First Grand-Challenge and Workshop on Human Multimodal Language (Challenge-HML)

Organizers:

Amir Zadeh
Louis-Philippe Morency
Paul Pu Liang
Soujanya Poria
Erik Cambria
Stefan Scherer
Zhun Liu
Continuous Theories of (Multimodal) Language

Throughout evolution language and nonverbal behaviors developed together.
Multimodal Language Modalities

“...I really like this vacuum cleaner, it was one of the few that has a large dust bag...”

Language (words)  Vision (gestures)  Acoustic (voice)
Our goal in the first Challenge-HML is to build models of sentiment and emotions for human multimodal language through the proxy of in-the-wild speech videos.
Challenge-HML Metrics

- Sentiment: Binary and Multiclass Sentiment Analysis

- Emotion Recognition: Binary and Multiclass Analysis of
  - Happiness
  - Sadness
  - Anger
  - Surprise
  - Disgust
  - Fear
Difficulties in Modeling Multimodal Language

- Modeling
  - Intra-modal Dynamics: difficulties in modeling each modality.

Language (words)  Vision (gestures)  Acoustic (voice)
Difficulties in Modeling Multimodal Language

- **Modeling**
  - **Intra-modal Dynamics**
    - Difficulties in modeling each modality
  - **Inter-modal Dynamics (fusion):** difficulties in modeling spatio-temporal relations between modalities
Difficulties in Modeling Multimodal Language

- **Modeling**
  - Intra-modal Dynamics
    - Difficulties in modeling each modality
  - Inter-modal Dynamics (fusion)

- **Idiosyncratic signal**
  - Speakers talk and behave differently
CMU-MOSEI Dataset

- CMU-Multimodal Sentiment and Emotion Intensity Analysis Dataset

  - Largest multimodal dataset of sentiment and emotion analysis.
  - More than 23,000 sentence utterances from YouTube.
  - More than 3,000 YouTube videos.
  - More than 1000 speakers.
  - More than 250 topics.
  - All three modalities.
  - Manual transcriptions.
  - Phoneme-level Alignment (semi-automatic)
  - Annotated for sentiment and emotions (Likert scales)
    - 7-scale sentiment (CMU-MOSI, SST)
    - 4-scale emotions (happiness, sadness, anger, disgust, surprise, fear)
## CMU-MOSEI Dataset

<table>
<thead>
<tr>
<th>Dataset</th>
<th># S</th>
<th># Sp</th>
<th>Mod</th>
<th>Sent</th>
<th>Emo</th>
<th>TL (hh:mm:ss)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CMU-MOSEI</td>
<td>23,453</td>
<td>1,000</td>
<td>{l, v, a}</td>
<td>✓</td>
<td>✓</td>
<td>65:53:36</td>
</tr>
<tr>
<td>CMU-MOSI</td>
<td>2,199</td>
<td>98</td>
<td>{l, v, a}</td>
<td>✓</td>
<td>✗</td>
<td>02:36:17</td>
</tr>
<tr>
<td>ICT-MMMA</td>
<td>340</td>
<td>200</td>
<td>{l, v, a}</td>
<td>✓</td>
<td>✗</td>
<td>13:58:29</td>
</tr>
<tr>
<td>YouTube</td>
<td>300</td>
<td>50</td>
<td>{l, v, a}</td>
<td>✓</td>
<td>✗</td>
<td>00:29:41</td>
</tr>
<tr>
<td>MOUD</td>
<td>400</td>
<td>101</td>
<td>{l, v, a}</td>
<td>✓</td>
<td>✗</td>
<td>00:59:00</td>
</tr>
<tr>
<td>SST</td>
<td>11,855</td>
<td>–</td>
<td>{}</td>
<td>✗</td>
<td>✓</td>
<td>–</td>
</tr>
<tr>
<td>Cornell</td>
<td>2,000</td>
<td>–</td>
<td>{}</td>
<td>✓</td>
<td>✓</td>
<td>–</td>
</tr>
<tr>
<td>Large Movie</td>
<td>25,000</td>
<td>–</td>
<td>{}</td>
<td>✓</td>
<td>✓</td>
<td>–</td>
</tr>
<tr>
<td>STS</td>
<td>5,513</td>
<td>–</td>
<td>{}</td>
<td>✓</td>
<td>✓</td>
<td>–</td>
</tr>
<tr>
<td>IEMOCAP</td>
<td>10,000</td>
<td>10</td>
<td>{l, v, a}</td>
<td>✗</td>
<td>✓</td>
<td>11:28:12</td>
</tr>
<tr>
<td>SAL</td>
<td>23</td>
<td>4</td>
<td>{v, a}</td>
<td>✗</td>
<td>✓</td>
<td>11:00:00</td>
</tr>
<tr>
<td>VAM</td>
<td>499</td>
<td>20</td>
<td>{v, a}</td>
<td>✓</td>
<td>✓</td>
<td>12:00:00</td>
</tr>
<tr>
<td>VAM-faces</td>
<td>1,867</td>
<td>20</td>
<td>{v}</td>
<td>✗</td>
<td>✓</td>
<td>–</td>
</tr>
<tr>
<td>HUMAIN</td>
<td>50</td>
<td>4</td>
<td>{v, a}</td>
<td>✓</td>
<td>✓</td>
<td>04:11:00</td>
</tr>
<tr>
<td>RECOLA</td>
<td>46</td>
<td>46</td>
<td>{v, a}</td>
<td>✓</td>
<td>✓</td>
<td>03:50:00</td>
</tr>
<tr>
<td>SEWA</td>
<td>538</td>
<td>408</td>
<td>{v, a}</td>
<td>✓</td>
<td>✓</td>
<td>04:39:00</td>
</tr>
<tr>
<td>SEMAINE</td>
<td>80</td>
<td>20</td>
<td>{v, a}</td>
<td>✓</td>
<td>✓</td>
<td>06:30:00</td>
</tr>
<tr>
<td>AFEW</td>
<td>1,645</td>
<td>330</td>
<td>{v, a}</td>
<td>✓</td>
<td>✓</td>
<td>02:28:03</td>
</tr>
<tr>
<td>AM-FED</td>
<td>242</td>
<td>242</td>
<td>{v}</td>
<td>✓</td>
<td>✓</td>
<td>03:20:25</td>
</tr>
<tr>
<td>Mimicry</td>
<td>48</td>
<td>48</td>
<td>{v, a}</td>
<td>✓</td>
<td>✓</td>
<td>11:00:00</td>
</tr>
<tr>
<td>AFEW-VA</td>
<td>600</td>
<td>240</td>
<td>{v, a}</td>
<td>✓</td>
<td>✗</td>
<td>00:40:00</td>
</tr>
</tbody>
</table>
Keynotes

Dr. Bing Liu
University of Illinois at Chicago (USA)

Dr. Sharon Oviatt
Monash University (Australia)

Dr. Roland Göcke
University of Canberra (Australia)
Thank you for joining us today