TOWARD VISUAL VOICE ACTIVITY DETECTION FOR UNCONSTRAINED VIDEOS
Rahul Sharma, Krishna Somandepalli and Shrikanth Narayanan
University of Southern California, USA

Problem: Visual VAD
Voice activity detection (VAD) from just the visual modality without the need for face detection and tracking.

Idea
Classify voice activity from video image frames ⇒ Such a model can learn to attend to speaking faces.

Approach
Supervised cross modal learning to map video image frames to audio VAD labels.
- We obtain coarse VAD labels from movies’ subtitles.
- We propose Hierarchical Context Aware (HICA) deep architecture that can capture short-term spatial-temporal context and long-term temporal context.
- The use of 3D CNNs makes HICA highly interpretable.
- We show that HICA attends on human faces (and persons) when there is speech activity.
- The VAD performance of HICA is moderate: accuracy 66.1%, F score: 55.7%.

Dataset
Media content: 97 Hollywood movies + Subtitles

<table>
<thead>
<tr>
<th></th>
<th>Speech (hours)</th>
<th>Non-speech (hours)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training Set</td>
<td>66.64</td>
<td>64.18</td>
</tr>
<tr>
<td>Validation Set</td>
<td>16.07</td>
<td>16.28</td>
</tr>
<tr>
<td>Test Set</td>
<td>15.86</td>
<td>15.23</td>
</tr>
</tbody>
</table>

Analysis of CAMs
- Expt1 (Face): Compare against all face detection regions.
- Expt2 (Human Body): Compare with all human body proposals.
- Expt3 (Person): Analyse non-face predictions.

Future Work
- Multimodal fusion to complement audio-VAD systems.
- Active speaker detection using the learned representations.