**Speech2Action:** Cross-modal Supervision for Action Recognition
Arsha Nagrani1,2  Chen Sun2  David Ross2  Rahul Sukthankar2  Cordelia Schmid2  Andrew Zisserman1,3

1VGG, Oxford  2Google Research  3DeepMind

### Problem Definition and Contribution

**Goal:** Action recognition in movies and TV shows using only the transcribed speech as supervision.

**Motivation:**
- Manual annotation of human actions is expensive and not scalable.
- The audio track is usually freely available for large video corpora.

**Key Contributions:**
- A Speech2Action model trained from literary screenplays that predicts actions from transcribed speech alone.
- By applying this Speech2Action model to a large unlabelled corpus of videos, we obtain weak action labels for over 800K video clips.
- An action classifier trained on these clips performs better than fully supervised models.

### Mining with the Speech2Action Model

**Main idea:** We train a text-based model to predict actions from transcribed speech alone. This model is trained on movie scripts from IMDb. This can be applied to the transcribed speech from unlabelled videos to automatically get labels for video clips.

**Speech2Action Model**
- We obtain speech-action paired data for 18 action classes from the IMDb data.
- We finetune a BERT model pretrained on English Wikipedia and the BooksCorpus.

**Mining Clips Automatically:**
- We apply the Speech2Action model to the subtitles of unlabelled movies and TV shows.
- We then assign the label for highly confident predictions of the model to the accompanying video clip.

In this manner we mined over 800K video clips and assign them with action labels based on the speech alone.

---

More details at: https://www.robots.ox.ac.uk/~vgg/research/speech2action/
Contact: arsha@robots.ox.ac.uk

### Results on Visual Action Recognition

#### Examples of clips mined using Speech2Action:

- Phone: “hello?...are you sure?...I need jeff on his secure line.”
- Drink: “Ah, I am sure you are drinking in the wrong sense.”
- Dance: “Do you want to see that up there?...let’s see.”

#### Examples of abstract actions mined using Speech2Action:

- Eat: “This chicken is very tasty...I am so hungry...and money on the cover more money.”
- Run: “Run!...run!...I am trying to follow him.”
- Count: “Two quarters, three dimes, one nickel...”

---

#### Results on 14 A V A mid and tail classes

<table>
<thead>
<tr>
<th>Data</th>
<th>drive</th>
<th>phone</th>
<th>kiss</th>
<th>dance</th>
<th>eat</th>
<th>drink</th>
<th>run</th>
<th>point</th>
<th>open</th>
<th>hit</th>
<th>shoot</th>
<th>push</th>
<th>bug</th>
<th>enter</th>
</tr>
</thead>
<tbody>
<tr>
<td>AVA fully supervised</td>
<td>0.63</td>
<td>0.54</td>
<td>0.22</td>
<td>0.46</td>
<td>0.67</td>
<td>0.27</td>
<td>0.66</td>
<td>0.02</td>
<td>0.49</td>
<td>0.62</td>
<td>0.08</td>
<td>0.09</td>
<td>0.29</td>
<td>0.14</td>
</tr>
<tr>
<td>S2A-mined (raw)</td>
<td>0.83</td>
<td>0.79</td>
<td>0.13</td>
<td>0.55</td>
<td>0.68</td>
<td>0.30</td>
<td>0.63</td>
<td>0.04</td>
<td>0.52</td>
<td>0.54</td>
<td>0.18</td>
<td>0.04</td>
<td>0.07</td>
<td>0.04</td>
</tr>
<tr>
<td>S2A-mined + AVA</td>
<td>0.86</td>
<td>0.89</td>
<td>0.34</td>
<td>0.58</td>
<td>0.78</td>
<td>0.42</td>
<td>0.75</td>
<td>0.03</td>
<td>0.65</td>
<td>0.72</td>
<td>0.26</td>
<td>0.13</td>
<td>0.36</td>
<td>0.16</td>
</tr>
</tbody>
</table>

#### Action Recognition Model
- We train an S3D-G model for 18-way classification on video clips labelled with the Speech2Action model.
- We evaluate on AVA with NO finetuning, on mid and tail classes. These actions occur rarely, and are hard to get manual supervision for. For 8 classes, we exceed fully supervised performance without a single manually labelled training example.
- On HMBDS1, we obtain a 17% improvement over training from scratch and also outperform previous self-supervised and weakly supervised works.

### Results on HMBDS1

<table>
<thead>
<tr>
<th>Method</th>
<th>Architecture</th>
<th>Pre-training</th>
<th>Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shuffle &amp; Learn</td>
<td>S3D-G (RGB)</td>
<td>UCF101</td>
<td>35.8</td>
</tr>
<tr>
<td>OPN</td>
<td>VGG-M-2048</td>
<td>UCF101</td>
<td>23.8</td>
</tr>
<tr>
<td>ClipOrder</td>
<td>R(2+1)D</td>
<td>UCF101</td>
<td>30.9</td>
</tr>
<tr>
<td>3DRoNet</td>
<td>S3D-G (RGB)</td>
<td>Kinetics</td>
<td>40.0</td>
</tr>
<tr>
<td>DPC</td>
<td>3DRoNet18</td>
<td>Kinetics</td>
<td>35.7</td>
</tr>
<tr>
<td>CBT</td>
<td>S3D-G (RGB)</td>
<td>Kinetics</td>
<td>44.6</td>
</tr>
<tr>
<td>Dislant (RGB) 2019</td>
<td>R(2+1)D</td>
<td>Kinetics++</td>
<td>54.8</td>
</tr>
<tr>
<td>Korbar et al. 2018</td>
<td>S3D-G (RGB)</td>
<td>Scratch</td>
<td>41.2</td>
</tr>
<tr>
<td>Ours</td>
<td>S3D-G (RGB)</td>
<td>S2A-mined</td>
<td>58.1</td>
</tr>
</tbody>
</table>

#### Acknowledgments
Arsha Nagrani is supported by a Google PhD Fellowship. We are grateful to Carl Vondrick for early discussions.