Motion Dynamics Improve Speaker-Independent Lipreading

Matteo Riva, Michael Wand, Jürgen Schmidhuber
matteo.riva@usi.ch, michael@idsia.ch, juergen@idsia.ch
What is Lipreading?

Audio-Visual Recognition systems

Historically Lipreading has been adopted to improve audio speech recognition in noisy environments: the first to use it was Petajan [1984].

Lipreading as a standalone problem

Being it a challenging task, as pointed out by Stork et al. [1992], it has also been studied as a standalone problem. The first to do so were Chiou and Hwang [1997].
Lipreading involves dealing with many diverse problems.

Automatic Lipreading systems do not generalize well over unseen speakers, as investigated among others by Cox et al. [2008]; Chung and Zisserman [2017]; Wand and Schmidhuber [2017].

Physical traits differ from speaker to speaker:
- Gender, age, ethnicity
- Mustaches, beard, lipstick
- Mouth conformation

Speaker-Independence is an open problem.
Goal

1. **Improve generalization** over speech uttered by **unknown speakers**

2. Evaluate our new method on a **word-level Lipreading** task

**How?**

Taking inspiration from Villegas et al. [2017], we want to build a system that also explicitly models the **motion dynamics of speech**.
Goal

1. Improve generalization over speech uttered by unknown speakers

2. Evaluate our new method on a word-level Lipreading task

How?

Taking inspiration from Villegas et al. [2017], we want to build a system that also explicitly models the motion dynamics of speech.
Data Corpus and Dataset preparation

GRID Data Corpus
- 34 speakers
- Strict grammar and sentence structure
  \text{command}\{4\} + \text{color}\{4\} + \text{preposition}\{4\} + \text{letter}\{25\} + \text{digit}\{10\} + \text{adverb}\{4\}
  Example: “Place green at g 6 again”
- 51 unique words, 6000 uttered by each speaker

20 development
34 speakers
13 evaluation
How to test Speaker-Independence

1. **SOURCE SPEAKERS**
   - Training the model
   - Setups: 1 / 4 / 8 speakers

2. **TARGET SPEAKER**
   - Testing the model
   - Always 1 speaker
Development Setup

Data splits
We divided data from each speaker into train, validation and test splits.
Validation and test sets are target balanced.

Cross-Speaker Validation
We took each development speaker as the target speaker one and only one time.
We report only average word classification accuracy.

<table>
<thead>
<tr>
<th>SPEAKER</th>
<th>TRAIN</th>
<th>VALIDATION</th>
<th>TEST</th>
</tr>
</thead>
<tbody>
<tr>
<td>s1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>s2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>s20</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>s1</td>
<td>s2</td>
</tr>
<tr>
<td>s2</td>
<td>s3</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>s20</td>
<td>s1</td>
</tr>
</tbody>
</table>

AVG   AVG
Baseline Definition (1)

We define a baseline system that does not explicitly model motion dynamics.

How?
**Baseline Definition (2)**

<table>
<thead>
<tr>
<th>Layers / Neurons</th>
<th>128</th>
<th>256</th>
<th>512</th>
</tr>
</thead>
<tbody>
<tr>
<td>(FF+DP)×1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LSTM×1</td>
<td>80.2%</td>
<td>80.6%</td>
<td>79.9%</td>
</tr>
<tr>
<td>LSTM×2</td>
<td>81.4%</td>
<td>81.1%</td>
<td>80.4%</td>
</tr>
<tr>
<td>LSTM×3</td>
<td>80.7%</td>
<td>81.0%</td>
<td>80.5%</td>
</tr>
<tr>
<td>(FF+DP)×2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LSTM×1</td>
<td>79.9%</td>
<td>80.2%</td>
<td>79.2%</td>
</tr>
<tr>
<td>LSTM×2</td>
<td>79.9%</td>
<td>80.1%</td>
<td>79.7%</td>
</tr>
<tr>
<td>LSTM×3</td>
<td>79.3%</td>
<td>79.3%</td>
<td>79.8%</td>
</tr>
<tr>
<td>(FF+DP)×3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LSTM×1</td>
<td>77.6%</td>
<td>78.8%</td>
<td>77.9%</td>
</tr>
<tr>
<td>LSTM×2</td>
<td>77.4%</td>
<td>78.4%</td>
<td>78.2%</td>
</tr>
<tr>
<td>LSTM×3</td>
<td>77.1%</td>
<td>77.5%</td>
<td>77.5%</td>
</tr>
<tr>
<td>(FF+DP)×4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LSTM×1</td>
<td>75.6%</td>
<td>76.9%</td>
<td>76.3%</td>
</tr>
<tr>
<td>LSTM×2</td>
<td>75.9%</td>
<td>76.7%</td>
<td>75.2%</td>
</tr>
<tr>
<td>LSTM×3</td>
<td>74.9%</td>
<td>76.0%</td>
<td>75.8%</td>
</tr>
</tbody>
</table>

**How?**

We *experimentally defined it altering meta-parameters* of base system by Wand and Schmidhuber [2017]:

- Feed-forward layers
- LSTM layers
- Hidden Units
Experiments

Dual-Pipeline MC Definition

JointLSTM (128 units)

(FF+DP)×2 + LSTM×1
(w/ 32 units bottleneck)
(w/ content downsampling)
Development Results

JointLSTM (128 units)

\[(\text{FF+DP}) \times 2 + \text{LSTM} \times 1\]

- Baseline (Content only)
- Dual-pipeline Motion&Content

<table>
<thead>
<tr>
<th>Speakers</th>
<th>Baseline (%)</th>
<th>Dual-pipeline (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>22.3</td>
<td>24.9</td>
</tr>
<tr>
<td>4</td>
<td>42.3</td>
<td>47.0</td>
</tr>
<tr>
<td>8</td>
<td>46.4</td>
<td>51.7</td>
</tr>
</tbody>
</table>
Evaluation Setup

Data splits
We divided data from each speaker into train, validation and test splits. Validation and test sets are target balanced.

Cross-Speaker Validation
We took each evaluation speaker as the target speaker one and only one time.
We report only average word classification accuracy.

<table>
<thead>
<tr>
<th>SPEAKER</th>
<th>TRAIN</th>
<th>VALIDATION</th>
<th>TEST</th>
</tr>
</thead>
<tbody>
<tr>
<td>s22-s23-s24-s25</td>
<td>s26</td>
<td>s23-s24-s25-s26</td>
<td>s27</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>s34-s22-s23-s24</td>
<td>s25</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

AVG  | AVG

T-Test
We measure statistical significance of improvements yielded by Dual-Pipeline MC w.r.t. the baseline system.

\[
H_0 : \mu_d = 0 \\
H_a : \mu_d > 0
\]
Target speaker word accuracies over the evaluation speakers

- 1 source speaker
- 4 source speakers
- 8 source speakers

Baseline vs Dual-Pipeline MC
## Results

<table>
<thead>
<tr>
<th></th>
<th>1 src speaker</th>
<th>4 src speakers</th>
<th>8 src speakers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Source</td>
<td>Target</td>
<td>Source</td>
</tr>
<tr>
<td>Baseline</td>
<td>80.7%</td>
<td>24.4%</td>
<td>78.6%</td>
</tr>
<tr>
<td>Dual-Pipeline MC</td>
<td>85.0%</td>
<td>26.3%</td>
<td>80.6%</td>
</tr>
<tr>
<td>(relative improvement)</td>
<td>+5.3%</td>
<td>+7.7%</td>
<td>+2.6%</td>
</tr>
<tr>
<td>(p-value)</td>
<td>6.9e−05</td>
<td>0.0215</td>
<td>0.0003</td>
</tr>
</tbody>
</table>

- Improvements both on source and target speakers
- Maintained when increasing the amounts of data used for training
- All improvements are statistically significant (p-values << 0.05)
- Motion Dynamics improve the model speaker-independence
Conclusion

Goal
We set out to build a word-level Lipreading model that improves on Speaker-Independence.

Results
Dual-Pipeline MC architecture yields improvements of $\approx 6.8\%$ on unseen speakers and of $\approx 3.3\%$ on known speakers.

How
We took inspiration from the work by Villegas et al. [2017] on decoupling motion and content.
Goal
We set out to build a word-level Lipreading model that improves on Speaker-Independence.

How
We took inspiration from the work by Villegas et al. [2017] on decoupling motion and content.

Results
Dual-Pipeline MC architecture yields improvements of $\approx 6.8\%$ on unseen speakers and of $\approx 3.3\%$ on known speakers.

Thank you for your attention

References