Toward trustworthy human-centered machine intelligence

Signal Analysis and Interpretation Lab Team:
Shrikanth (Shri) Narayanan (PI)
Tiantian Feng, Rahul Sharma, Raghuveer Peri, Amrutha Nadarajan, Rajat Hebbar

Amazon Mentors/POC:
Rahul Gupta, Anil Ramakrishna

USC-Amzon Center Kickoff Meeting, Sept/2021
Goals
- Understand the human condition: traits, state, behavior, interaction
- Support and enhance human experiences

**Human centered view**: characterizing data about, from and for people
- includes knowledge about how people perceive, process and use (human) data
  - constructs capturing human expression & experience
  - constructs characterizing human “perception”
  - some constructs (e.g., “labels”) may be available explicitly, others have to be (self) learned implicitly
  - other constructs may not be human “label-able”
Elements of Trustworthy Human-centered Machine Intelligence

- INCLUSION
- PRIVACY
- SAFETY

- e.g., safeguard protected information
- e.g., protect against malicious attacks
- e.g., enable broad access

A human conversation example: rich verbal and nonverbal behavior and interaction

Speech and language provide access to assessing intent, emotions, and a variety of information about personal demographic traits (age, gender,…), physical/psychological/health state, and interaction context. These attributes/constructs are often intricately related.

Other biobehavioral data streams share similar “loaded” characteristics: e.g., ECG
Tight links between generation and processing of signals by human system and their interaction with the world
- the system is characterized by traits/states/behavior expression
- in turn the human perception/experience affects the system and shapes future behavior/actions
How about supporting interaction with/through an AI agent?

- Relies on knowledge regarding aspects of individual traits, state, behavior for providing the desired user experience
  - while lots of information is available-and possibly retrievable-from human signals not all of it is essential or should be used

**For example:**
- a specific use case may rely on *what* was said but need not know *who* said it, or what their *affective* state was
- another use case may rely on *age* information and aspects of *health state* but doesn’t need to track the specific *linguistic content* of an interaction
Example human-centric applications at SAIL

Large-scale workplace behavioral study

Psychotherapy Interaction
Child-inclusive interaction

Clinically-relevant features
❖ Empathy
❖ Entrainment/
    Synchrony
❖ Visual Gaze
❖ Emotion-state

Sensitive features
❖ Gender
❖ Age
❖ Ethnicity
❖ Language content

**TILES study: different rooms in the hospital,**
different privacy constraints  https://tiles-data.isi.edu/

**CARE/DEPTH: Behavioral Modeling in Human Interactions**
https://sail.usc.edu/care
Elements of Trustworthy Human-centered Machine Intelligence

- **Inclusion**: e.g., enable broad access in the presence of individual heterogeneity
- **Privacy**: e.g., safeguard desired information to protect (PII: age, gender, race, etc.)
- **Safety**: e.g., protect against malicious attacks of vulnerable protected variables

---

- [Link](https://design.google/library/designing-global-accessibility-part-1/)
Our initial efforts in human-centric trustworthy AI:

1. **Privacy**
   a. Sensitive attribute obfuscation for speech emotion recognition
   b. Federated human activity detection from wearable sensors

2. **Inclusion**
   a. Fairness in automatic speaker verification

3. **Safety**
   a. Exploring strategies for defense against adversarial attacks in speaker recognition

**Conclusion:** Summary of active/ongoing research threads and milestones
Elements of Trustworthy AI: PRIVACY
Privacy in human centric applications

Privacy is personal and context dependent

Mode of acquisition

Primary Task

Sensitive Attributes

Speech Signal

Speaker ID

Language content, Location, Emotional state, etc.

Speech Recognition

Demographics, Gender, Accent etc.
Privacy in human centric applications

Newer ways of sensing - e.g., wearable sensors, IoT devices - have varying privacy demands.

Egocentric sensing for stress regulation

Mode of acquisition

Primary Task
Detect foreground speech

Common sensitive attributes
location, demographics, gender

Speech Signal

Detect desired affect state

Task-dependent sensitive attributes
Specific emotional states

Language content
Attribute inference in Speech Emotion Recognition (SER)

- Required application
  - Demographic
  - Sensitive Emotions
- Speech signal
  - Speech feature
    - Feature Extraction
    - Feature Transformation
  - Query: Is the person sad?
  - Model
    - Yes/No
  - Adversary
    - Model
      - Demographic
Sensitive emotion obfuscation in SER

Replacement Autoencoder

Speech Feature For an Utterance

X

Enc1

Enc2

Embedding

Dec1

Dec2

X’

Transformed Feature For an Utterance

Sensitive emotion obfuscation in SER

\[ U (\text{Utility score}) = \frac{\text{Correct Predictions Using Both } X \text{ and } X'}{\text{Correct Predictions Using Only } X} \]

<table>
<thead>
<tr>
<th>Set of Inference</th>
<th>X F1</th>
<th>X' F1</th>
<th>U</th>
</tr>
</thead>
<tbody>
<tr>
<td>(N = {\text{neutral}})</td>
<td>56.0%</td>
<td>43.0%</td>
<td>81.0%</td>
</tr>
<tr>
<td>(R = {\text{sad}})</td>
<td>64.8%</td>
<td>61.0%</td>
<td>82.3%</td>
</tr>
<tr>
<td>(S = {\text{happy, angry}})</td>
<td>78.5%</td>
<td>3.6%</td>
<td>1.8%</td>
</tr>
</tbody>
</table>

Required application

- Sad
- Happy
- Angry

Sensitive emotions

- Happy
- Angry
- Neutral

Feng & Narayanan, Privacy and Utility Preserving Data Transformation for Speech Emotion Recognition, ACII 2021
Gender obfuscation in SER

Required Application

<table>
<thead>
<tr>
<th></th>
<th>X</th>
<th>X’</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1</td>
<td>64.8%</td>
<td>58.4%</td>
</tr>
<tr>
<td>Sad</td>
<td>😞</td>
<td>✓</td>
</tr>
</tbody>
</table>

Gender

Idea: Minimize mutual information between gender and transformed data

Adversarial gradients

Gradient Reversal Layer

Gender Inference

<table>
<thead>
<tr>
<th></th>
<th>X</th>
<th>X’</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>69.7%</td>
<td>55.9%</td>
</tr>
<tr>
<td>Random guess: 52.1%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gender</td>
<td>♂ ♂</td>
<td>❌</td>
</tr>
</tbody>
</table>
A multimodal human subject study on the clinical population
  ○ Understand workplace stressors
  ○ How do they affect wellbeing and productivity

Nurse population
  ○ High burnout, high stress population
  ○ Work long shifts

Passive egocentric sensing using portable, lightweight sensors

Study easy to run, replicate

Fig (Clockwise) : ECG shirt, TILES study app, location sensors, humidity/temperature sensor, Fitbit and TILES Audio Recorder
Federated human activity detection

Wearable Sensor Data ➔ Data Mining ➔ Application Model

Central Server

Heart rate time series

Motif Extraction (Motif: repetitive pattern)

Wake up

Sleep

......
Federated human activity detection

Motif Data → Application Model

Application Model → Adversary Model

Adversary Model → Demographics, Identity

Sleep, Wake up

Trustworthy human-centered machine intelligence
Federated human activity detection

- Model trained on local data at each individual edge device.
- Model updates are transferred to the global server.
- Federated models (FedAVG) perform comparable to centralized model while preserving personal information.

<table>
<thead>
<tr>
<th>Training setup</th>
<th>f1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Centralized training</td>
<td>0.81</td>
</tr>
<tr>
<td>Federated training</td>
<td>0.80</td>
</tr>
</tbody>
</table>
Elements of Trustworthy AI: INCLUSION
Fairness and Inclusion

Artificial data-balancing or training on class-balanced data is seldom enough to ensure fairness and inclusion in learning.

Fig 1: Outputs from an (withdrawn) object recognizer that labels same object differently in the context of skin tone

Fig 2: She vs he occupations output from a word embedding system

1. homemaker
2. nurse
3. receptionist
4. librarian
5. socialite
6. hairdresser
7. nanny
8. bookkeeper
9. stylist
10. housekeeper
11. interior designer
12. guidance counselor

Extreme she occupations

1. maestro
2. skipper
3. protege
4. philosopher
5. captain
6. architect
7. financier
8. warrior
9. broadcaster
10. magician
11. fighter pilot
12. boss

https://algorithmwatch.org/en/google-vision-racism/
https://arxiv.org/abs/1607.06520
Towards inclusive speech-centric interaction technologies

Speaker Verification

Applications
- Access control
- Audio indexing
- Transaction authentication

Alex’s enrolment utterance
I'm Alex
I'm Alex

Trustworthy human-centered machine intelligence.key - September 28, 2021
Genuine: Test speaker same as claimed

Impostor: Test speaker different from claimed

- Similarity scores between enrolment and test utterance embeddings
- State-of-the-art speaker embeddings: x-vectors
- Models trained using gender-balanced data
- Evident skew in scores between male and female populations

Data balancing alone does not ensure fairness in ASV
Improving fairness using Adversarial Training

- Required attribute: Speaker id
- Sensitive attribute: Gender

Goal

Learn speaker discriminative representations while discarding gender information
Qualitative results

Quantitative results (lower is better)

<table>
<thead>
<tr>
<th></th>
<th>Male</th>
<th>Female</th>
<th>Delta</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>% EER</td>
<td>6.50</td>
<td>4.80</td>
<td>1.70</td>
<td>5.82</td>
</tr>
<tr>
<td>x-vector</td>
<td>6.60</td>
<td>5.37</td>
<td>1.23</td>
<td>6.05</td>
</tr>
</tbody>
</table>

- Adversarial training mitigates skew in similarity scores
- Difference (delta) in EER between Male and Female populations reduced with Adversarial training

Future research directions

- Intersectionality: Age, Gender, Accent etc.
- Holistic evaluation metrics

Adversarial training succeeds in reducing gender bias in ASV
Elements of Trustworthy AI: SAFETY
Security risks in Machine Learning Systems

Train stage

- **Data poisoning**
  - Manipulate training data to learn incorrect correspondences
  - Can compromise privacy and confidentiality

Inference stage

- **Adversarial attack**
  - Add malicious noise to input
  - Force pre-trained system to make incorrect predictions
  - Difficult to detect

- **Output attack**
  - Attacks result display setup
  - Manipulate output prediction presentations
Security of Speaker Recognition systems

Adversarial attacks

- Malicious attacker crafts noise to modify DNN system predictions
- Added noise can be imperceptible
- Attacker can force classifier to output any desired incorrect prediction
- Compromise security of any DNN-based system
- Impacts trustworthiness

DNN-based systems (including speaker recognition) are vulnerable to adversarial attacks

FGSM attack

- Uses sign of gradient to push prediction away from true output

Defense against adversarial attacks

- Use adversarially added noise during speaker recognition model training
- On-the-fly data augmentation

Speaker recognition performance

<table>
<thead>
<tr>
<th></th>
<th>Benign</th>
<th>FGSM* attack</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without defense</td>
<td>94.0</td>
<td>25.0</td>
</tr>
<tr>
<td>With defense</td>
<td>92.0</td>
<td>73.0</td>
</tr>
</tbody>
</table>

FGSM: Fast Gradient Sign Method

Defense significantly improves performance under attack

Minimal impact on benign performance


### Summary/Milestones: Toward trustworthy human-centered machine intelligence

<table>
<thead>
<tr>
<th>Project Title</th>
<th>Description</th>
<th>Research Area</th>
<th>Milestones</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Emotion Recognition</strong></td>
<td>Privacy and utility preserving data transformation for Speech Emotion Reco.</td>
<td>Privacy preserving representation learning</td>
<td>1. Learn data transformations for speech to hide sensitive emotions</td>
</tr>
<tr>
<td></td>
<td>Using intrinsic relationships between label/annotation noise to improve emotion recognition</td>
<td>Federated learning, learning with noisy labels</td>
<td>2. Suppress demographic info. (e.g., gender)</td>
</tr>
<tr>
<td><strong>Egocentric wearable health attribute modelling</strong></td>
<td>Modelling health attributes from in-situ egocentric wearable audio</td>
<td>Wearable sensing, Egocentric audio, Behavior modelling</td>
<td>1. Incorporating annotator reliability/label noise into federated learning setup</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2. Having a central model that is generalizable enough for different datasets</td>
</tr>
</tbody>
</table>
## Summary/ Milestones

<table>
<thead>
<tr>
<th>Project Title</th>
<th>Description</th>
<th>Research Area</th>
<th>Milestones</th>
</tr>
</thead>
<tbody>
<tr>
<td>Federated human activity recognition</td>
<td>Understanding human activity using physiological signals, in trustworthy fashion.</td>
<td>Personalisation, Federated learning</td>
<td>1. Extended federated learning to bio-behavioural signals</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2. Showed effectiveness of personalisation in label sparse regimes</td>
</tr>
<tr>
<td>Fairness in speaker verification</td>
<td>Understand issues of biases and fairness in ASV systems at different stages of the pipeline.</td>
<td>Inclusion, Fairness</td>
<td>1. Used adversarial training methods to improve fairness</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2. Incorporate intersectional fairness by jointly modelling multiple demographic attributes</td>
</tr>
<tr>
<td>Defense against adversarial attacks on speech technology systems</td>
<td>Improve reliability of speech technologies such as ASR and Speaker recognition by defending against adversarial threats</td>
<td>Security, Robustness against adversarial attacks</td>
<td>1. Developed defenses against black-box and white-box adversarial attacks using preprocessors</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2. Improve defense performance by incorporating adversarial training techniques</td>
</tr>
</tbody>
</table>

**Collaborators:**
Anil Ramakrishna, Rahul Gupta

**amazon**
Trustworthy AI Alexa NU group
Elements of Trustworthy Human-centered Machine Intelligence

Strive for an integrative approach across the machine intelligence ecosystem

- **PRIVACY**
  - e.g., safeguard protected information

- **SAFETY**
  - e.g., protect against malicious attacks

- **INCLUSION**
  - e.g., enable broad access