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].

Problem

- **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**.

Spatial Layout + Motion Dynamics

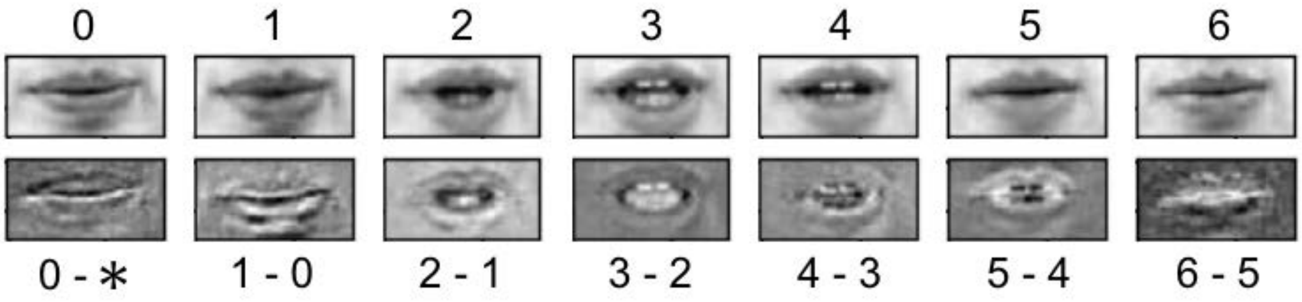
Goal

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

  command{4} + color{4} + preposition{4} + letter{25} + digit{10} + 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

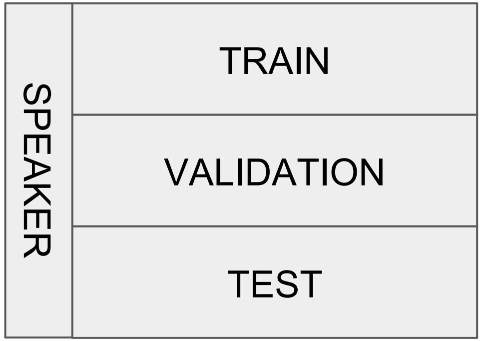
Always 1 speaker

Development Setup

Data splits

We divided data from each speaker into train, validation and test splits.

Validation and test sets are target balanced.



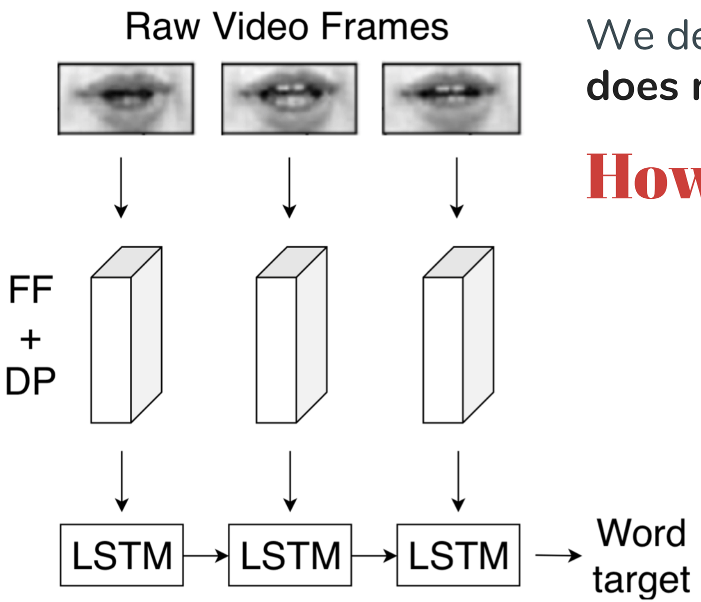
Cross-Speaker Validation

We took each development speaker as the **target speaker one and only one time.**

We report only **average word classification accuracy**.

Source	Target
s1	s2
s2	s3
…	…
s20	s1
AVG	AVG

Baseline Definition (1)



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

How?

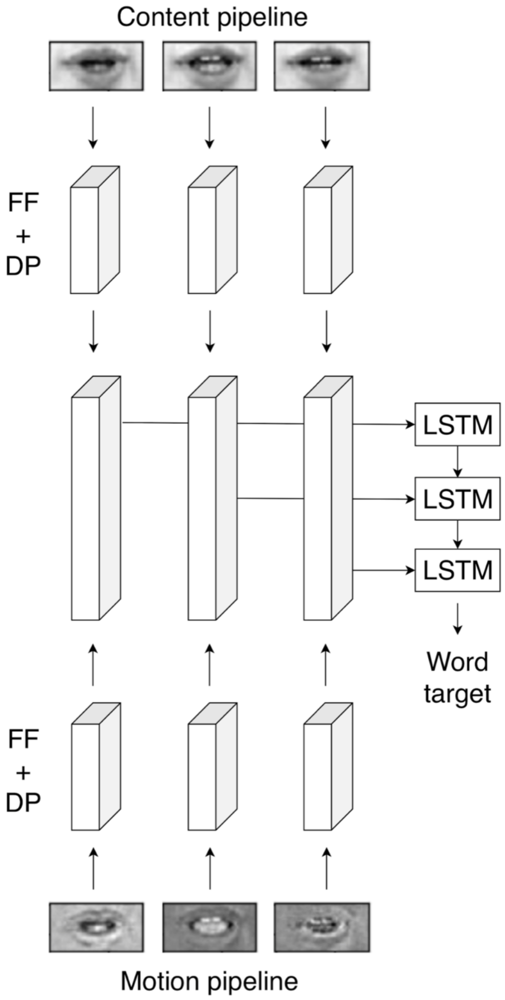
Baseline Definition (2)

Layers / Neurons		128	**256**	512
(FF+DP)×1	LSTM×1	80.2% / 41.0%	80.6% / 41.2%	79.9% / 41.6%
	LSTM×2	81.4% / 42.2%	81.1% / 41.7%	80.4% / 41.3%
	LSTM×3	80.7% / 41.4%	81.0% / 41.1%	80.5% / 41.9%
(FF+DP)×2	**LSTM×1**	79.9% / 41.2%	**80.2% / 42.3%**	79.2% / 41.6%
	LSTM×2	79.9% / 41.6%	80.1% / 42.2%	79.7% / 40.9%
	LSTM×3	79.3% / 41.6%	79.3% / 42.2%	79.8% / 42.1%
(FF+DP)×3	LSTM×1	77.6% / 41.9%	78.8% / 41.3%	77.9% / 41.1%
	LSTM×2	77.4% / 40.9%	78.4% / 41.7%	78.2% / 41.1%
	LSTM×3	77.1% / 41.2%	77.5% / 41.1%	77.5% / 41.4%
(FF+DP)×4	LSTM×1	75.6% / 41.8%	76.9% / 41.3%	76.3% / 40.8%
	LSTM×2	75.9% / 40.1%	76.7% / 40.2%	75.2% / 40.8%
	LSTM×3	74.9% / 40.5%	76.0% / 40.0%	75.8% / 39.9%

How?

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

- Feed-forward layers
- LSTM layers
- Hidden Units



Dual-Pipeline MC Definition

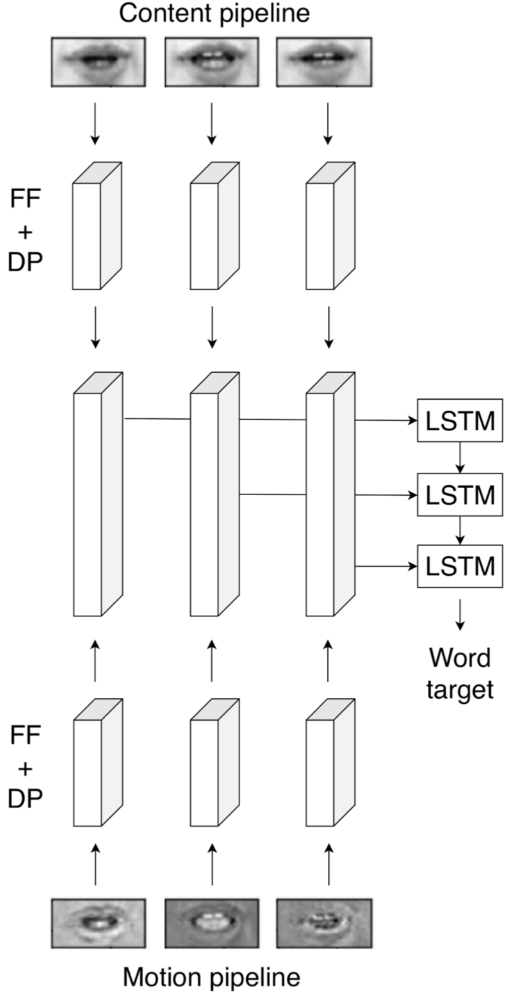
Experiments

JointLSTM (128 units)

(FF+DP)×2 + LSTM×1

(w/ 32 units bottleneck)

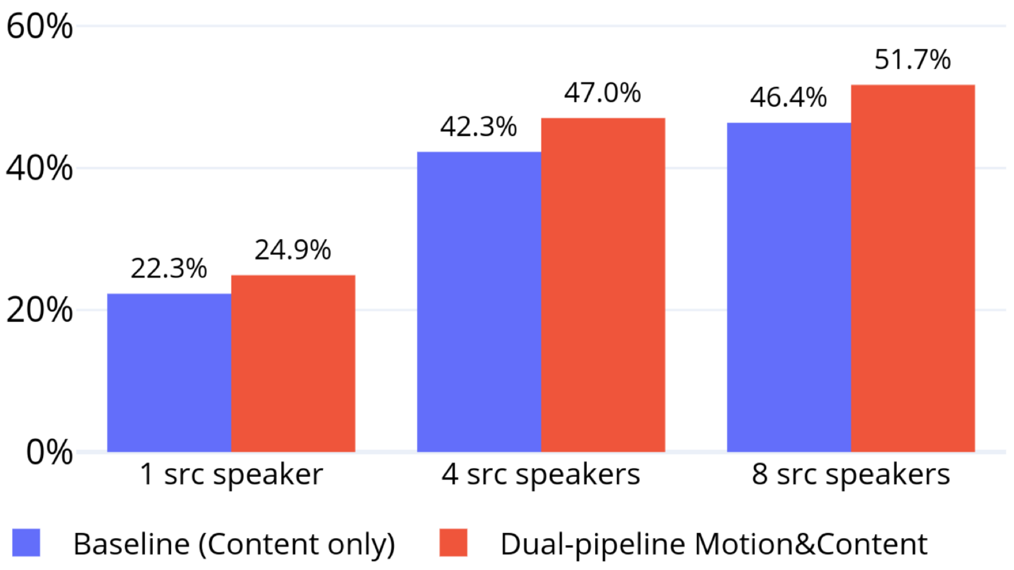
(w/ content downsampling)



JointLSTM (128 units)

(FF+DP)×2 + LSTM×1

Development Results

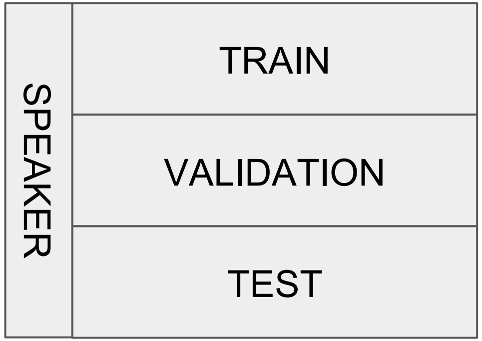


Evaluation Setup

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.**

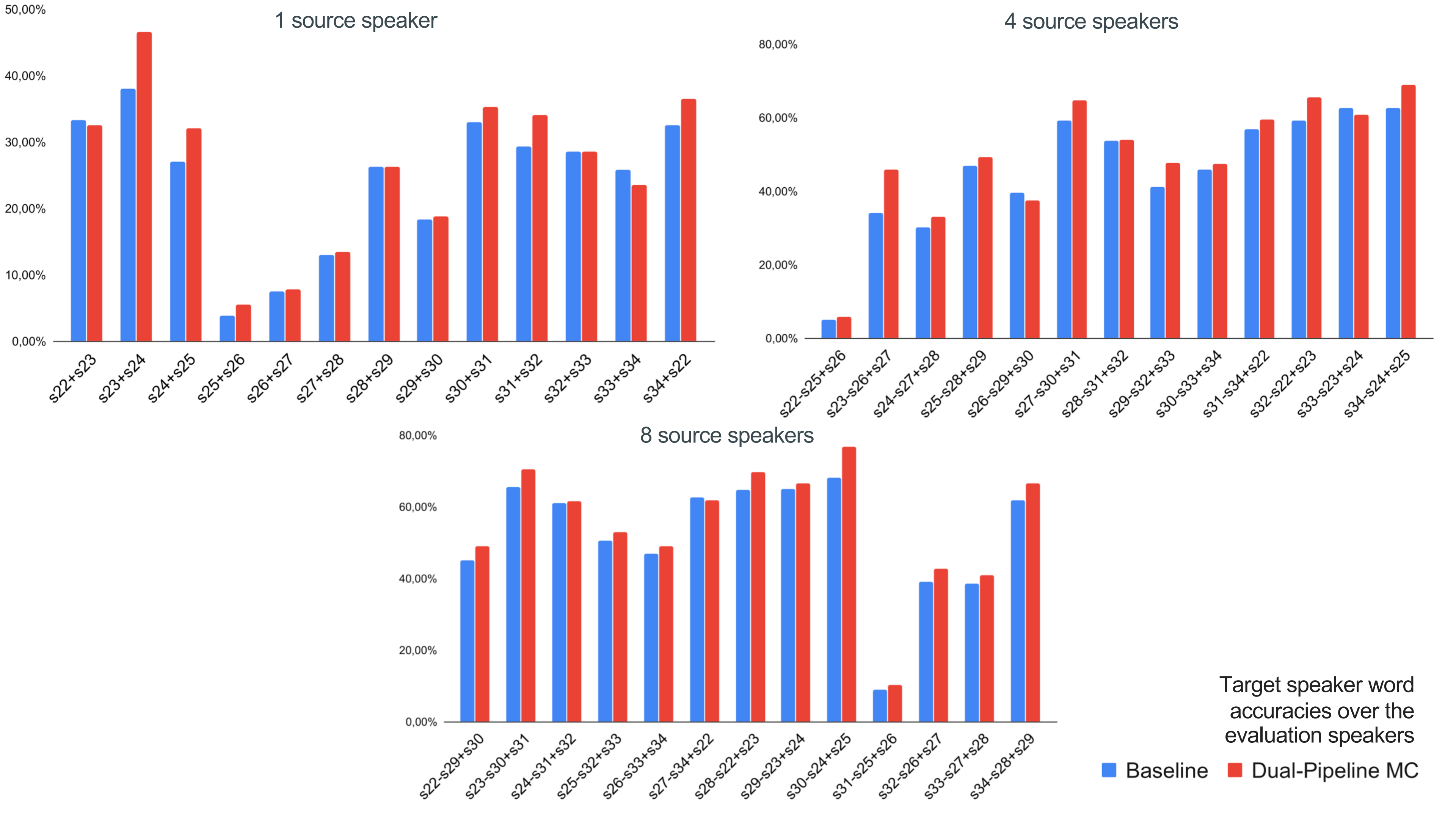
We report only **average word classification accuracy**.

Source	Target
s22-s23-s24-s25	s26
s23-s24-s25-s26	s27
...	...
s34-s22-s23-s24	s25
AVG	AVG

T-Test

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

$$H_0 : \mu_d = 0$$
$$H_a : \mu_d > 0$$


1 source speaker
50,00%
40,00%
30,00%
20,00%
10,00%
0,00%
s22+s23
s23+s24
s24+s25
s25+s26
s26+s27
s27+s28
s28+s29
s29+s30
s30+s31
s31+s32
s32+s33
s33+s34
s34+s22
4 source speakers
80,00%
60,00%
40,00%
20,00%
0,00%
s22-s25+s26
s23-s26+s27
s24-s27+s28
s25-s28+s29
s26-s29+s30
s27-s30+s31
s28-s31+s32
s29-s32+s33
s30-s33+s34
s31-s34+s22
s32-s22+s23
s33-s23+s24
s34-s24+s25
8 source speakers
80,00%
60,00%
40,00%
20,00%
0,00%
s22-s29+s30
s23-s30+s31
s24-s31+s32
s25-s32+s33
s26-s33+s34
s27-s34+s22
s28-s22+s23
s29-s23+s24
s30-s24+s25
s31-s25+s26
s32-s26+s27
s33-s27+s28
s34-s28+s29
Target speaker word accuracies over the evaluation speakers
Baseline
Dual-Pipeline MC

Results

	1 src speaker		4 src speakers		8 src speakers	
	Source	**Target**	**Source**	**Target**	**Source**	**Target**
Baseline	80.7%	24.4%	78.6%	46.0%	76.3%	52.2%
Dual-Pipeline MC	85.0%	26.3%	80.6%	49.3%	77.7%	55.3%
(relative improvement)	*+5.3%*	*+7.7%*	*+2.6%*	*+7.2%*	*+1.9%*	*+5.6%*
(p-value)	$6.9e-05$	0.0215	0.0003	0.0047	$5.4e-05$	0.0004

- 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.**

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 ≈ 6.8% on unseen speakers and of ≈ 3.3% on known speakers.**

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 ≈ 6.8% on unseen speakers and of ≈ 3.3% on known speakers.

Thank you for your attention

References

- **Eric Petajan. *Automatic Lipreading To Enhance Speech Recognition.*** PhD thesis, University of Illinois at Urbana-Champaign, 1984.
- **David G Stork, Greg Wolff, and Earl Levine. *Neural Network Lipreading System For Improved Speech Recognition***. In Proc. IJCNN, volume 2, pages 289–295, 1992.
- **Greg I Chiou and Jenq-Neng Hwang. *Lipreading From Color Video***. IEEE Transactions on Image Processing, 6(8):1192–1195, 1997.
- **Stephen J. Cox, Richard Harvey, Yuxuan Lan, Jacob L. Newman, and Barry-John Theobald. *The Challenge Of Multispeaker Lip-reading***. In Proc. AVSP, pages 179–184, 2008.
- **Joon Son Chung and Andrew Zisserman. *Lip Reading In The Wild***. In Proc. ACCV, pages 87–103, 2017.
- **Michael Wand and Jürgen Schmidhuber. *Improving Speaker-independent Lipreading With Domain-adversarial Training***. In Proc. Interspeech, pages 3662–3666, 2017.
- **Ruben Villegas, Jimei Yang, Seunghoon Hong, Xunyu Lin, and Honglak Lee. *Decomposing Motion And Content For Natural Video Sequence Prediction***. In Proc. ICLR, 2017.