)

- Estimation of anxiety from acoustic measures via linear regression (VerBIO dataset)
- Statistically significant differences in ground truth anxiety not preserved in estimated anxiety

Mean of self-reported and estimated anxiety per demographic group



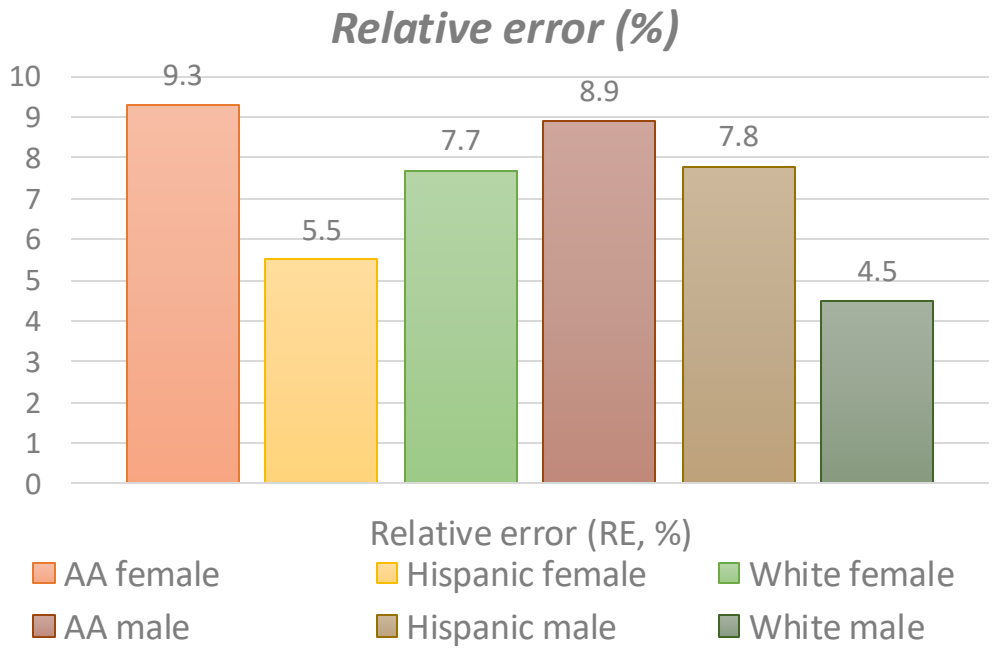
University of Colorado **Boulder**

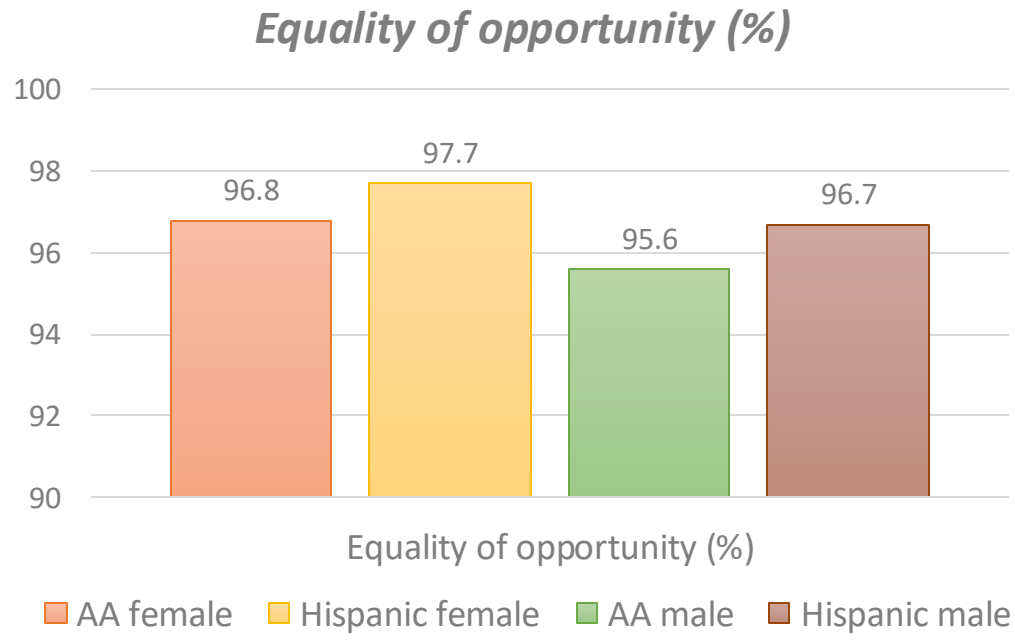
HUBBS HUman Bio-Behavioral Signals Lab

Bias in speech-based machine learning

Task 3: To examine performance differences among groups (VerBIO dataset)

- Differences wrt anxiety estimation performance in terms of biological sex and race/ethnicity
- Hispanic female speakers have lower equality of opportunity than African American female speakers
- African American male speakers have lower equality of opportunity than Hispanic male speakers





Yang, M., El-Attar, A., & Chaspari, T. Deconstructing demographic bias in human-centered speech-based machine learning models. Submitted.

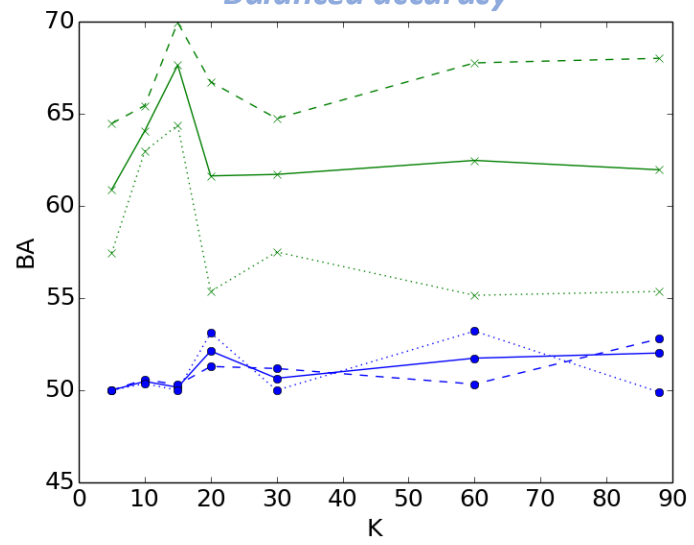
University of Colorado **Boulder**

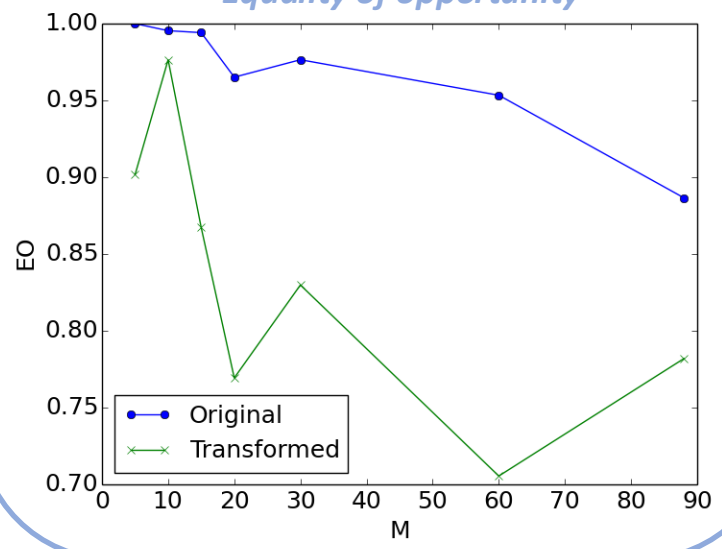
Bias in speech-based machine learning

Task 3: To examine performance differences among groups and explore the effectiveness of de-biasing methods (DAIC-WOZ)

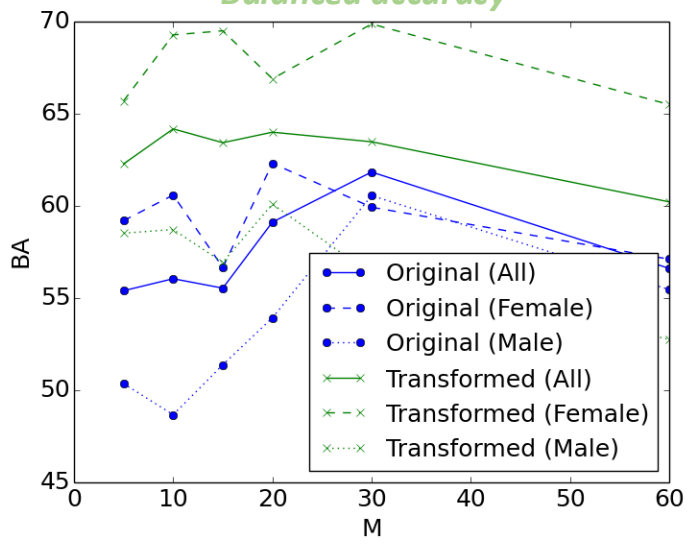
- Preserving the K=10 most relevant features to depression, followed by feature transformation via adversarial learning, yields the best tradeoff between accuracy and equality of opportunity
- Removing the M=20-30 most relevant features to biological sex improves depression classification accuracy, further improved after transforming via adversarial learning
- EO=1-|Error(sensitive group)-Error(non-sensitive group)|

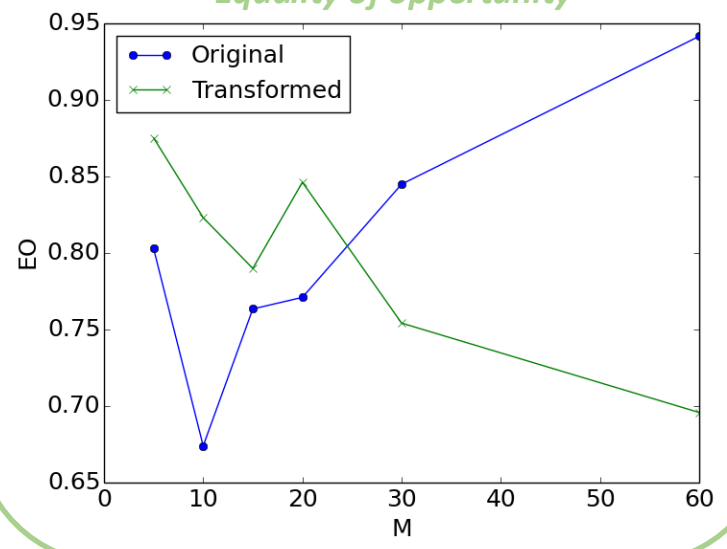
Including K most relevant features to depression





Removing M most relevant features to biological sex



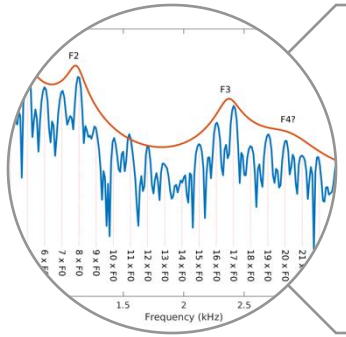


University of Colorado **Boulder**

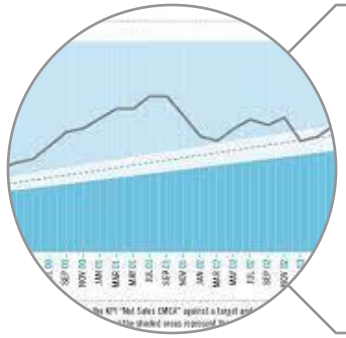
HUBBS
HUman Bio-Behavioral Signals Lab

Bias in speech-based machine learning

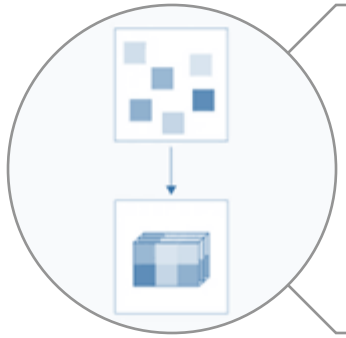
Summary of findings



Acoustic features are indicative of socio-demographic information related to biological sex and race/ethnicity contributing to variations in the performance of the machine learning (ML) models



ML models might fail to retain potentially meaningful differences among populations at the output



Feature transformation via adversarial learning and removing features relevant to socio-demographic information can result in reduced performance differences among groups

University of Colorado **Boulder**

HUBBS
Human Bio-Behavioral Signals Lab

Integrating domain expertise into the design of explainable AI for supporting clinician decision-making and training

Kexin Feng, Ph.D. Student

Clinician-AI collaboration to augment decision-making

- AI can provide decision support tools that can offer treatment recommendations and assist clinicians during training

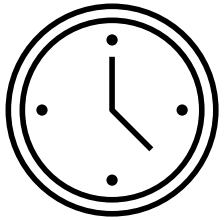
Clinician-AI collaboration to augment decision-making

Depression affects 5%
adults worldwide

30-50% female
patients over-diagnosed

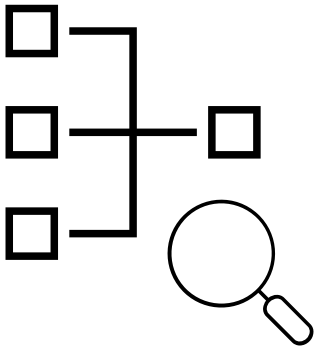
Primary care physicians recognize depression
50% times

System factors impacting human trust in AI



Timing of AI feedback

- Concurrent, sequential, on demand, time delayed



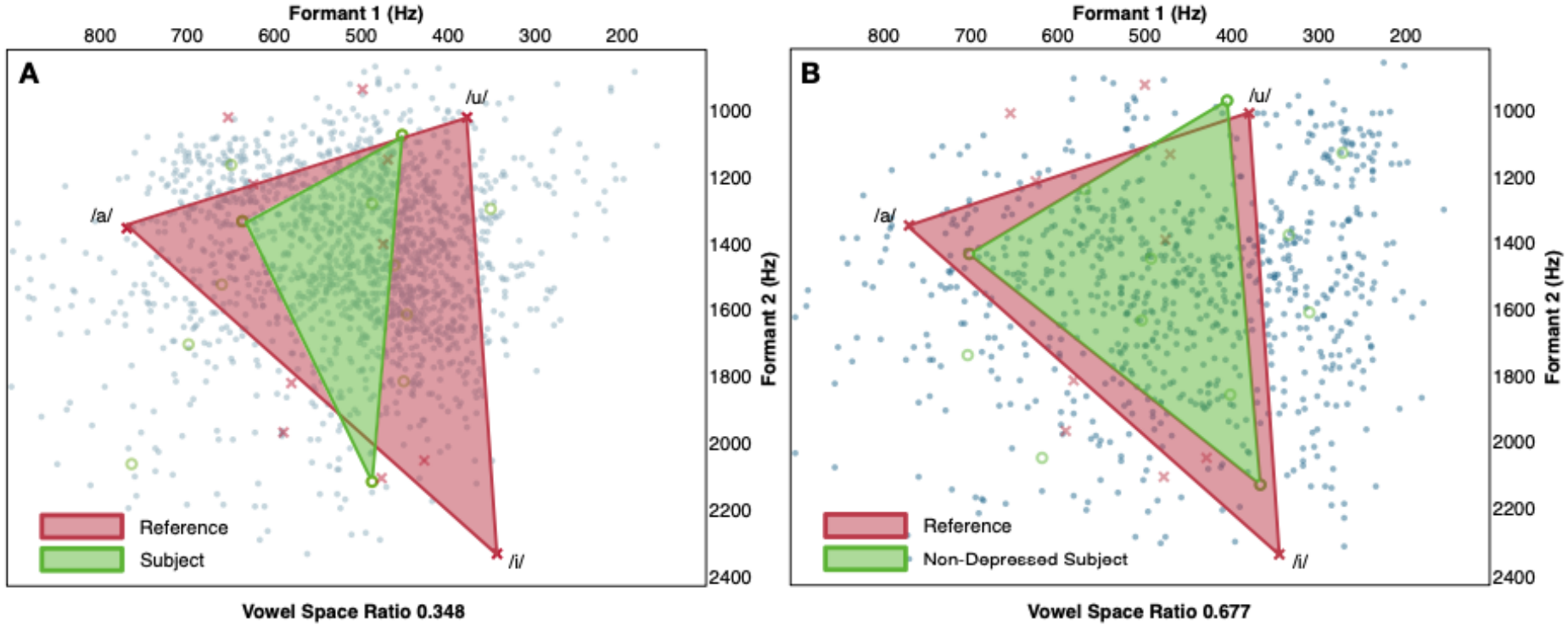
Type of AI feedback

- Decision output
- Confidence
- Explanation
  - Feature attribution
  - Input attribution
  - Showcasing similar examples
  - Counterfactuals
  - Natural language dialog

Steyvers, M., & Kumar, A. (2023). Three challenges for AI-assisted decision-making. Perspectives on Psychological Science, 17456916231181102.

Clinician-AI collaboration to augment decision-making

Observation 1: Patients with depression depict reduced 'vowel space' (the frequency range between the first and second speech formant)

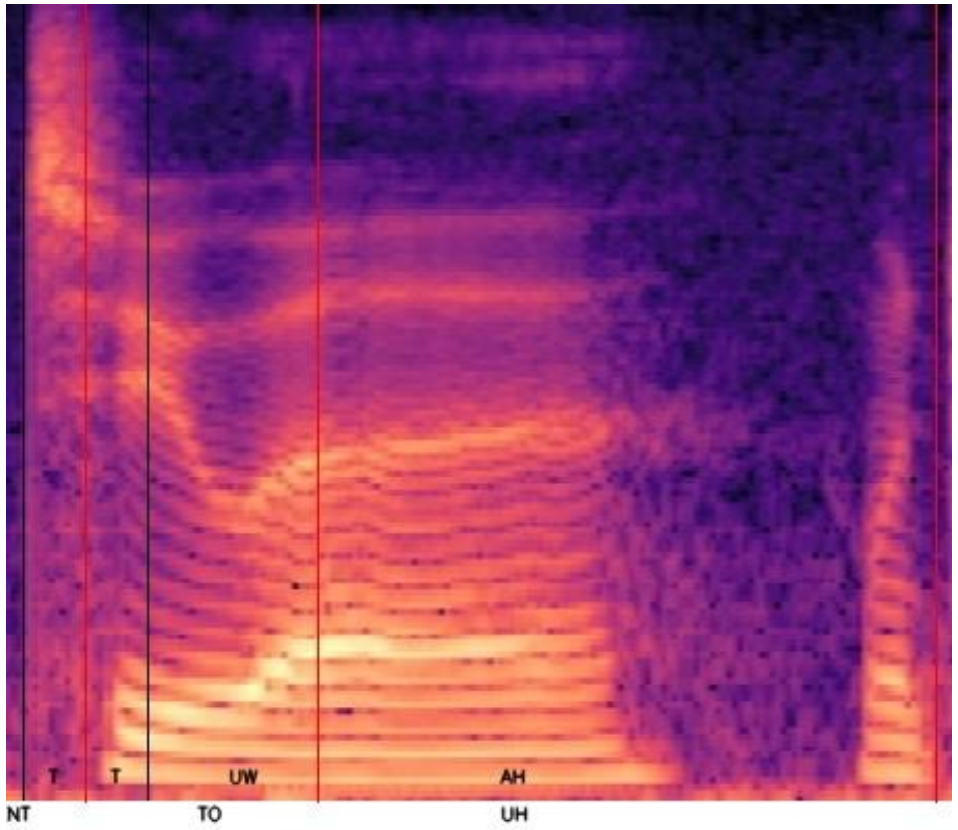


Source: Scherer, Stefan, et al. "Self-reported symptoms of depression and PTSD are associated with reduced vowel space in screening interviews." IEEE Transactions on Affective Computing 7.1 (2015): 59-73

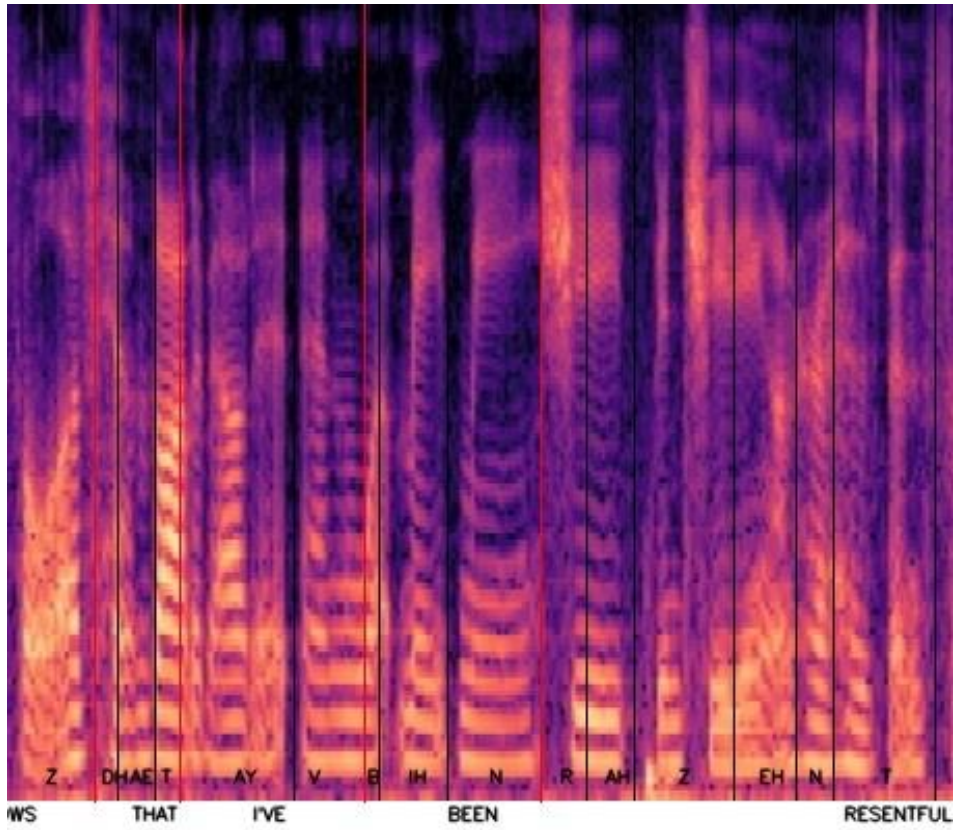
Clinician-AI collaboration to augment decision-making

Observation 2: Patients with depression are more likely to lack ‘sustained’ energy over high speech frequencies

Healthy participant



Participant with depression



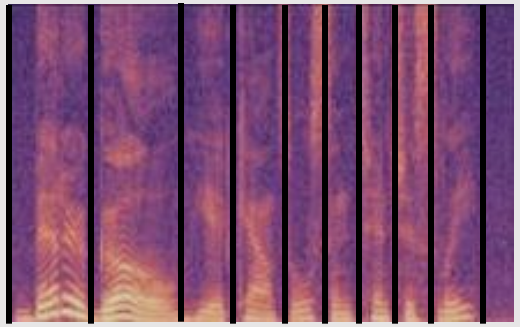
An explainable speech-based system for depression detection

Speech spectrogram: time-frequency representation of speech

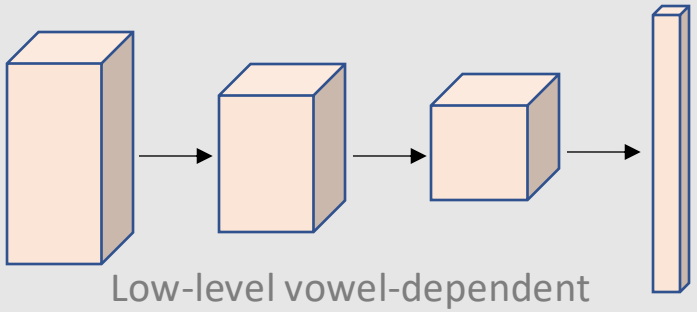
Low-level explanation of depression-specific spectrotemporal variations over semantically meaningful linguistic units (vowels)

High-level explanation showing the contribution of each spoken utterance to the final ML decision

Phoneme segmentation based on speech spectrogram

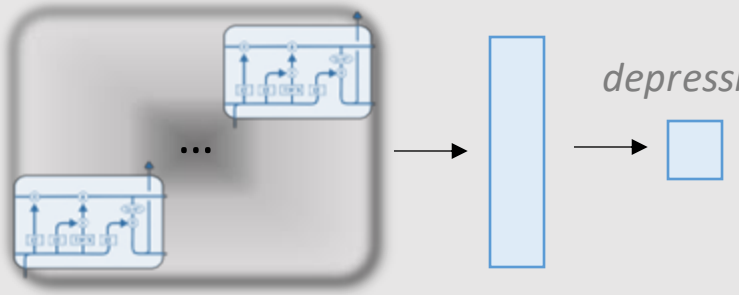


Neural network-based phoneme classification

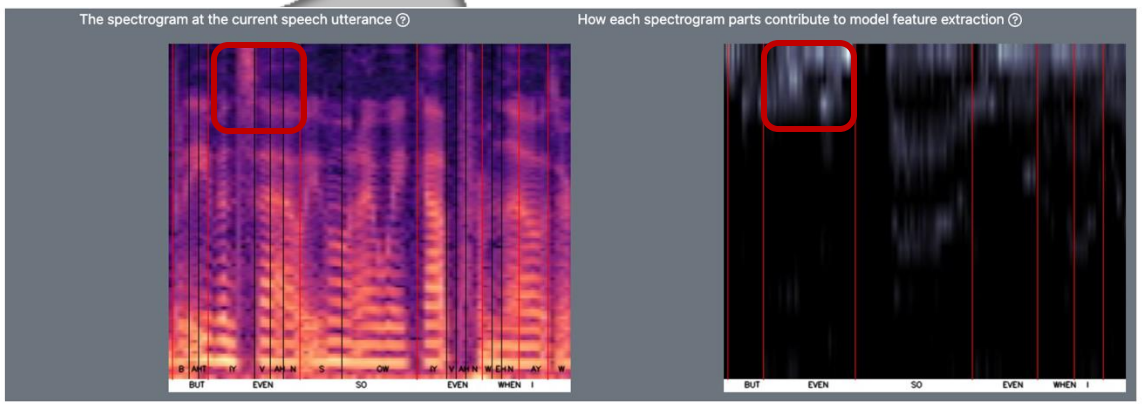


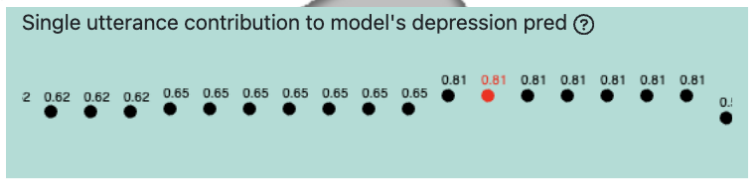
Low-level vowel-dependent embeddings capturing spectrotemporal variations at the vowel region

***High-level descriptors:** Long short-term memory neural network followed by fully-connected layer*



High-level temporal dependencies across speech segments





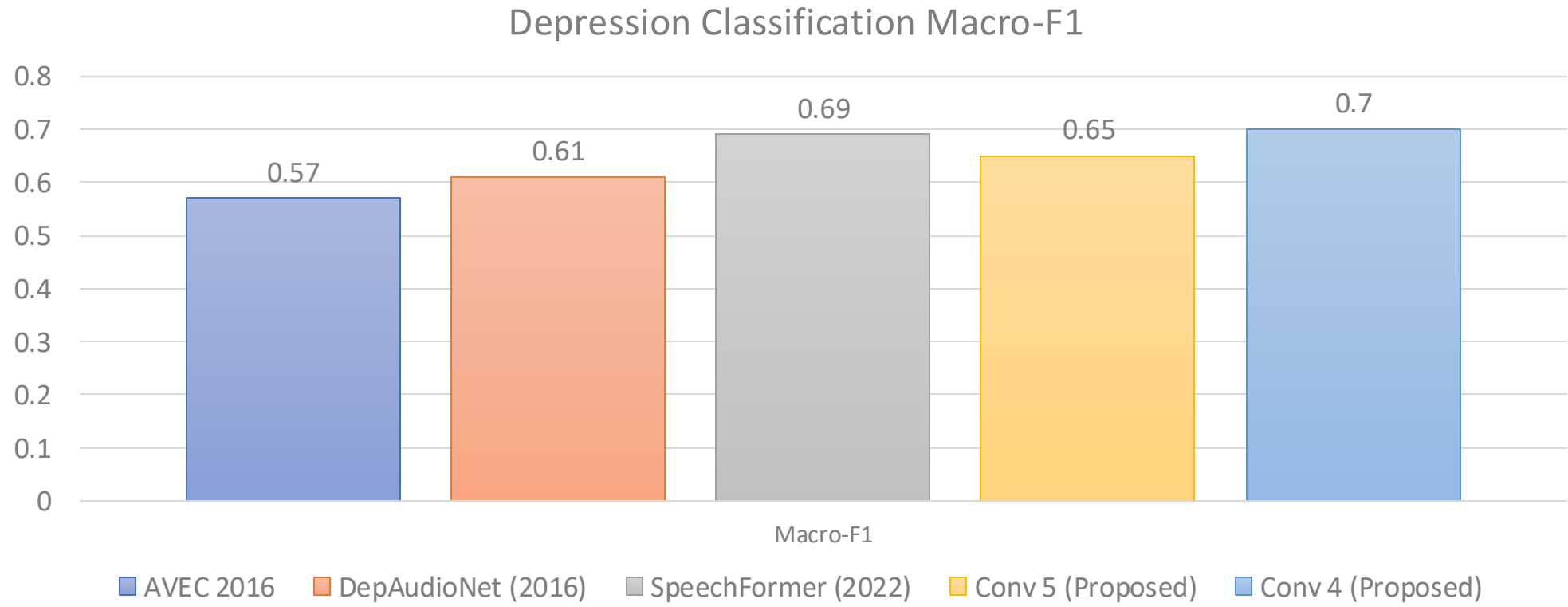
Feng, K., & Chaspari, T. (2022). Toward knowledge-driven speech-based models of depression: Leveraging spectrotemporal variations in speech vowels. In 2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI) (pp. 01-07). IEEE.

University of Colorado **Boulder**

HUBBS HUman Bio-Behavioral Signals Lab

Clinician-AI collaboration to augment decision-making

A knowledge-driven machine learning (ML) method that integrates spectrotemporal information of speech at the vowel-level to identify the depression

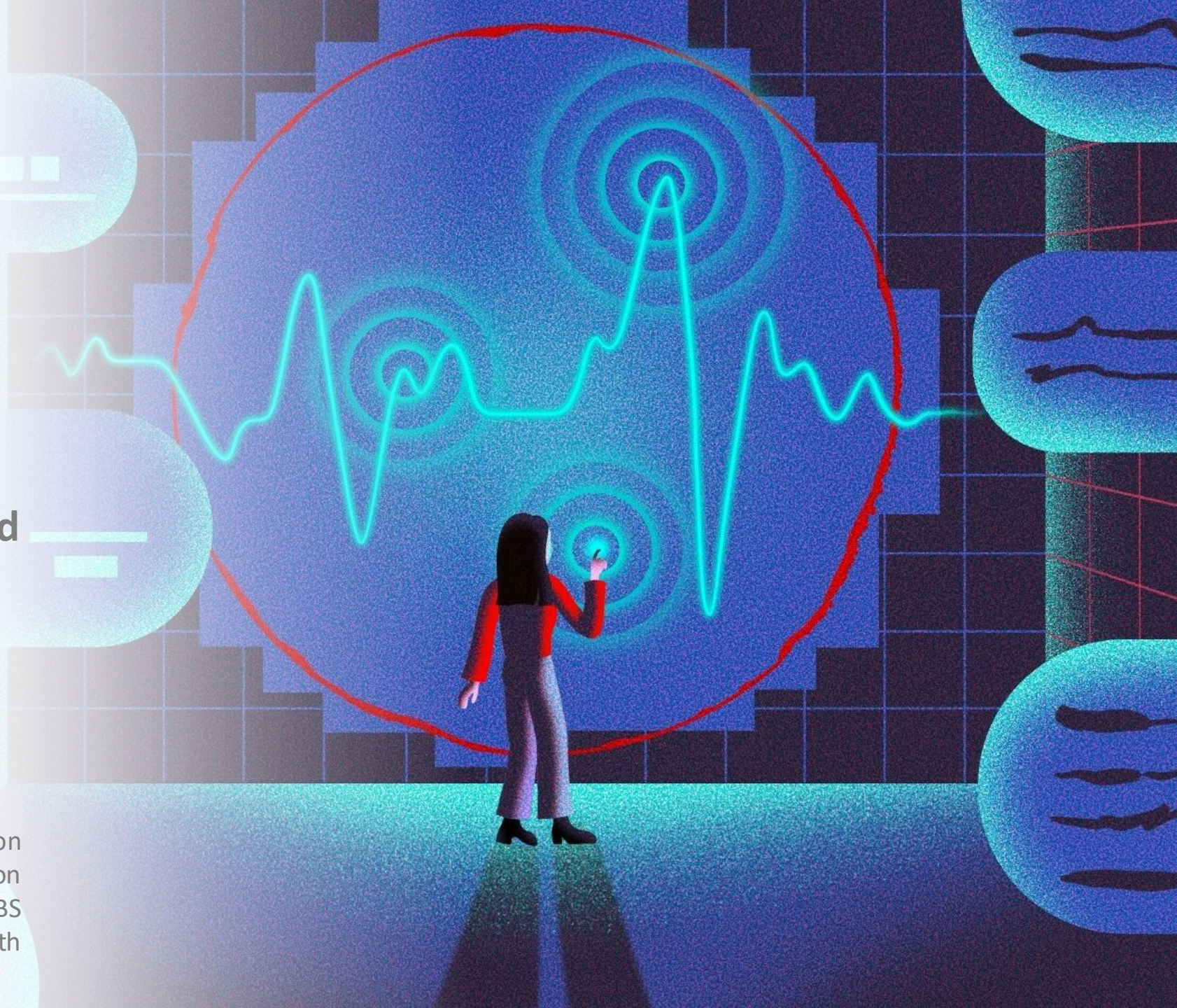


Feng, K., & Chaspari, T. (2024, November). Robust and Explainable Depression Identification from Speech Using Vowel-Based Ensemble Learning Approaches. Accepted in 2024 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI). IEEE. (8-page paper)

Feng, K., & Chaspari, T. (2023, June). A knowledge-driven vowel-based approach of depression classification from speech using data augmentation. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1-5). IEEE.

Feng, K., & Chaspari, T. (2022, September). Toward knowledge-driven speech-based models of depression: Leveraging spectrotemporal variations in speech vowels. In 2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI) (pp. 01-07). IEEE.

Clinician-AI collaboration to augment decision-making



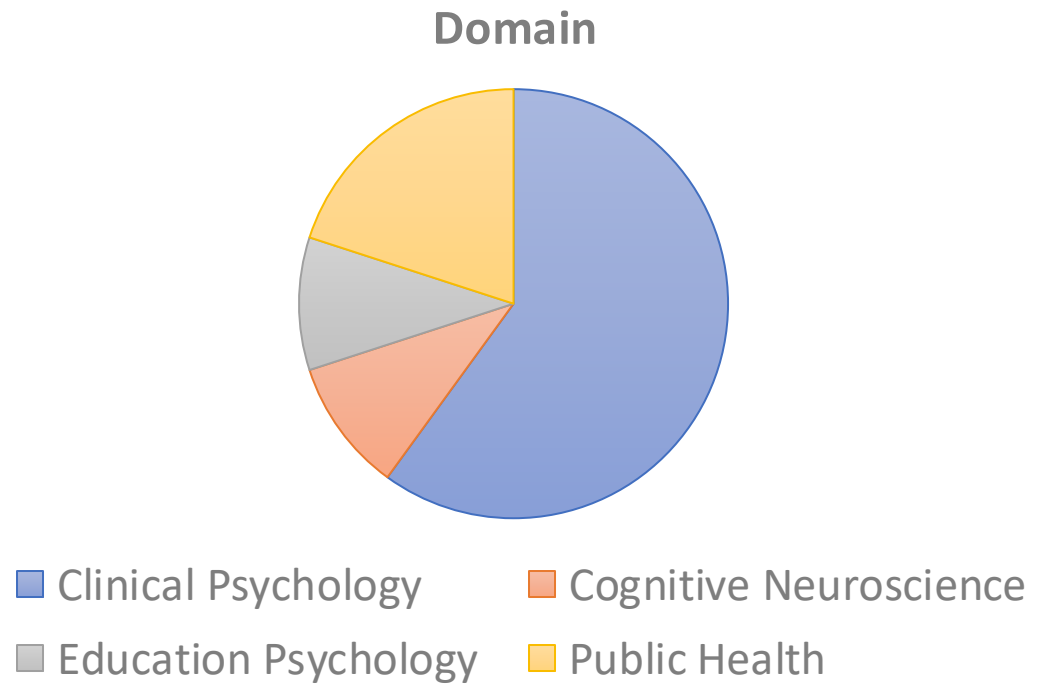
To evaluate clinicians' perceived usability of and trust in the AI system, and usability of design factors associated with the various system components

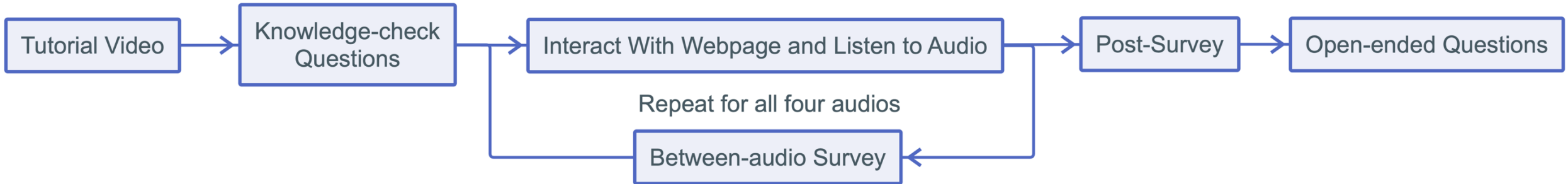
Feng, K., & Chaspari, T. (2024). A Pilot Study on Clinician-AI Collaboration in Diagnosing Depression from Speech. Accepted in 2024 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI). IEEE. (8-page paper)

Clinician-AI collaboration to augment decision-making

User study

- Eligibility criteria: Being graduate student in clinical psychology or related majors, OR licensed psychologist or psychiatrist
- 10 participants (all female) between 20-35 years old, 9 doctoral students
- Each participant reviewed audio files from 4 clinical interviews (~20-35 minutes)
- Between-subjects design to assess usability of various system components (Conditions 1-3)



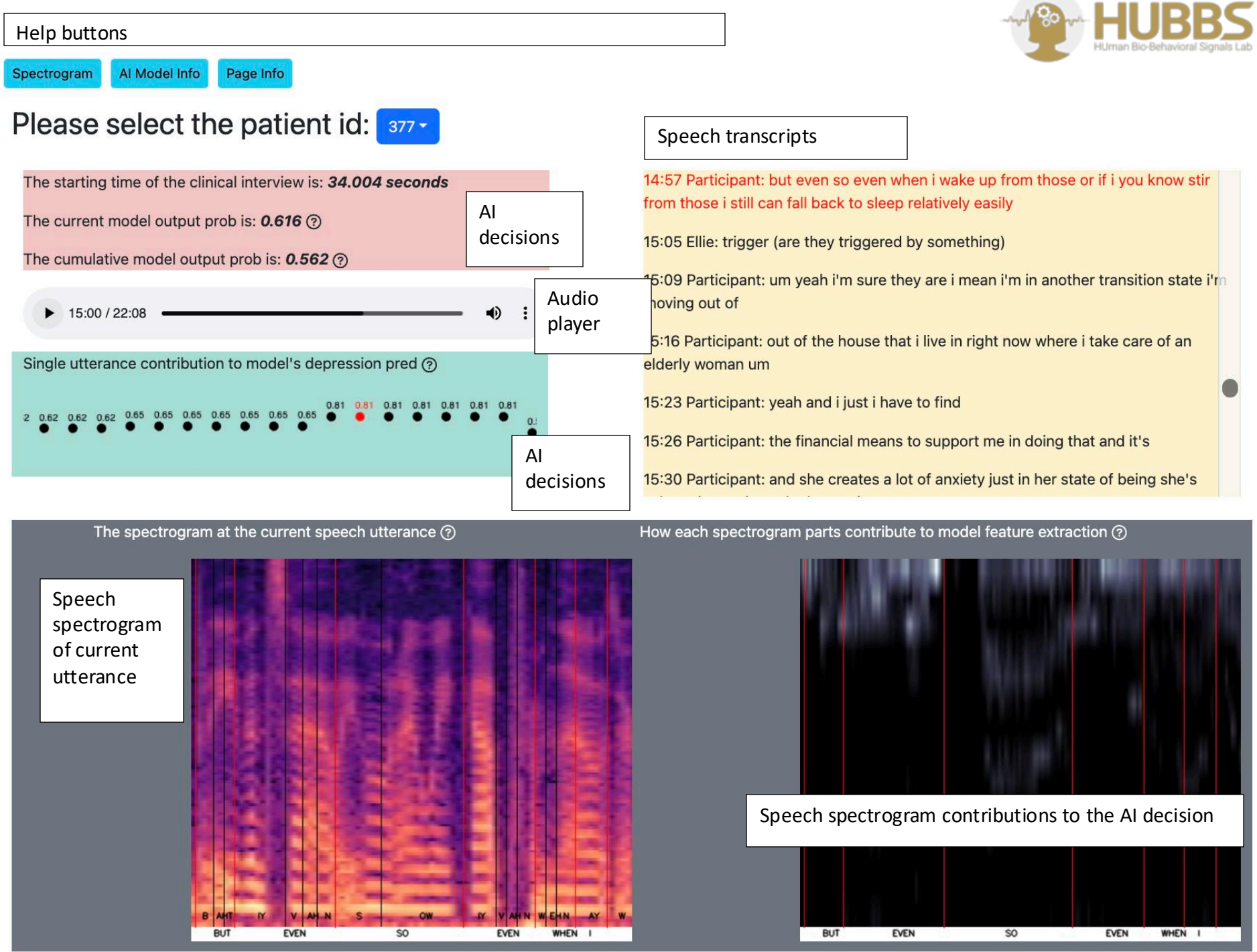


University of Colorado **Boulder**

HUBBS
HUman Bio-Behavioral Signals Lab

Clinician-AI collaboration to augment decision-making

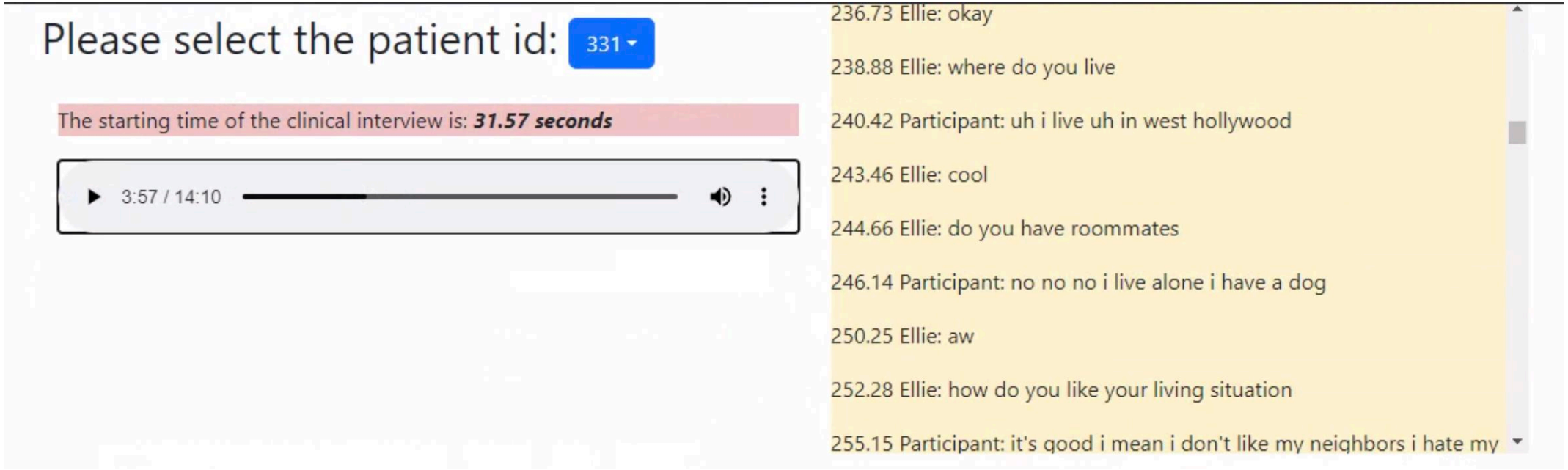
- Identify key utterances in the clinical encounter that are important for diagnosis
- Identify key parts in an utterance depicting spectrotemporal variation indicative of depression



Clinician-AI collaboration to augment decision-making

Condition 1 (no AI)

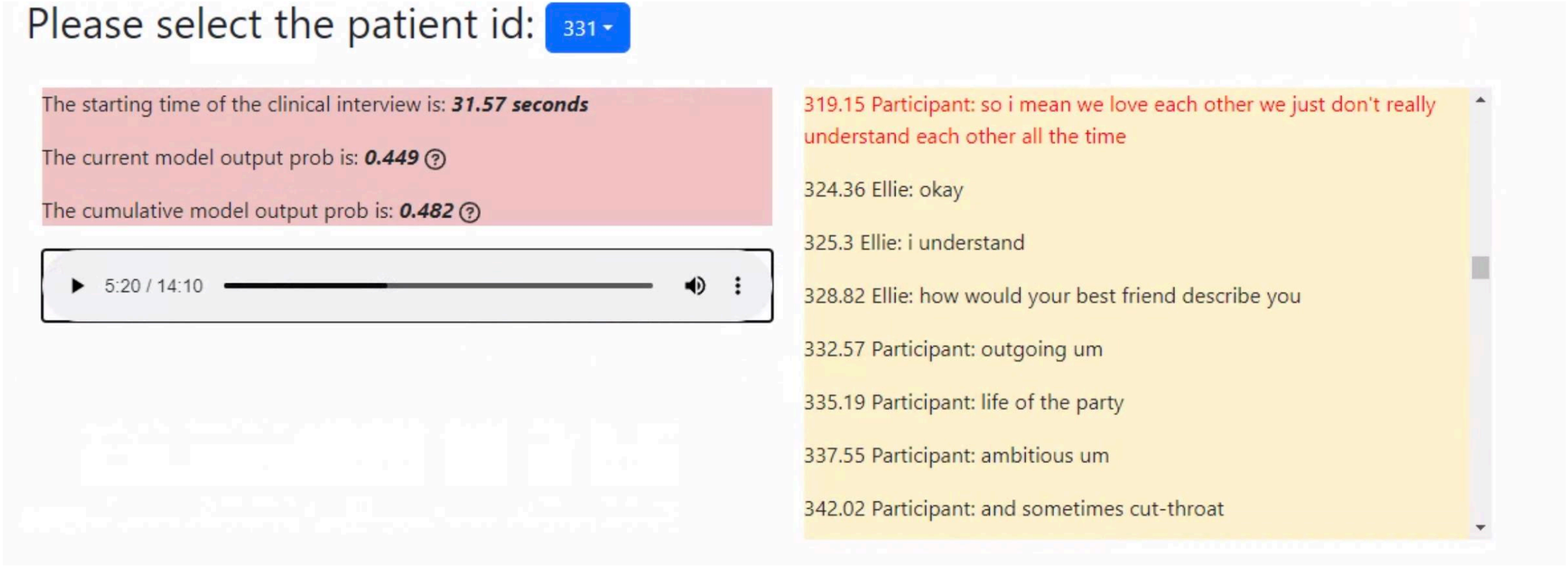
- The participant listened to an audio, while also viewing the corresponding subtitles



Clinician-AI collaboration to augment decision-making

Condition 2 (AI without explanation)

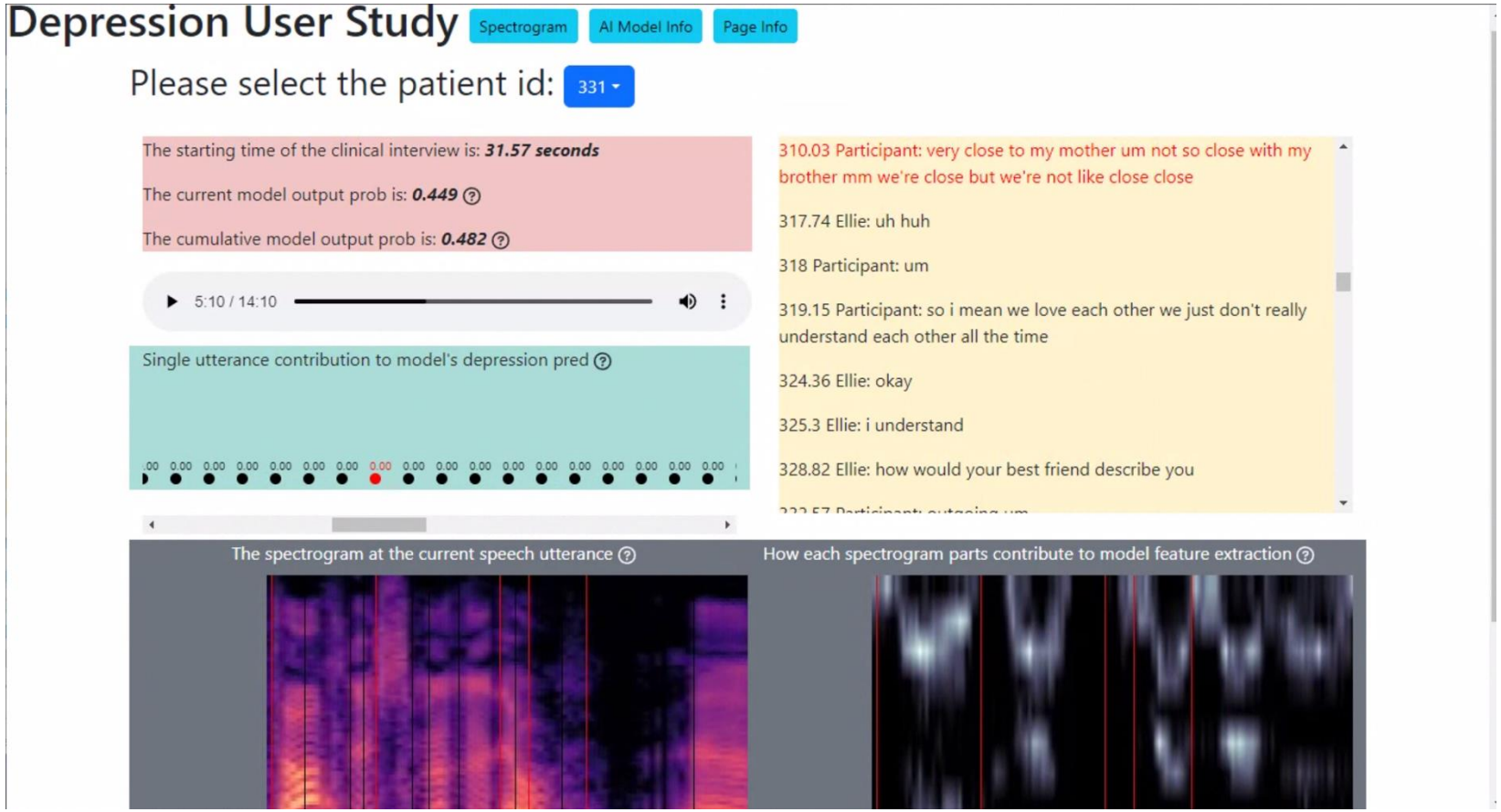
- Same as Condition 1 & participant viewed the probability of depression as estimated by the AI



Clinician-AI collaboration to augment decision-making

Condition 3 (AI with explanation)

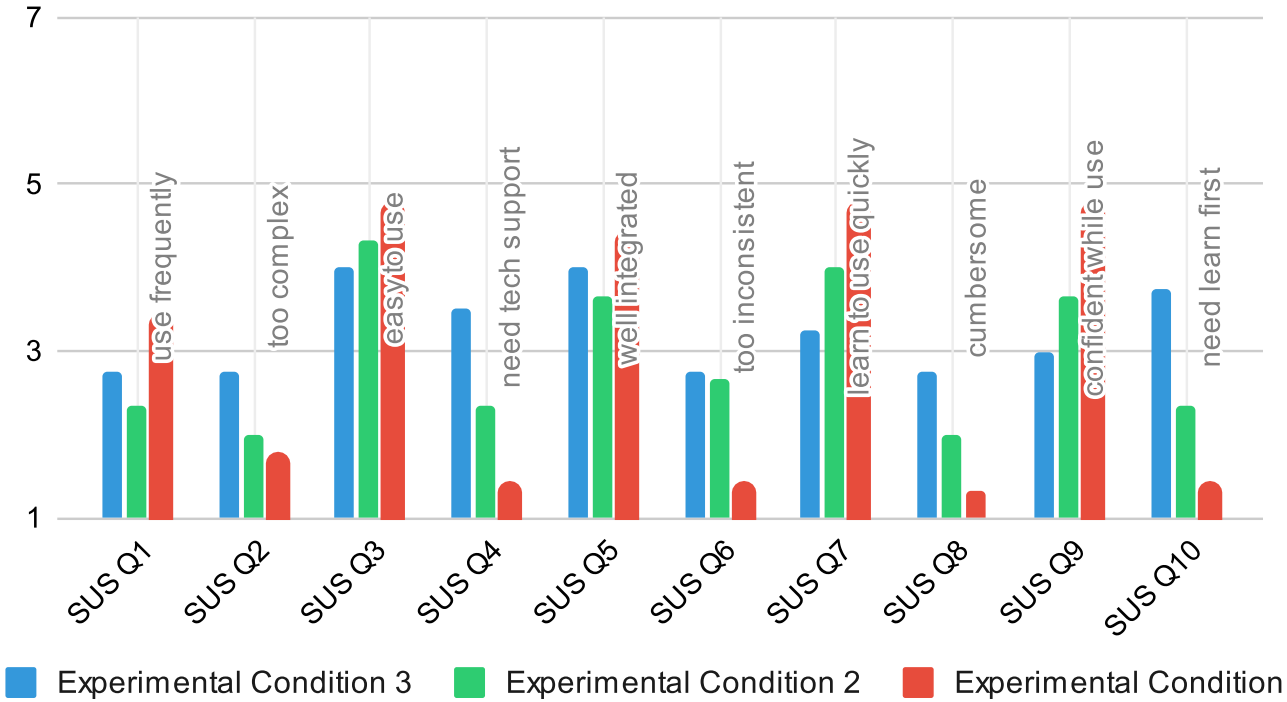
- Same as Condition2 & participant viewed the importance that each utterance had on the decision provided by the AI, had access to the spectrogram of the and the regions of the spectrogram that contributed the most to the AI decision



User study with mental health clinicians

System usability Scale (SUS)

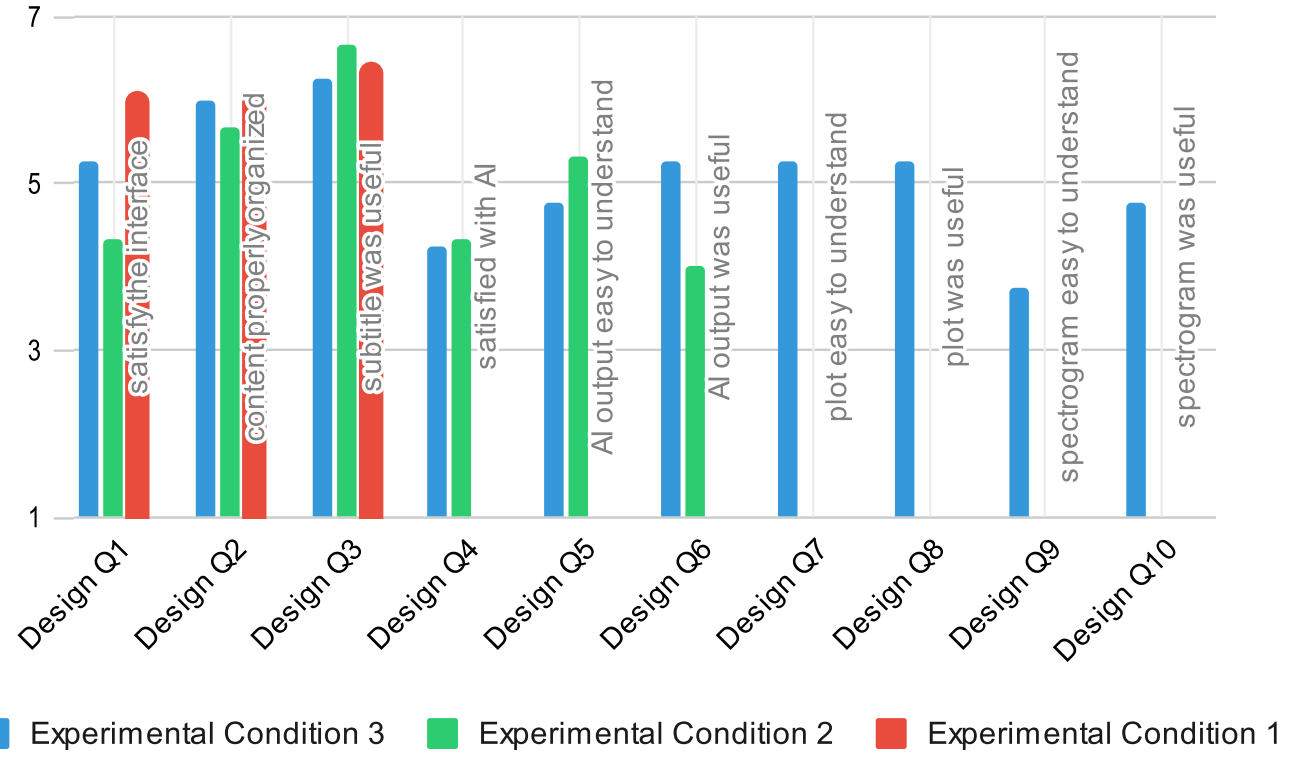
- Condition 1 (no AI) was perceived as the least complex(Q2)/cumbersome(Q8), easiest to use(Q3), and fastest to learn(Q7)
- Participants reported that they would use Condition 3 (AI w/ explanation) more frequently(Q1) than Condition 2 (AI w/o explanation), and that Condition 3 functions are better integrated(Q5) than Condition 2 functions
- Participants reported that Condition 3 was more complex(Q2)/cumbersome(Q8), slower to learn(Q7), and less consistent(Q6) than Condition 2



User study with mental health clinicians

System Design Survey

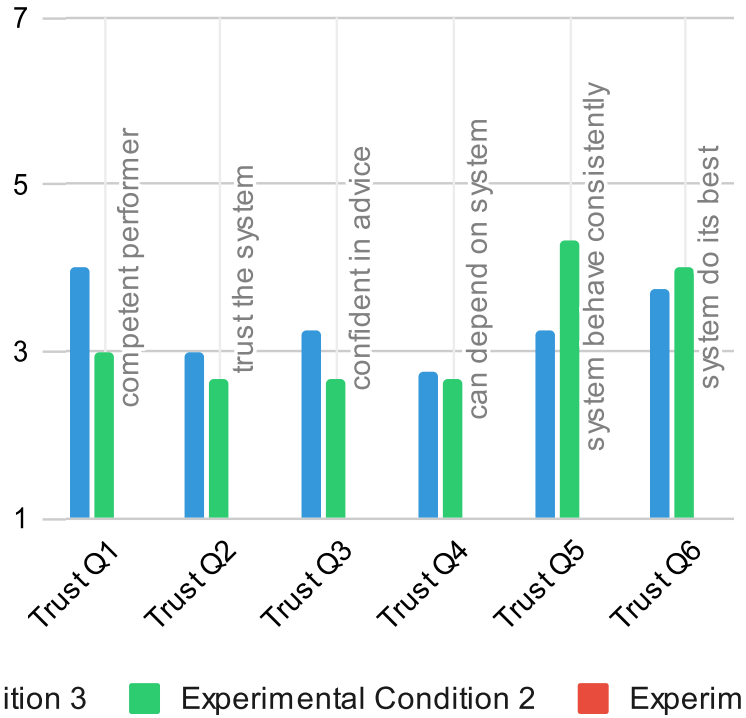
- Transcripts were rated useful(Q3) across all conditions
- The interface was rated the most satisfactory(Q1) in Condition 1, followed by Condition 3 and Condition 2
- Condition 2 (AI w/o explanation) output was easier to understand(Q5) than Condition 3 (AI w/ explanation)
- Condition 3 (AI w/ explanation) was rated more useful(Q6) than Condition 2 (I w/o explanation)
- Spectrogram function was overall difficult to understand(Q9), but was deemed useful(Q10)



User study with mental health clinicians

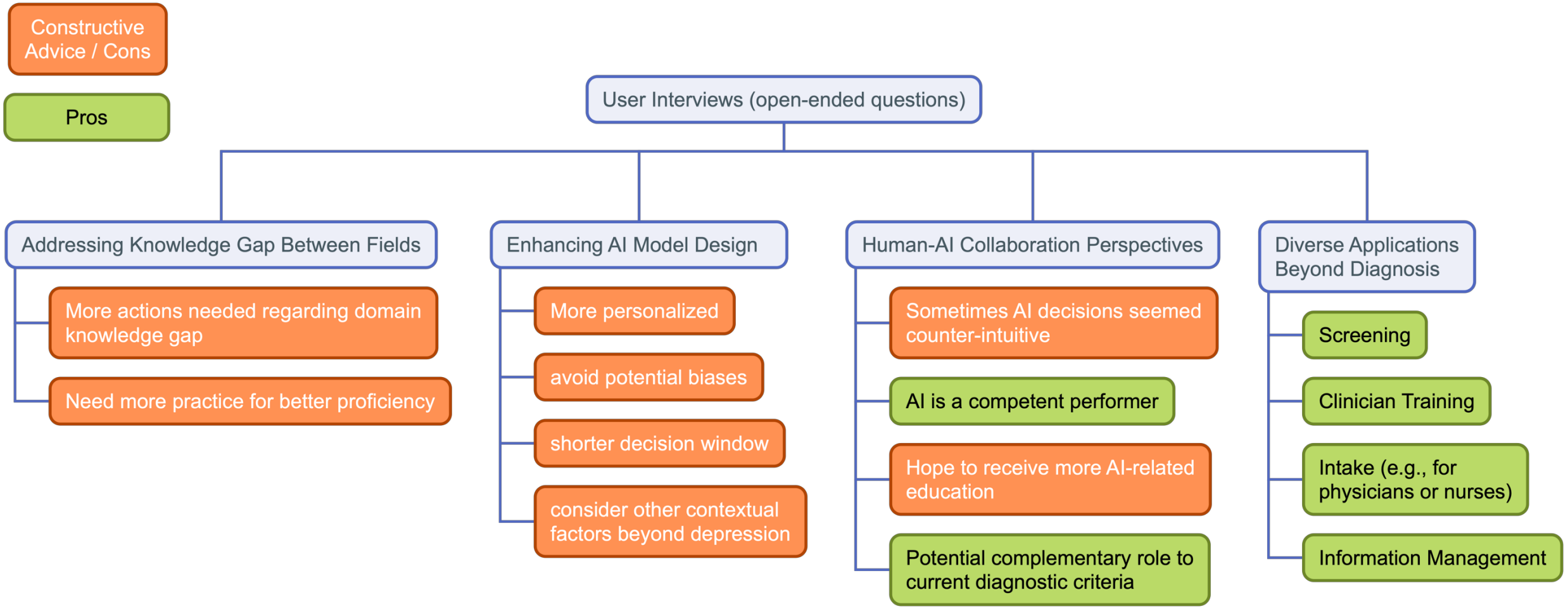
Merit Scale Survey

- Users assigned to Condition 3 (AI w/ explanation) depicted higher trust in the system(Q2) and were more confident in their decision(Q3) than the ones assigned to Condition 2 (AI w/o explanation)
- Users in Condition 2 (AI w/o explanation) perceived the system more consistent in its decisions(Q5) and reported that they can depend on it(Q4) to perform its best compared to users in Condition 3 (AI w/ explanation)



Clinician-AI collaboration to augment decision-making

Coding clinicians' open-ended responses



Clinician-AI collaboration to augment decision-making

Theme 1: Addressing the knowledge gap between the clinical field and AI systems

- Participants characterized the system as "advanced" and expressed a desire for more "laymen's terms"
- "I felt like you're spitting out new things in that (tutorial) video, you're providing new information to somebody who is not familiar with the realm of a spectrogram and these prediction models."
- "the definition (of spectrogram) is clear to me, but it was not useful because I would need more practice with the inner workings to truly comprehend what it is telling me."
- "I don't really like looking to the spectrogram. I feel like that one I need so much knowledge to understand that."

Theme 2: Adding more design features to the AI model

- Personalization: "So if people have a different cadence when they talk, it may be difficult to diagnose them until the AI learns their speech patterns."
- Regional biases related to accent and language use
- Shorter analysis windows, e.g., change rate to depression probability was perceived low
- Contextual information beyond acoustics

Clinician-AI collaboration to augment decision-making

Theme 3: Perspectives on human-AI collaboration

- AI performance: “when people are discussing something positive, and then at that point, the AI is indicating the depression probability is like 0.65, which is above 0.5”, “I feel like some part of the website is kind of accurate”, “AI is a competent performer”
- Supplementary role to diagnosis: AI can perceive acoustic details that are hard to capture by humans, “I think I would use it to supplement the diagnostic criteria based on what this person is telling me. I would use this system or this website as a quantitative supplement”
- Established evaluation : “Even though there's the prediction of this person probably having it (depression), I still would have to confirm, even if it's giving me this probability, I would still have to confirm based on content that meets DSM-5 criteria.”

Theme 4: Applications beyond diagnosis

- Screening: “It would be helpful in terms of screening and also assisting non-mental health professionals in screening, enabling them to refer individuals for follow-up. It would catch a lot of people that you were unsure of.”
- Patient pre/post treatment evaluation
- Training, particularly for inexperienced clinicians
- Information management: notes for reference

University of Colorado **Boulder**

Trustworthy human-centered machine intelligence for augmenting healthcare

HUBBS
HUman Bio-Behavioral Signals Lab

Summary of findings

Embedding expert knowledge at the local and global level in ML models can potentially improve the degree of explainability, while preserving the predictive performance of the models

It is important to consider new XAI paradigms, incorporate value-sensitive design and participatory design, adopt a holistic view of the problem, and evaluate with target users.

Discussion

University of Colorado **Boulder**

Trustworthy human-centered machine intelligence for augmenting healthcare

HUBBS
HUman Bio-Behavioral Signals Lab

Discussion

Privacy preservation should be the guiding principle when designing AI technologies for health. Yet, the inherent presence of PII in data can limit the effectiveness of widely used privacy mitigation techniques.

Despite the recent interest in equitable human-centered technologies, existing work is mostly developed in labs in isolation from the people and communities that they serve.

It is important to consider new XAI paradigms, incorporate value-sensitive design and participatory design, adopt a holistic view of the problem, and evaluate with target users.

It is necessary to start thinking on how to give human-centered technologies ethical principles and enable them to function in a responsible manner.

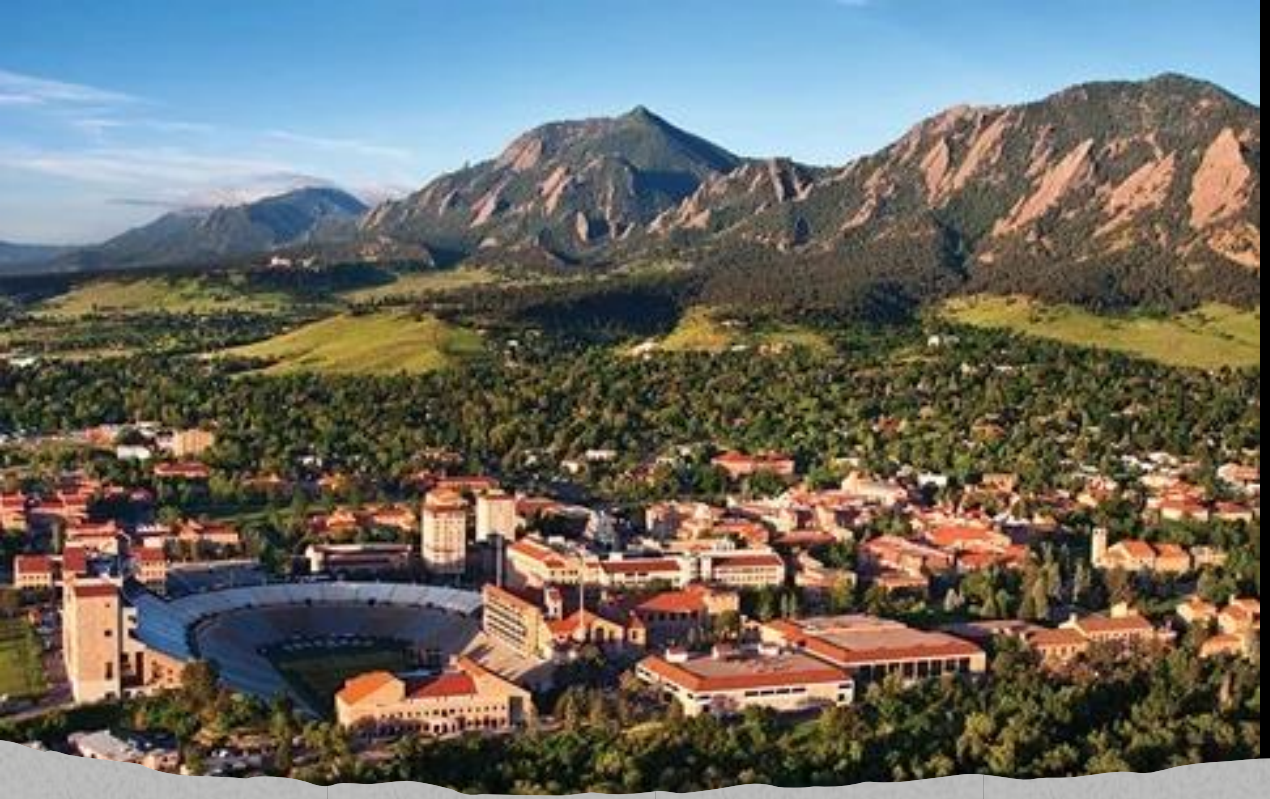
University of Colorado **Boulder**

HUBBS
Human Bio-Behavioral Signals Lab

Thank you

This work is supported by the National Science Foundation (CAREER: Enabling Trustworthy Speech Technologies for Mental Health Care: From Speech Anonymization to Fair Human-centered Machine Intelligence, #2046118, PI: Chaspari), the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA Contract No. 2017-17042800005 (Narayanan, PI), and the NIMH Grant No. R42MH123368 (PI: Timmons, Co-I: Comer). The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to re-produce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

NSF NIH NASA IARPA BSF gm W. M. KECK FOUNDATION



Institute of Cognitive Science (ICS) @ University of Colorado Boulder

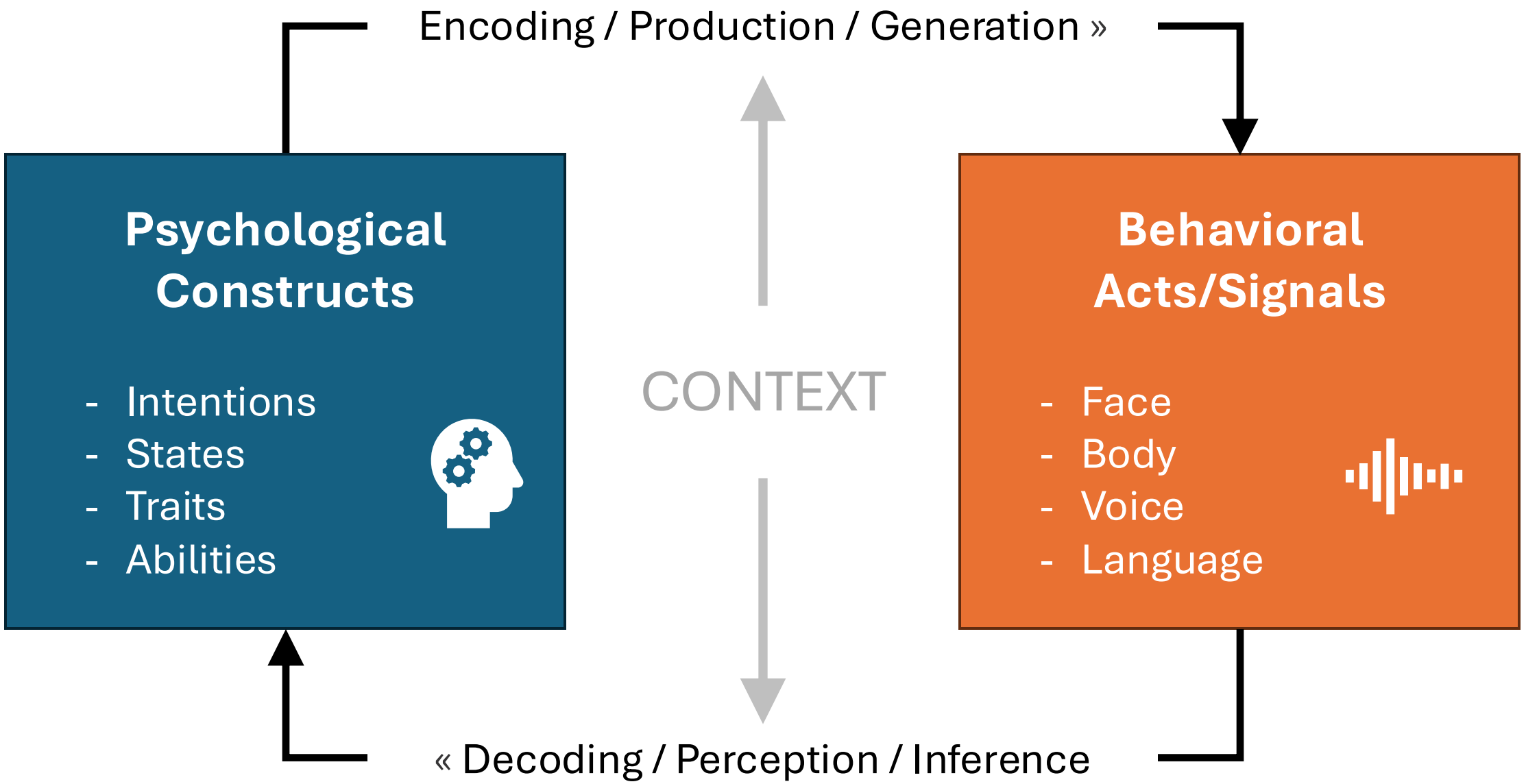
CONTEXTUALIZING SOCIAL AI

FOR GENDER, CULTURE, AND MORE

Dr. Jeffrey Girard, Psychology and Data Science, University of Kansas

Goals of the Presentation

1. Define key aspects of communication
2. Argue for the importance of context
3. Establish gender and culture as context
4. Define descriptive and prescriptive norms
5. Propose large-scale observational studies
6. Quantify descriptive smiling norms in two studies
7. Model norms across countries and genders
8. Discuss challenges, concerns, and open questions


Encoding / Production / Generation »
Psychological Constructs
- Intentions
- States
- Traits
- Abilities
CONTEXT
Behavioral Acts/Signals
- Face
- Body
- Voice
- Language
« Decoding / Perception / Inference

Social AI Needs Contextual Knowledge

- Improved understanding of situations
- Accurate construct perception/inference
- Nuanced behavior generation/synthesis
- Adaptability and flexibility to environment
- Increased authenticity and human-likeness
- Enhanced user experience

The encoding and decoding of a given behavior is influenced by its **surrounding context**. For example, the same behavior may differ depending on the **demographics of the person** producing it.

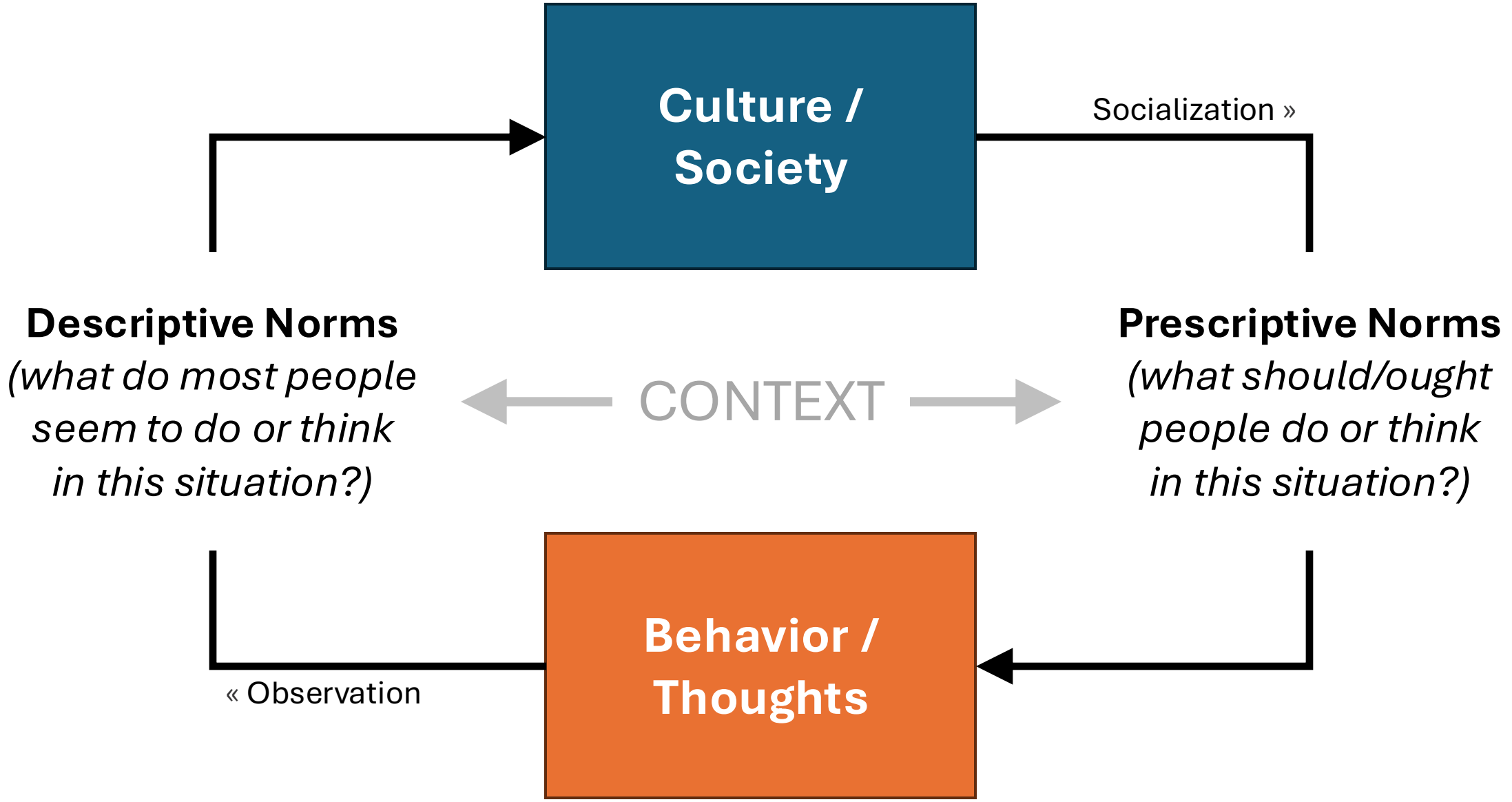
Demographics as Context

Coarse Categorizations

- Sex & Gender
- Age & Generation
- Race & Ethnicity
- Country & Culture
- Wealth & Social Status
- Education & Occupation
- Sexual Orientation & Identity
- Relationship & Parental Status
- Religion & Region & Politics
- Health & Disability

Presumed Mechanisms

- Hormones (testosterone, estrogen)
- Neurobiology (amygdala, PFC)
- Appraisal Patterns and Regulation Strategies
- Personality Traits and Motives/Goals
- Norms and Display Rules / Socialization
- Interpersonal Context / Dynamics
- Social Roles and Relationships
- Cultural Models of the Self
- Cultural and Religious Values
- Cultural Scripts and Rituals
- Historical and Societal Context
- Abilities and Disabilities


Culture / Society
Socialization »
Descriptive Norms
(what do most people seem to do or think in this situation?)
CONTEXT
Prescriptive Norms
(what should/ought people do or think in this situation?)
« Observation
Behavior / Thoughts

Social AI Needs Normative Knowledge

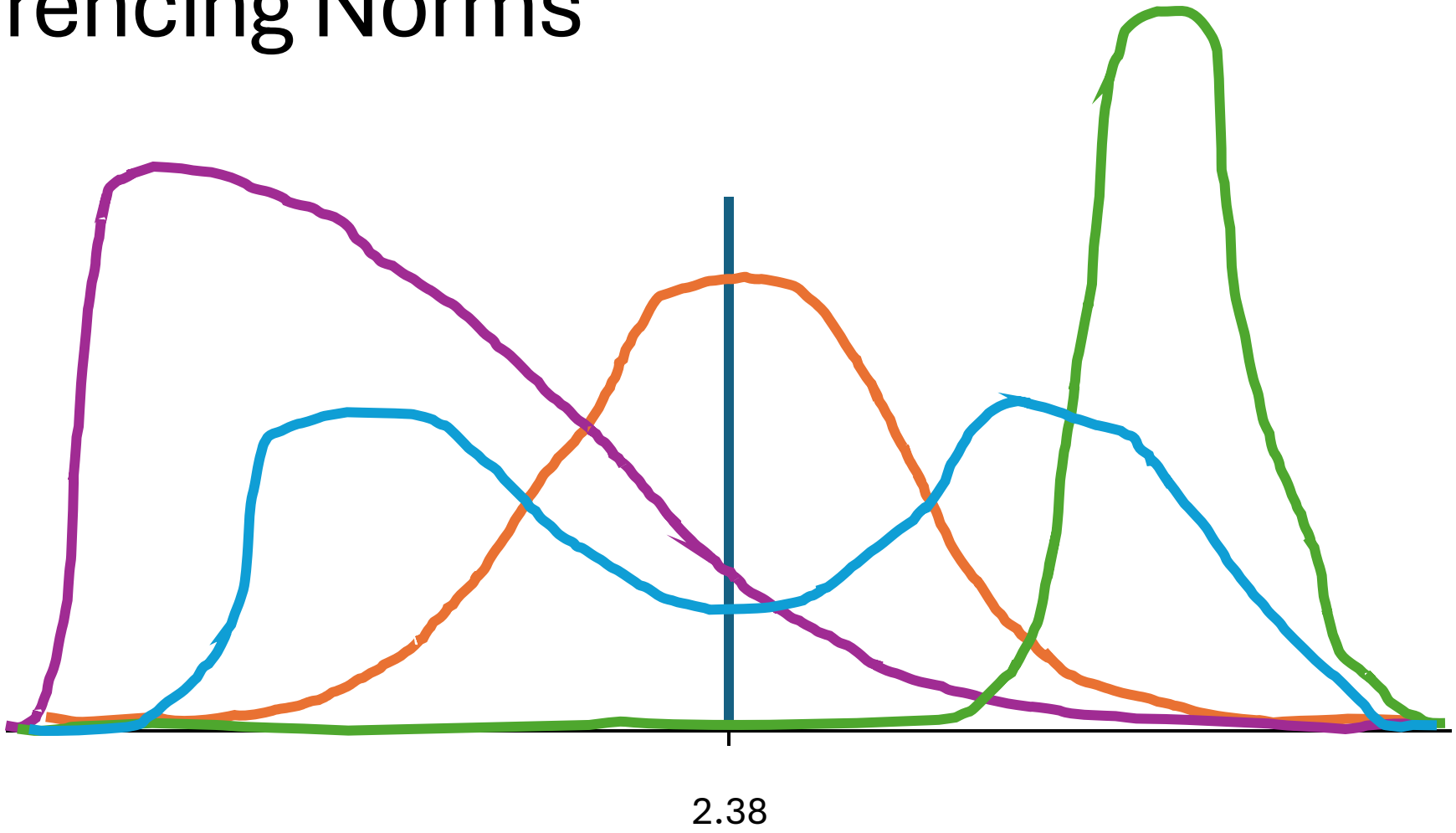
- Anticipating User Behavior
- Social Appropriateness
- Better User Interaction
- Adaptive Problem-Solving
- Cultural Sensitivity

- Avoiding Miscommunication
- Avoiding Confusion
- Avoiding Cultural Offense
- Avoiding Awkwardness
- Avoiding Uncanny Valley

The Importance of Norms

- How should our decoding/interpretation of behavior X change after learning that the person is part of a demographic group?
  - *What behaviors are normal or prescribed for this group in this situation?*
  - *How does the observed behavior compare to those expectations?*
- How should our encoding/generation of construct X into behavior change when synthesizing for a member of a different group?
  - *What behaviors are normal or prescribed for this group in this situation?*
  - *How will the generated behavior compare to those expectations?*
- How can we collect, quantify, and use normative knowledge?

Referencing Norms



Large-scale observational research

- Previous studies of descriptive norms relied on self-reports of what participants believe about normal and abnormal behavior
- But we also need to compare this to observational measures of how people actually behave, e.g., *empirical* descriptive norms
- But observational research is expensive and time-consuming
  - Previous studies typically collected observational records of dozens or hundreds of participants from around 2 to 4 convenient countries
- Data science and affective computing can help us scale up!

DATASET 1

ONLINE CELEBRITY IMAGES

Focus on celebrities as they are “trend setters” and public figures

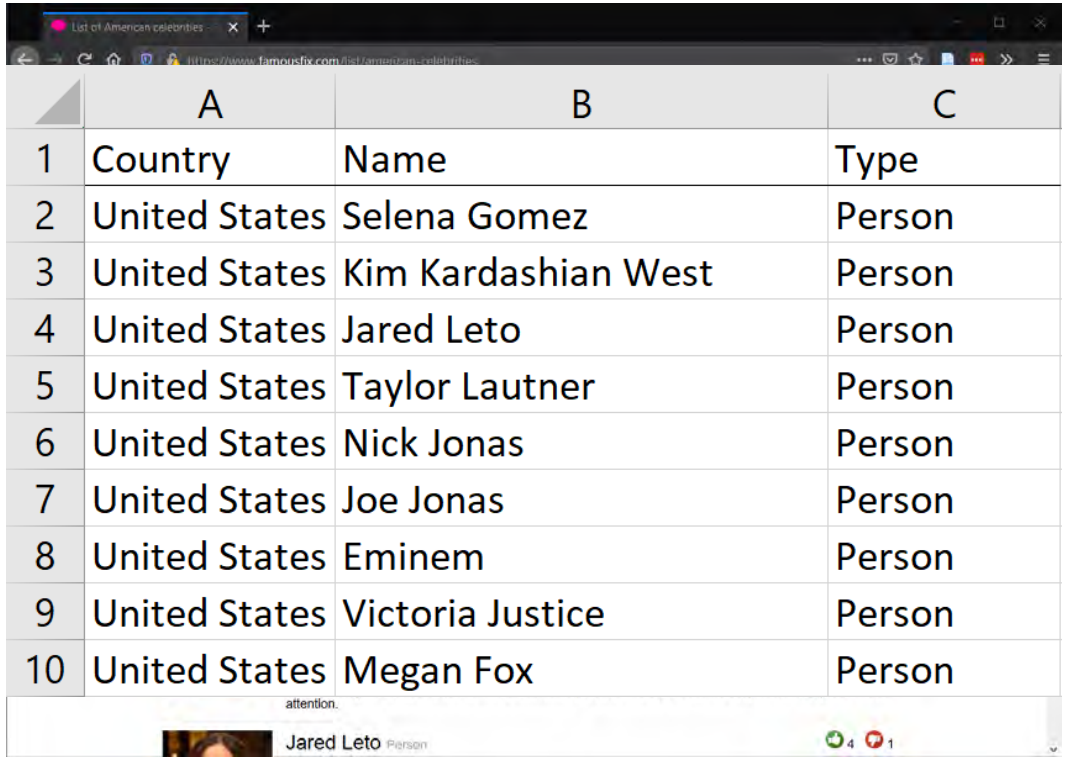
Attempt to maximize number of countries and individuals

Focus on smile intensity as an important social/affective signal

Data Collection (Text Scraping)

Goal: Construct a list of celebrities from as many countries as possible

1. Find list of celebrities by country
2. Loop through each country list
3. Parse HTML as structured text
4. Loop through top 1000 results
5. Extract country, name, type
6. Format into tidy data frame



	A	B	C
1	Country	Name	Type
2	United States	Selena Gomez	Person
3	United States	Kim Kardashian West	Person
4	United States	Jared Leto	Person
5	United States	Taylor Lautner	Person
6	United States	Nick Jonas	Person
7	United States	Joe Jonas	Person
8	United States	Eminem	Person
9	United States	Victoria Justice	Person
10	United States	Megan Fox	Person

Data Collection (Image Scraping)

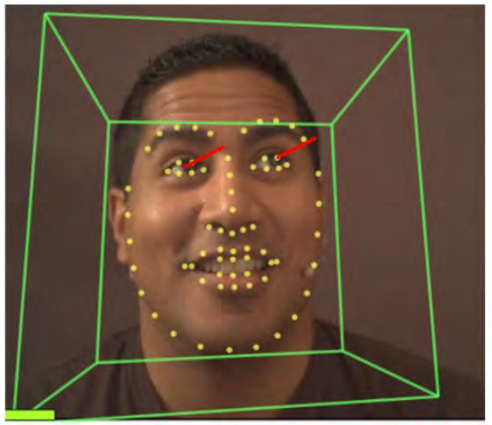
Goal: Search for and download photos of all celebrities on the list

1. Use Microsoft Bing Search API
2. Search for each celebrity name
3. Filter results for facial photos
4. Filter top 12 results per name
5. Extract and save image data

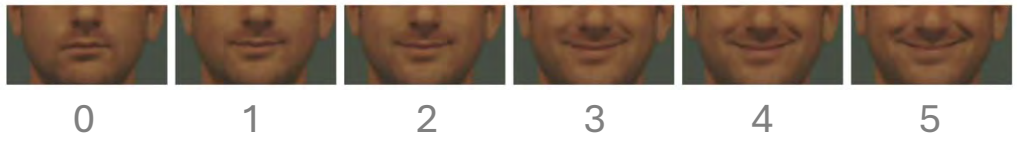
Data Processing (Smile Estimation)

- **Goal:** Estimate the absence or intensity of the smile on each face

1. Detect faces in each image
2. Locate facial landmark points
3. Extract visual features
4. Apply trained ML model
5. Predict smile intensity level



OpenFace Output Visualization



Smile Intensity Level Examples

Data Processing (Smile Validation)

Goal: Provide evidence that the smile intensity estimates are trustworthy

1. Select a subset of 300 images balanced by country and gender
2. Recruit 5 crowdworkers to rate images' positivity and smile
3. Recruit 1 expert FACS coder to code the smile (AU12) intensity
4. Compare human ratings to estimates from OpenFace

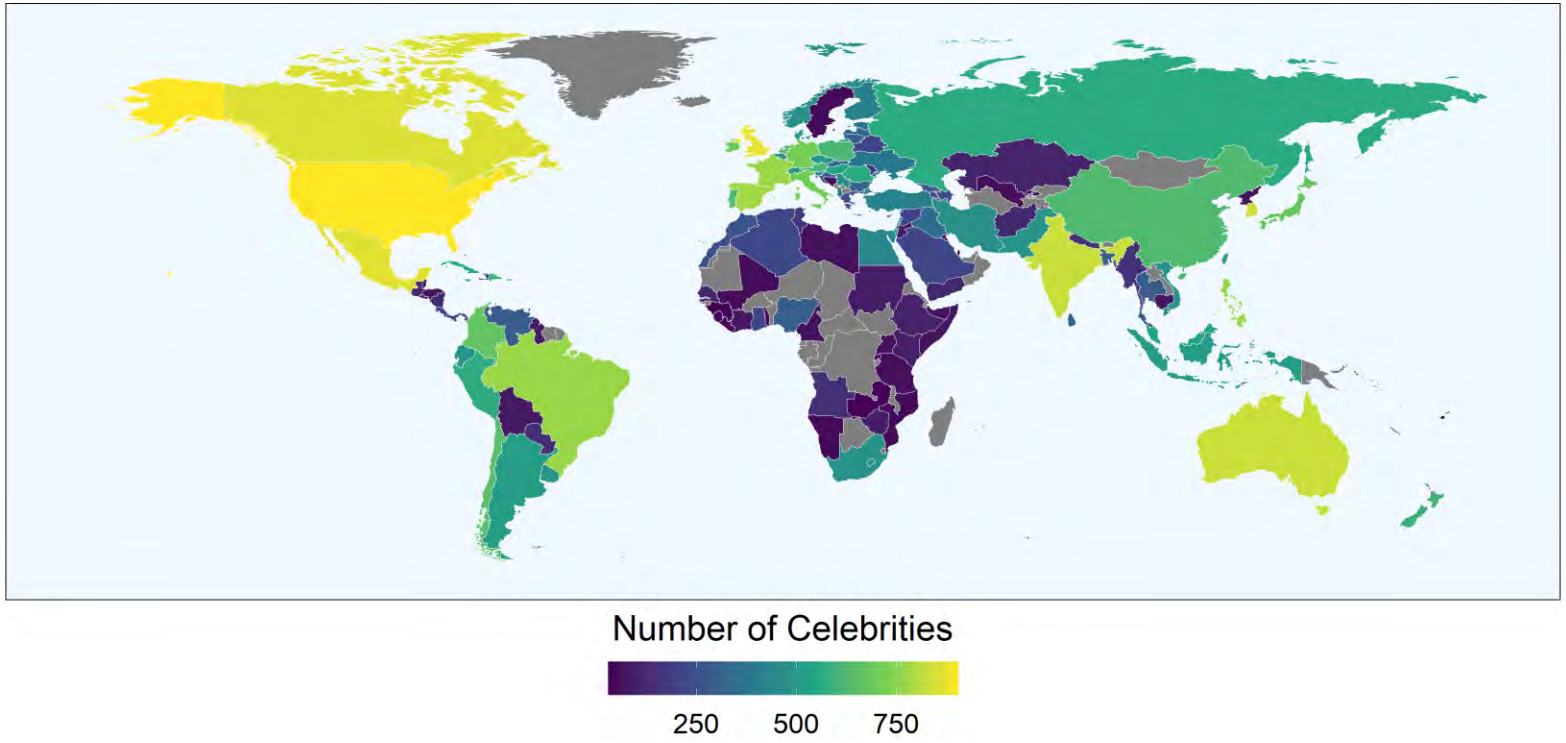
Inter-Rater Reliability Estimates

Measure	ICC(A,5)	95% CI
Positive Rating	0.90	[0.88, 0.92]
Smile Rating	0.90	[0.88, 0.92]

Heterogeneous Correlation Matrix

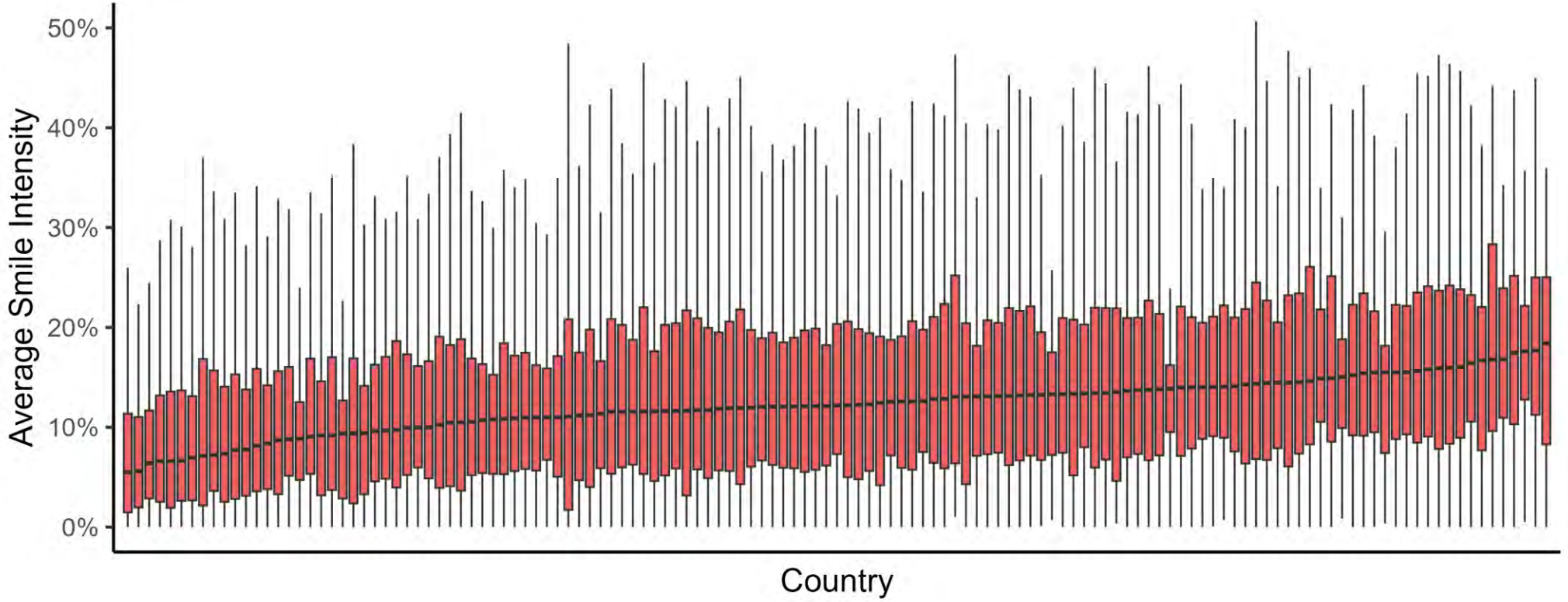
Measure	OpenFace	Positive	Smile
Positive Rating	0.79		
Smile Rating	0.78	0.94	
Expert FACS	0.87	0.97	0.94

Data Exploration (Counts by Country)



Final Count: 276,811 images; 44,602 celebrities; 133 countries

Country Distributions (Ignoring Gender)

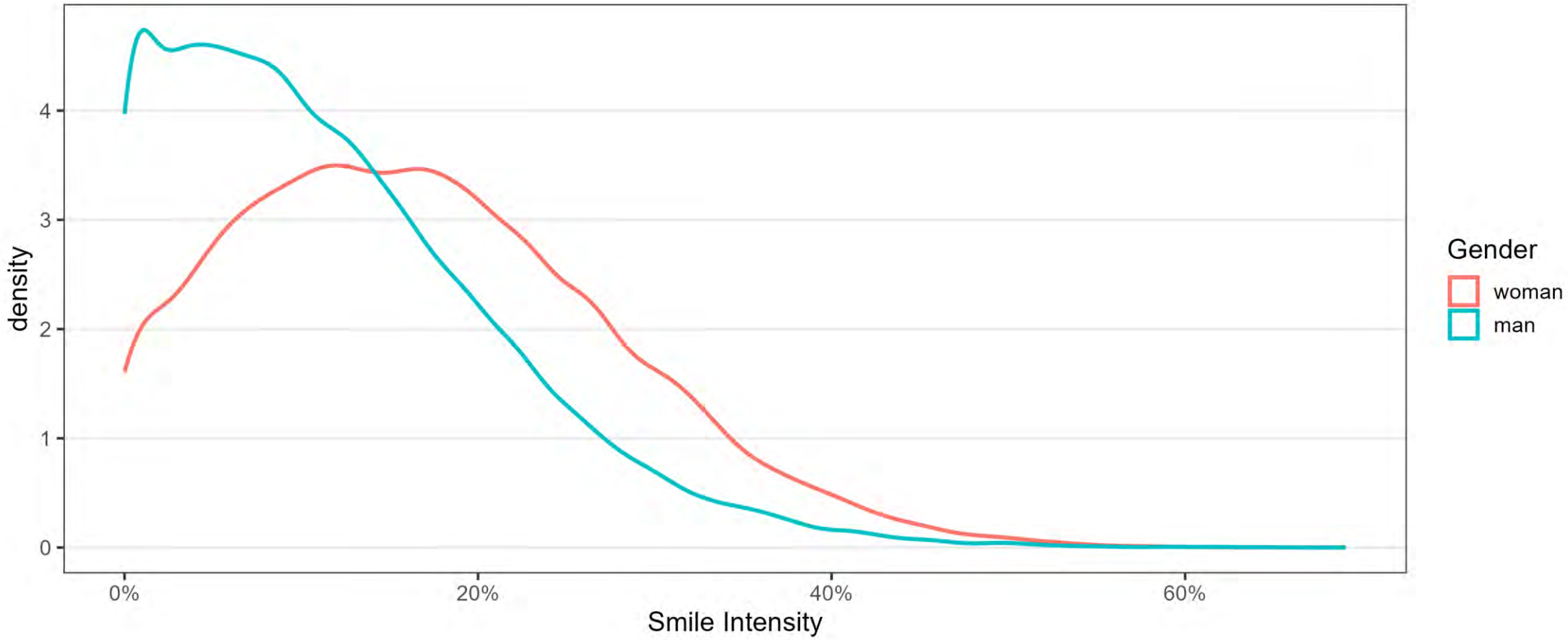


Low: Azerbaijan, Georgia

Mid: Ireland, Thailand

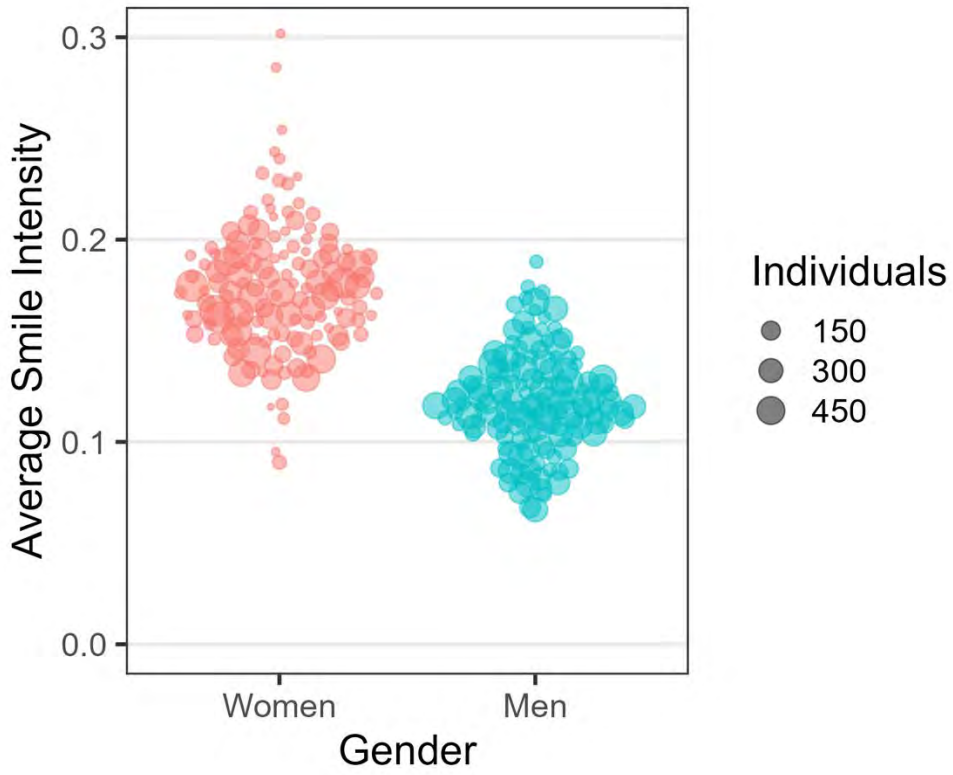
High: Venezuela, Philippines

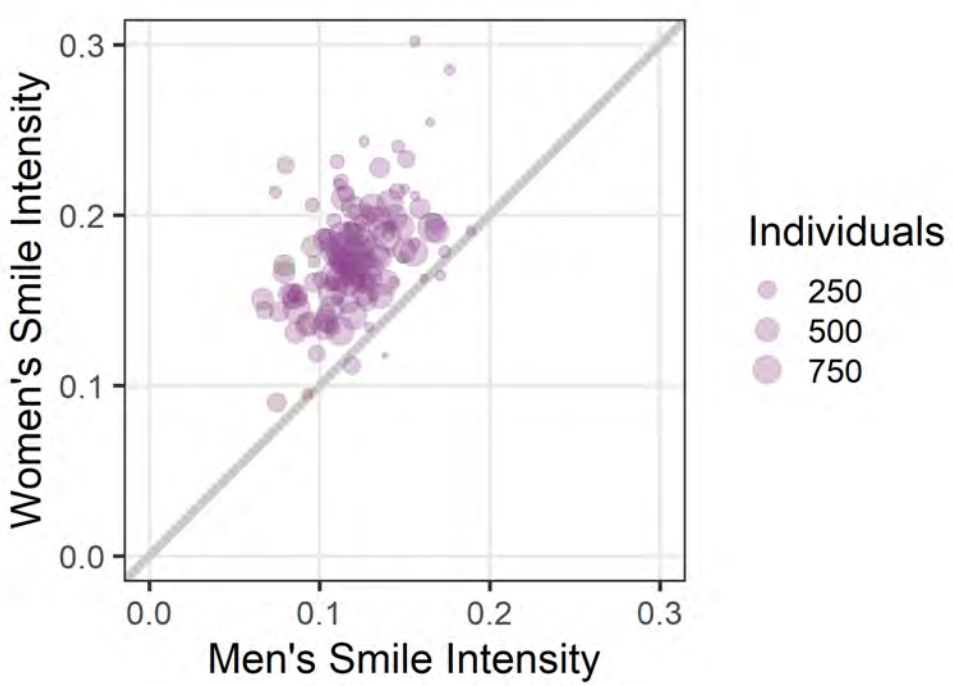
Gender Distributions (Ignoring Country)



Gender Distributions by Country

Note: In these visualizations, each bubble is one country (sized by the number of individuals in the sample).





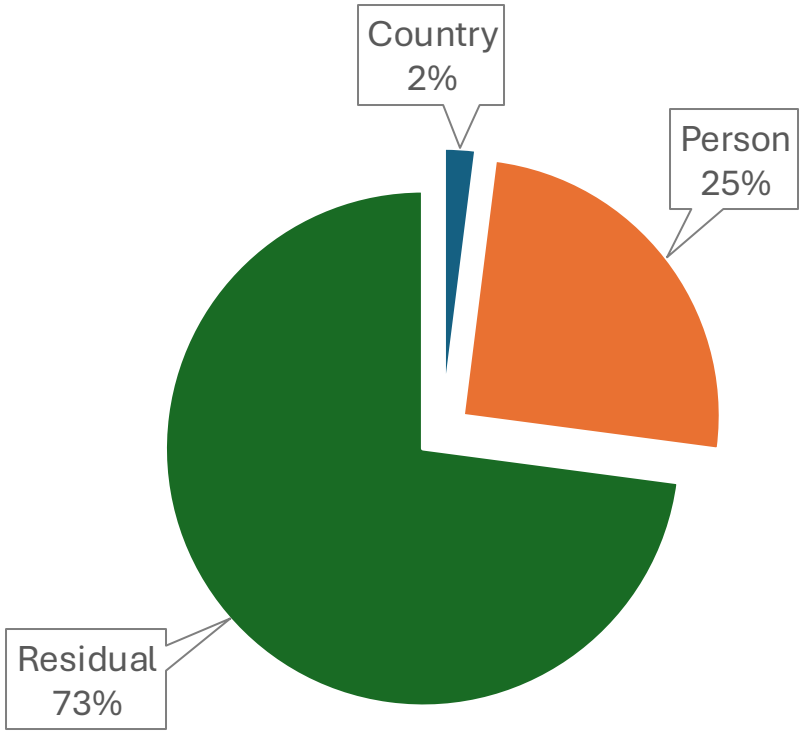
Variance Decomposition

$$\sigma_x^2 = \sigma_c^2 + \sigma_p^2 + \sigma_{i,e}^2$$

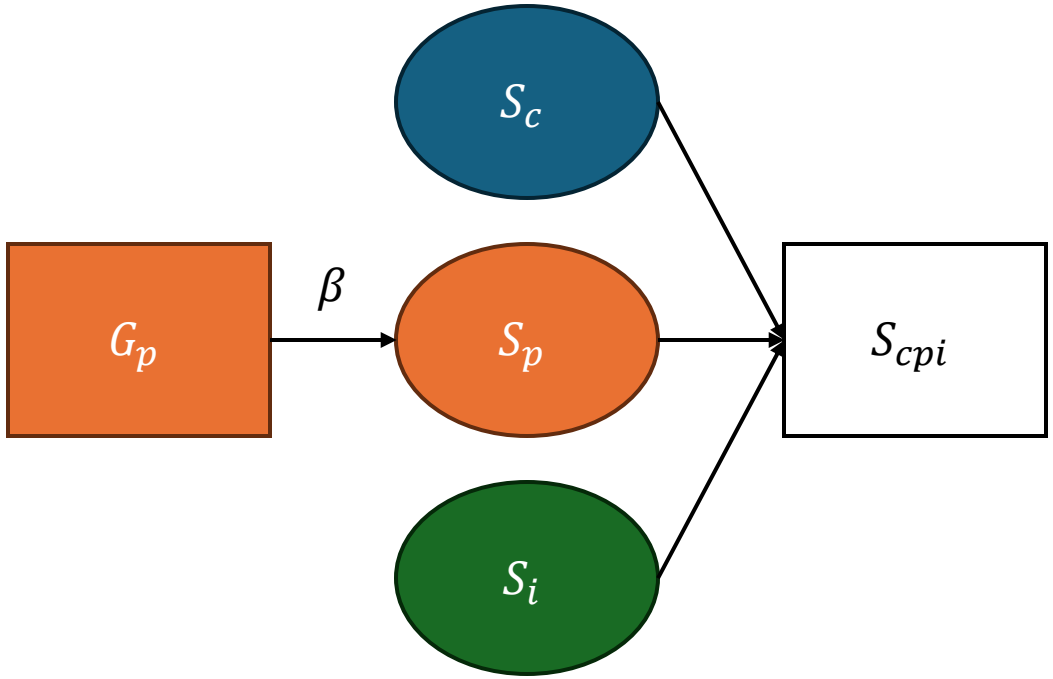
Most of the variance was residual (within-people, unexplained, error)

Countries explained very little variance, which means cultural influences on smiling are weak or "country" is too coarse a proxy

These results highlight the importance of person-level and image-level context



Variance Explained by Gender



$$\beta = 0.33 \quad p < .001$$

$$R^2 = .026$$

Women's average smile intensity was 0.33 SDs higher than Men's

Gender explained 2.6% of the total (and 10.4% of the person) variance.

DATASET 2

INSTAGRAM INFLUENCER IMAGES

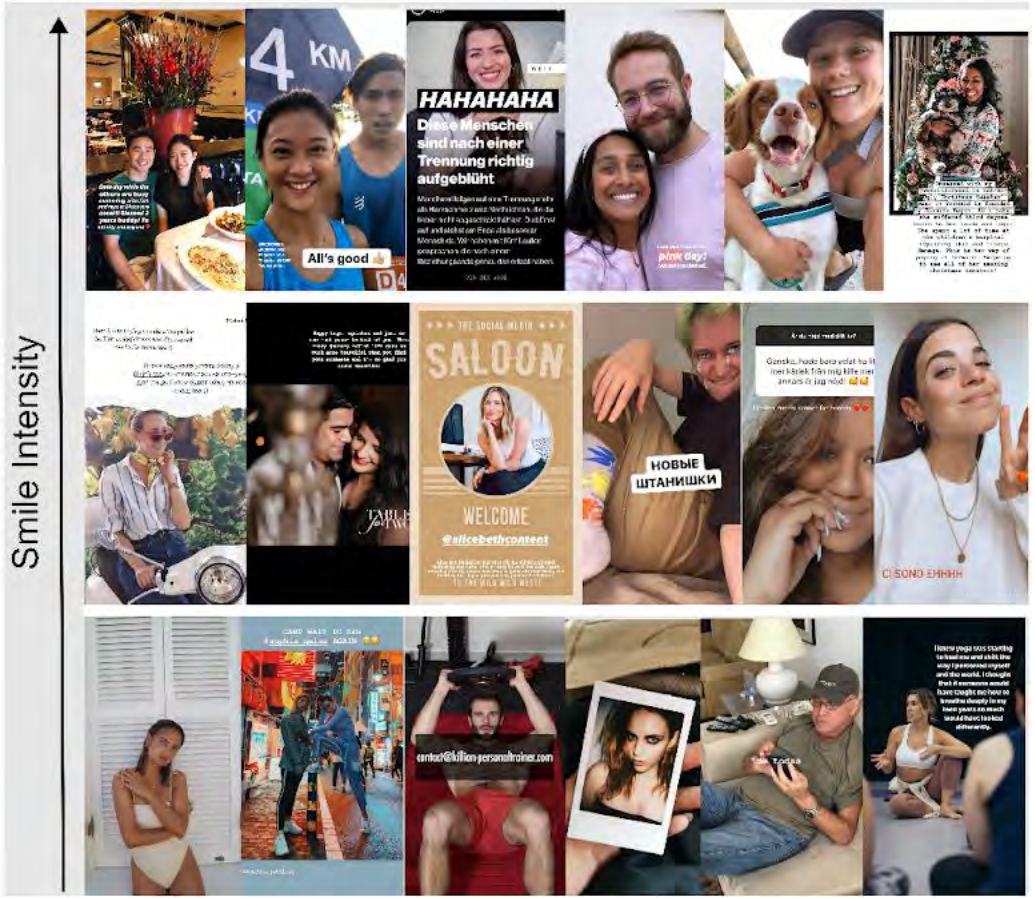
Follow individuals over time to assess within-person variability

Extend findings from celebrities to social media influencers

Extend images from headshots to more varied image types

Data Sourcing

- Partnered with an international influencer management agency
- Gained access to their internal database of Instagram influencers (with self-reported gender/country)
- Downloaded all images uploaded between May 2019 and Oct. 2021
- Images were analyzed for smile intensity using OpenFace
  (criterion validity of r = 0.41 with human ratings in a subsample of 595 images)



Data Exploration (Counts by Country)

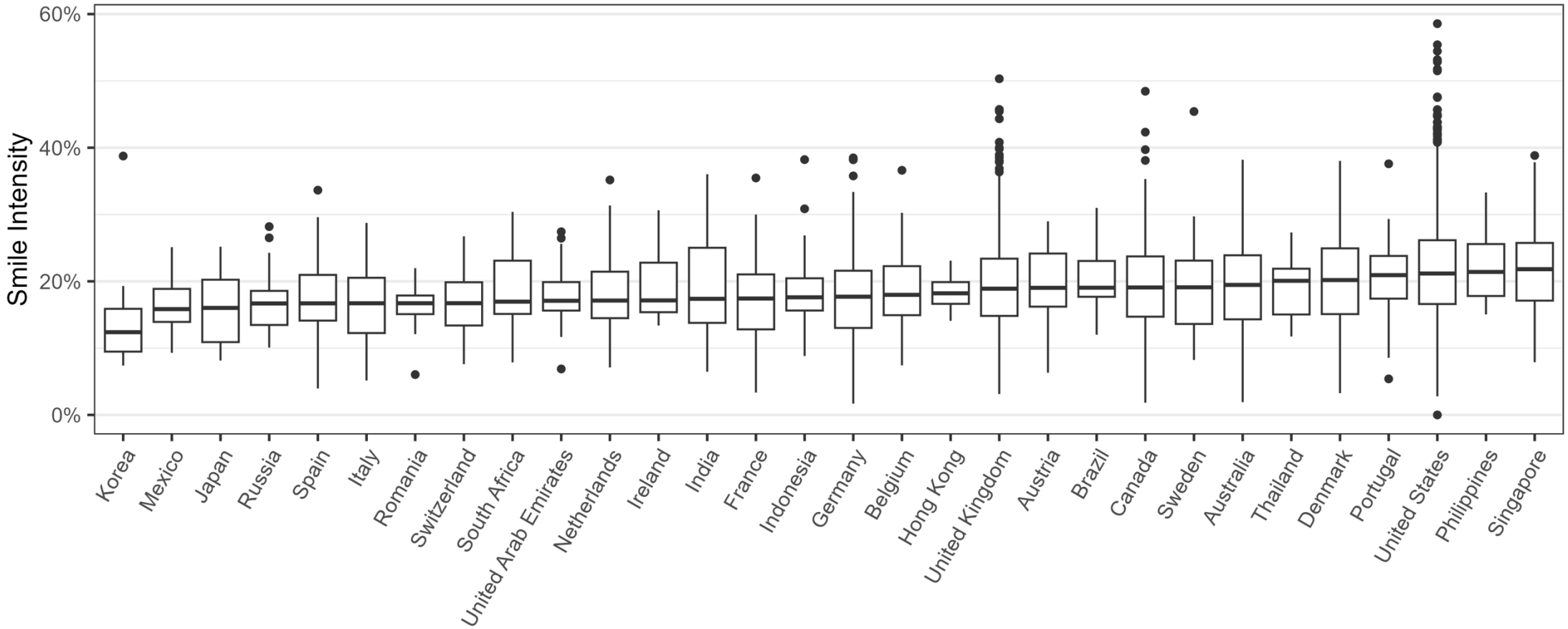
Country	People
United States	3096
United Kingdom	1536
Germany	199
Canada	151
France	144
Italy	100
Spain	87
Australia	74
Singapore	71
Netherlands	61

Country	People
Brazil	49
Sweden	45
Russia	40
India	33
Mexico	32
Philippines	32
Portugal	30
Belgium	26
Indonesia	23
Thailand	20

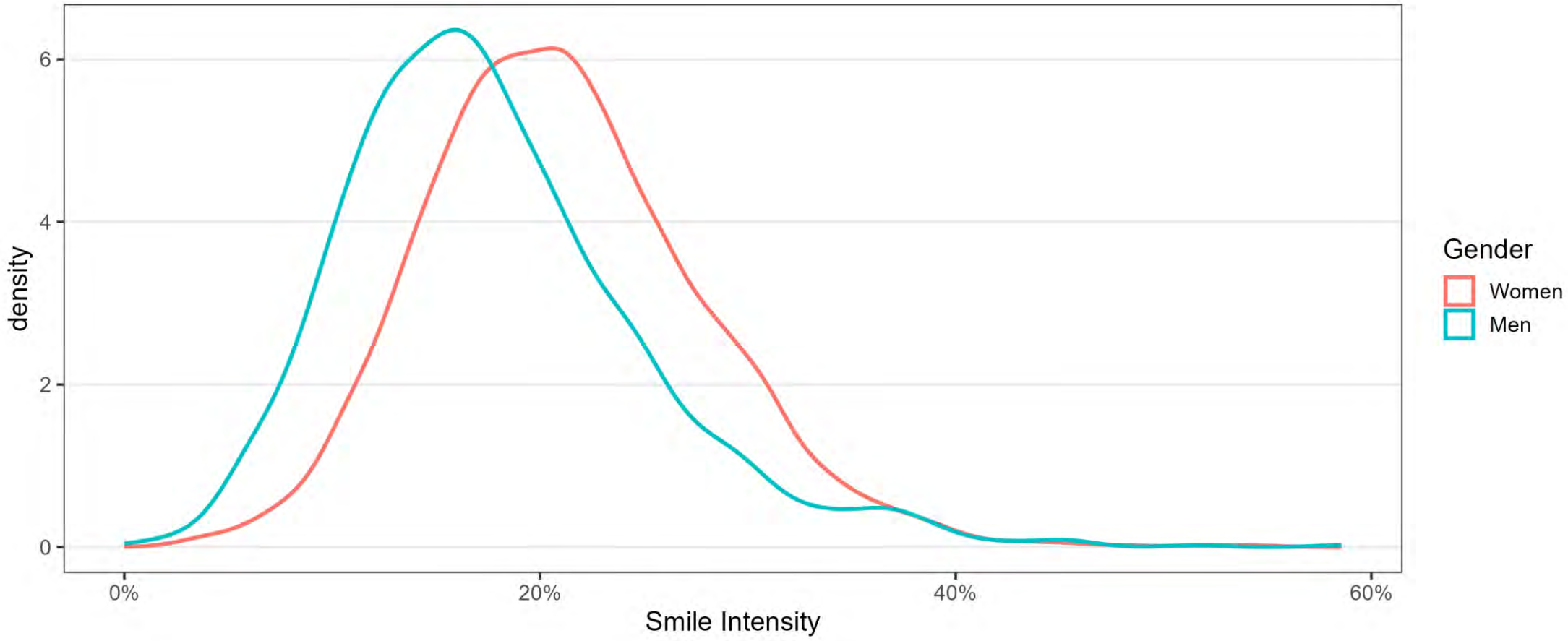
Country	People
South Africa	19
United Arab Em.	19
Ireland	19
Hong Kong	17
Japan	16
Romania	14
Denmark	14
Austria	13
Switzerland	12
Korea	12

Final Count: 1,930,376 images; 6,004 celebrities; 30 countries

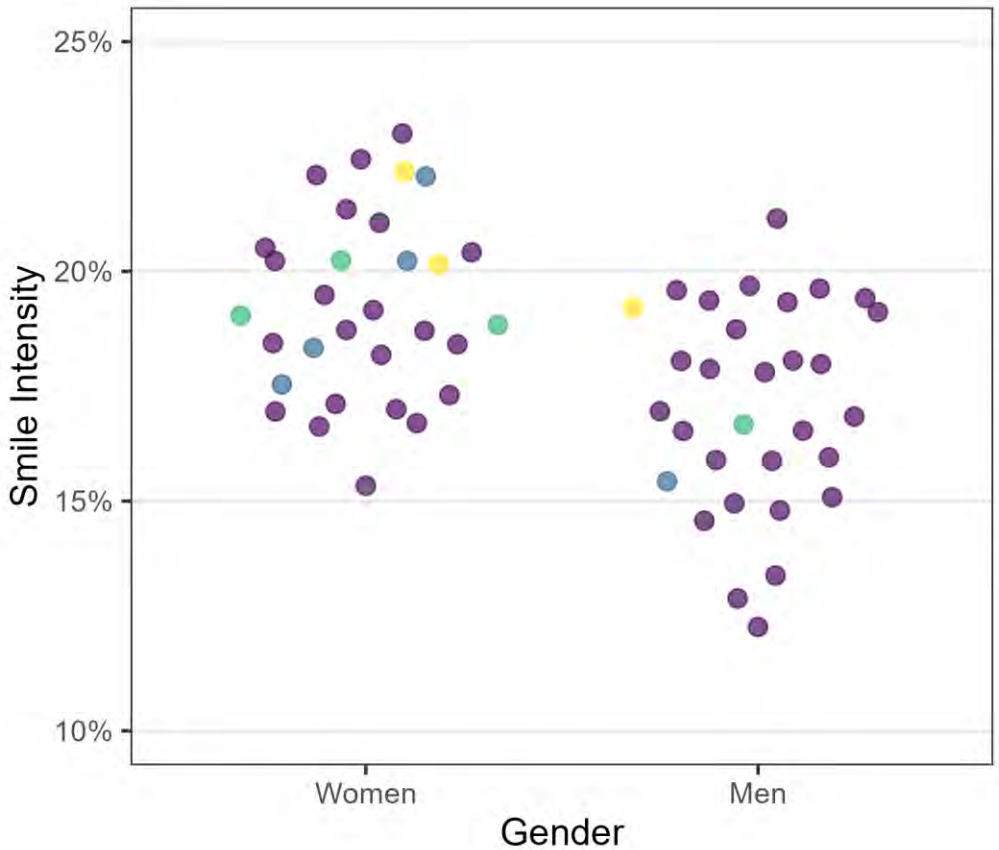
Country Distributions (Ignoring Gender)

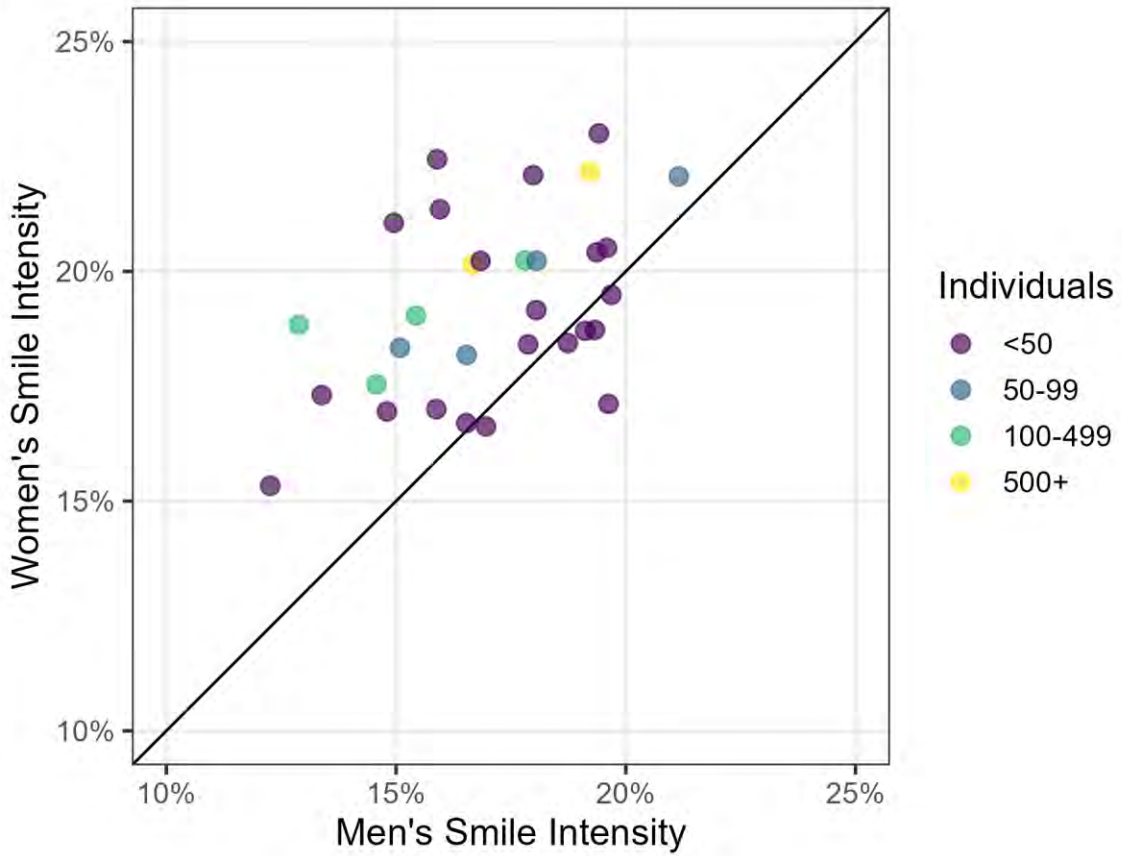


Gender Distributions (Ignoring Country)



Gender Distributions by Country





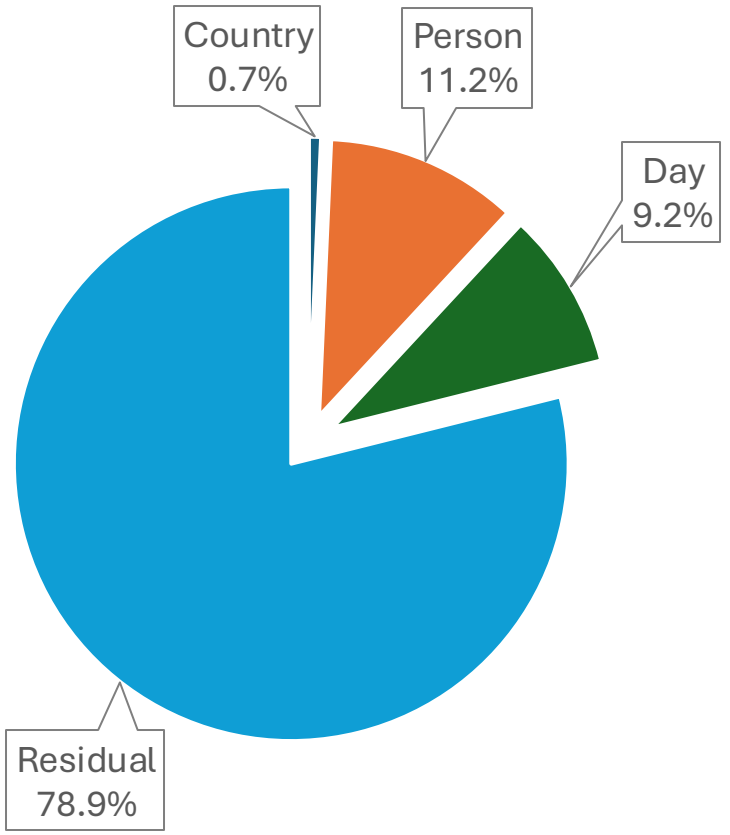
Variance Decomposition

$$\sigma_x^2 = \sigma_c^2 + \sigma_{p:c}^2 + \sigma_{d:p:c}^2 + \sigma_{i,e}^2$$

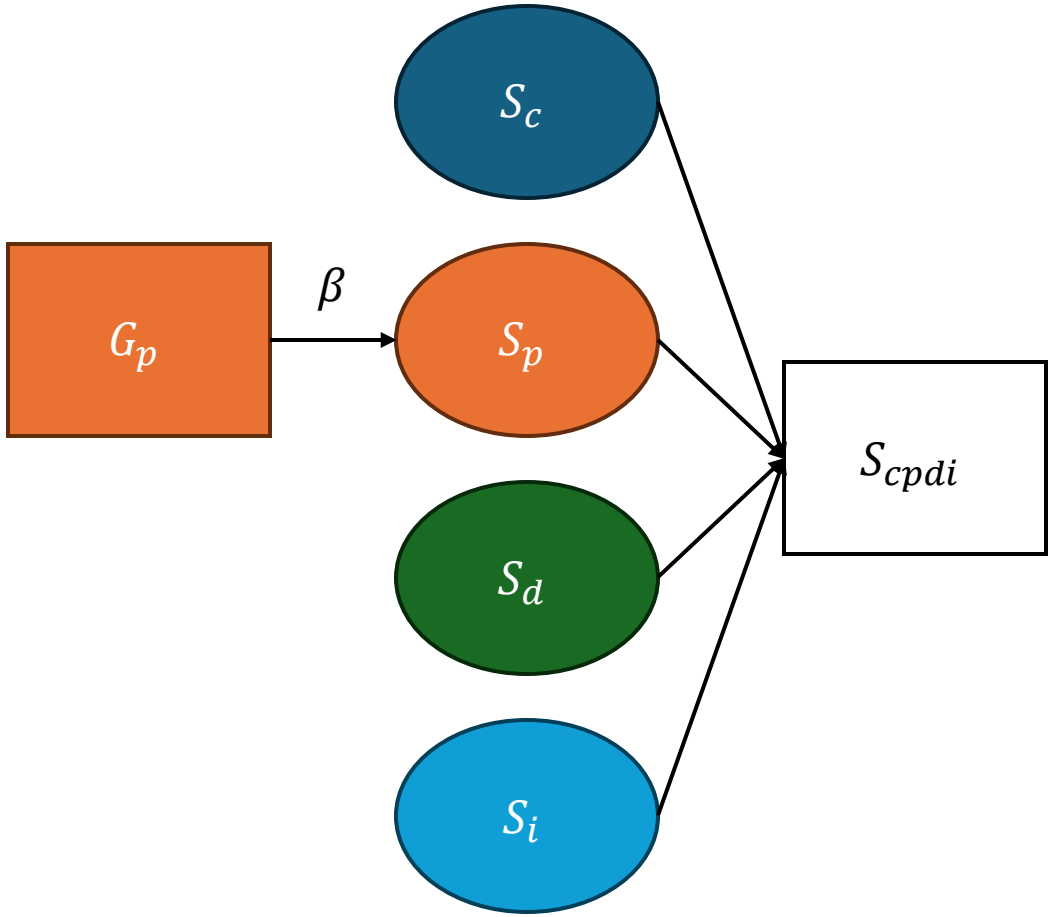
Most of the variance was residual (within-days, unexplained, error)

Countries and people explained less variance this time and day emerged as an important cluster

These results highlight the primary importance of *momentary context*



Variance Explained by Gender



$$\beta = 0.18 \quad p < .001$$

Women's average smile intensity was 0.18 SDs higher than Men's

$$R^2 = .004$$

Gender explained 0.4% of the total (and 3.6% of the person) variance.

CLOSING

Conclusion and Take-Aways

Open Challenges and Concerns

References and Discussion Questions

Conclusions

- Country explained disappointingly little of the total variance (<2%)
  - The impact of culture may be relatively weak on (posed) smiling
  - Or country may be too coarse a proxy to represent culture
- Individuals differed considerably within countries (11-25%)
  - Gender had a significant, global, and replicable effect on smiling
  - However, the vast majority of person variance and total variance remain
  - There is ample room for further research into individual differences
- Images differed substantially within individuals and within days
  - Local, momentary contextual features may be the most impactful

Challenges and Concerns

- Biases in Normative Data
  - Avoid reinforcing or reifying harmful or outdated stereotypes
- Privacy and Confidentiality
  - Personal data must be handled with care to protect privacy
- Transparency
  - Users should know when and why the system is adapting to norms
- Generalizability
  - Spontaneous behavior from non-celebrities may be quite different

References

- McDuff, D., Girard, J. M., & El Kaliouby, R. (2017). Large-scale observational evidence of cross-cultural differences in facial behavior. *Journal of Nonverbal Behavior*, *41*(1), 1–19.
- Girard, J. M., & McDuff, D. (2017). Historical heterogeneity predicts smiling: Evidence from large-scale observational analyses. *Proceedings of the 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, 719–726.
- McDuff, D., & Girard, J. M. (2019). Democratizing psychological insights from analysis of nonverbal behavior. *Proceedings of the 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, 220–226.
- Girard, J. M., El Kaliouby, R., Campbell, C., Rosengren, S., & McDuff, D. (in preparation). Quantifying descriptive smiling norms in multiple contexts around the world using large-scale observational methods.

affcom.ku.edu psych.ku.edu data.ku.edu

THANK YOU!
QUESTIONS?

Special Thanks to:

- Daniel McDuff
  (Google, Univ. Washington)
- Social AI Group
  (Univ. Glasgow)

Discussion Questions

1. What are some examples of prescriptive or descriptive norms related to emotion or interpersonal communication that you have noticed?
2. How can we most effectively build knowledge of prescriptive and descriptive norms into social artificial intelligence (SAI) systems?
3. What are some examples of situations in which we might want an SAI system to deliberately generate/produce “abnormal” behaviors?
4. How can we further improve large-scale observational research?
5. How can SAI systems handle bias, privacy, and complexity concerns?