Speechreading is Difficult

**speech-reading n.** use of non-auditory clues as to what is being said, acquired by observing the speaker's facial expressions, lip and jaw movements. Formerly called lip reading

- Two approaches for automating speechreading:
  1. Classification: Output is word from predefined vocabulary, phoneme
  2. Regression: Output is audio signal
- Regression advantages:
  - No input pre-segmentation
  - Vocabulary-agnostic
  - Learn using “natural supervision”
  - Ability to output emotion, prosody

Visual Representation (Input)

- Speaker’s face cropped and rescaled to 128 x 128 pixels
- K consecutive grayscale frames (K=9 worked best)

⇒ CNN input volume of 128 x 128 x 9 numbers

Speech Representation (Output)

- Audio downsampling to 8 kHz
- 8th order LPC and LSP on half-overlapping 40ms waveform segments
- Concatenate every 2 successive LSP vectors

⇒ Network output of 18 numbers

VID2SPEECH: SPEECH RECONSTRUCTION FROM SILENT VIDEO

Ariel Ephrat        Shmuel Peleg
The Hebrew University of Jerusalem

**Visual Representation (Input)**

- Audio-visual recordings of 34 speakers
- Each has 1000 3-second videos @ 25 FPS containing 6-word sequence of form shown above

**Evaluation**

- Used Amazon MTurk for human intelligibility testing
- Followed protocol used by [10]
- Workers were given GRID vocabulary
- Each job was transcription of either:
  1. Audio only - reconstructed speech with no video
  2. Audio-visual - reconstructed speech with original video frames
  3. Out-of-vocabulary Audio-Visual
- Over 400 videos (38 distinct) transcribed by 23 people

Results

**Visualization of Original vs. Reconstructed Speech**

- Vertical columns of (a) are actual output of CNN
- In (c), unvoiced excitation causes lack of formants (horizontal lines)

1. Reconstruction from full dataset

Dataset:
- Train on 800 videos from one GRID speaker (60K frames)
- Test on remaining 200 videos

<table>
<thead>
<tr>
<th>Dataset</th>
<th>[10]</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>S4</td>
<td>84%</td>
<td>85%</td>
</tr>
<tr>
<td>S2</td>
<td>82%</td>
<td>83%</td>
</tr>
</tbody>
</table>

2. Reconstruction of out-of-vocabulary (OOV) words

Dataset:
- Train on (S4) videos containing only 8 spoken digits
- Test on videos containing 2 OOV digits
- Results averaged across 5 splits

<table>
<thead>
<tr>
<th>Dataset</th>
<th>OOV Digits 0-9</th>
<th>None out</th>
<th>Chance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Audio-visual</td>
<td>53.6%</td>
<td>93.4%</td>
<td>10.0%</td>
</tr>
</tbody>
</table>

3. Learning from mouth only vs. full face

- Face region error is 40% lower than mouth only
- Disambiguation effect of using temporal context is clear