LipNet

End-to-End Sentence-level Lipreading

Yannis Assael, Brendan Shillingford, Shimon Whiteson, Nando de Freitas

UNIVERSITY OF OXFORD

NVIDIA GTC San Jose 2017

Outline

1. Introduction
2. Background
3. LipNet
4. Analysis

UNIVERSITY OF OXFORD

How easy do you think lipreading is?

- McGurk effect (McGurk & MacDonald, 1976)
- Phonemes and Visemes (Fisher, 1968)
- Human lipreading performance is poor

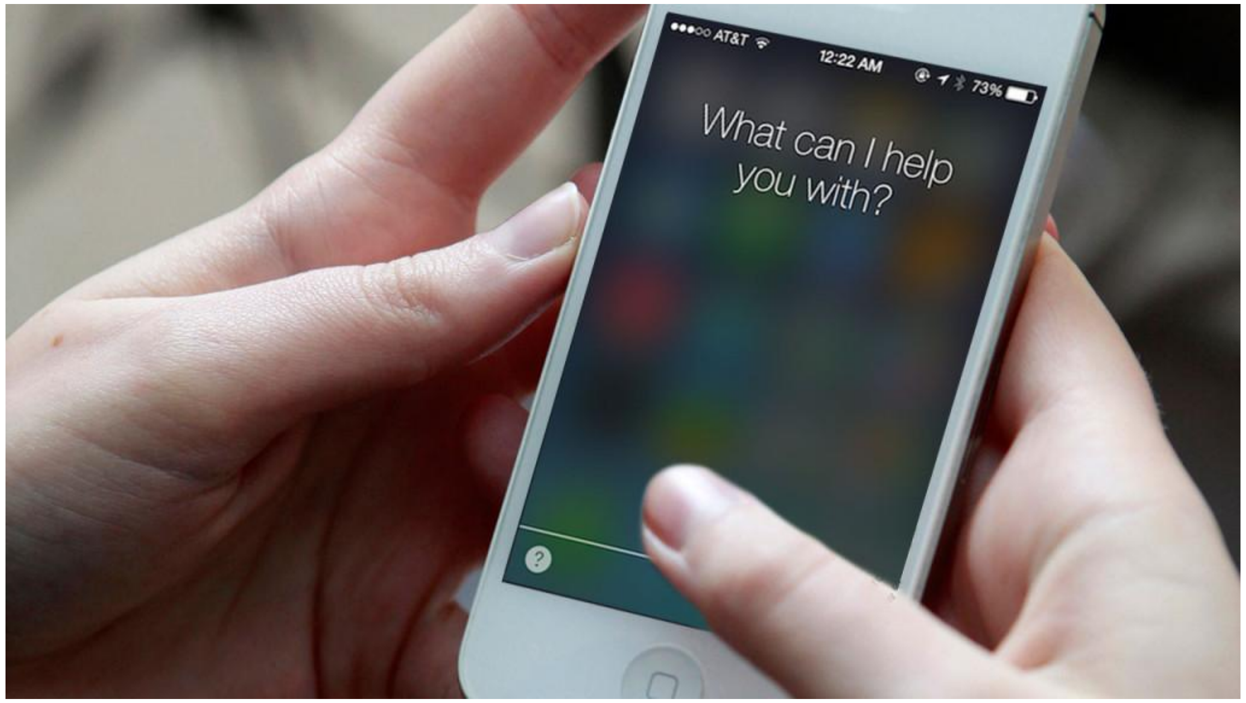
We can improve it…

https://goo.gl/hyFBVQ

Why is lipreading important?

Among others:

- Improved hearing aids
- Speech recognition in noisy environments (e.g. cars)
- Silent dictation in public spaces
- Security
- Biometric identification
- Silent-movie processing

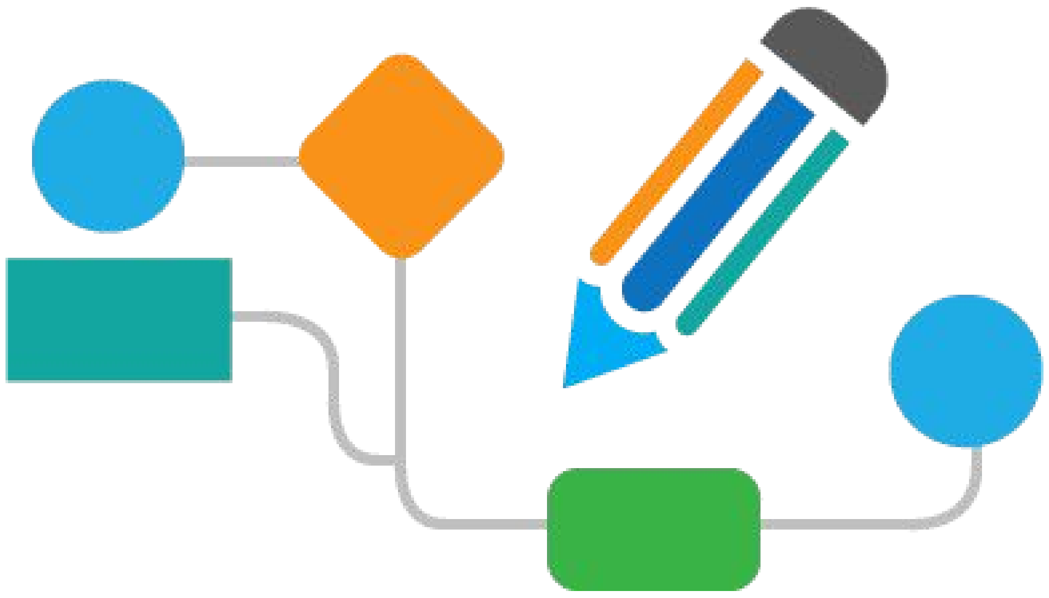


UNIVERSITY OF OXFORD

https://goo.gl/RTXh9Q

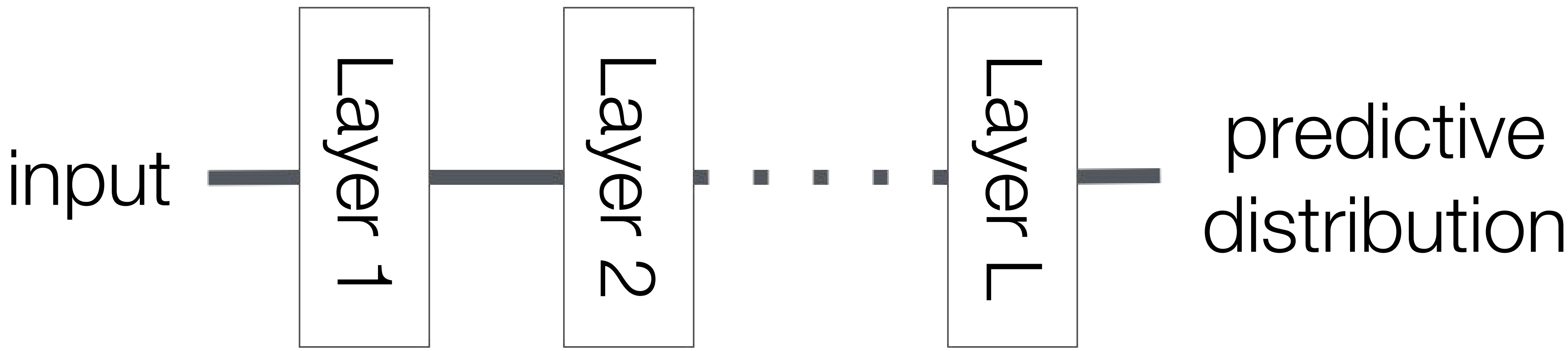
Automated lipreading

- Most existing work does not employ deep learning
- Heavy preprocessing
- Open problems:
  - generalisation across speakers
  - extraction of motion features

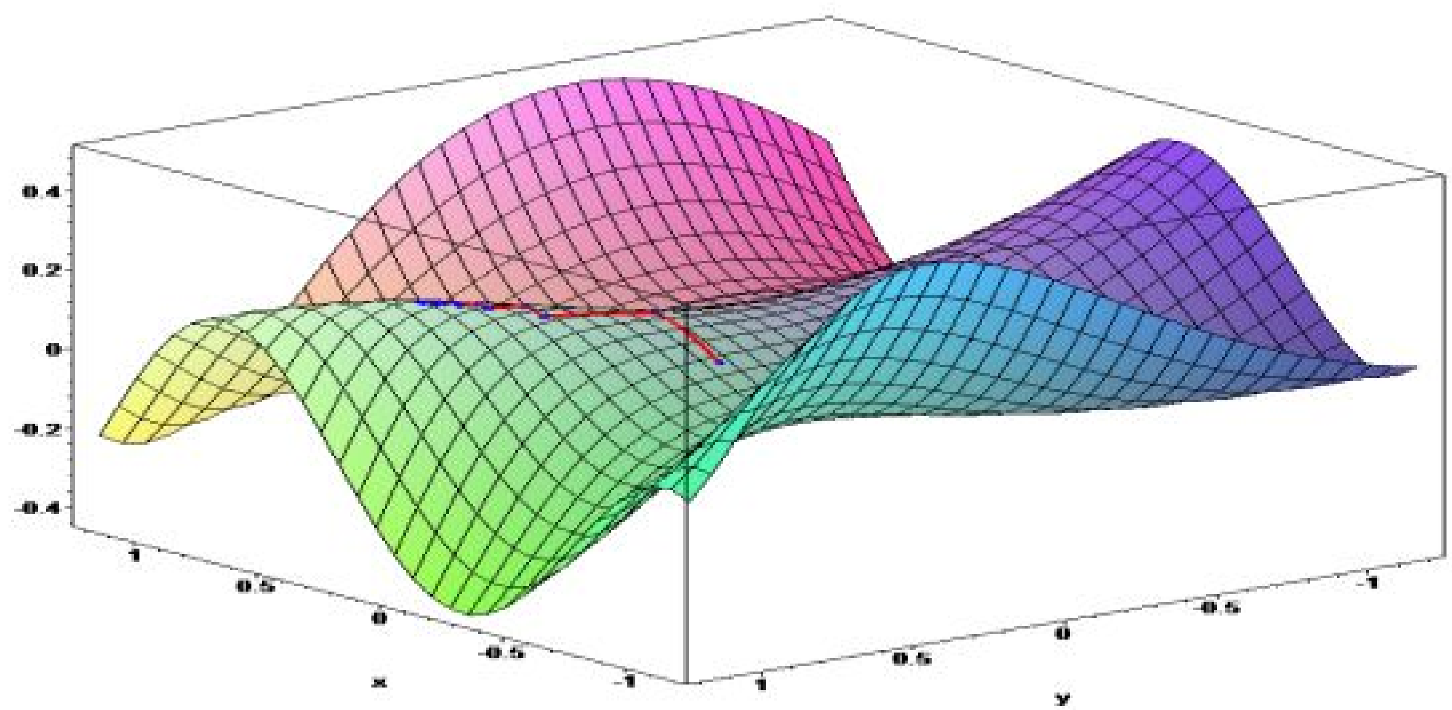


End-to-end supervised learning using NNs

1. Hierarchical, expressive, differentiable function

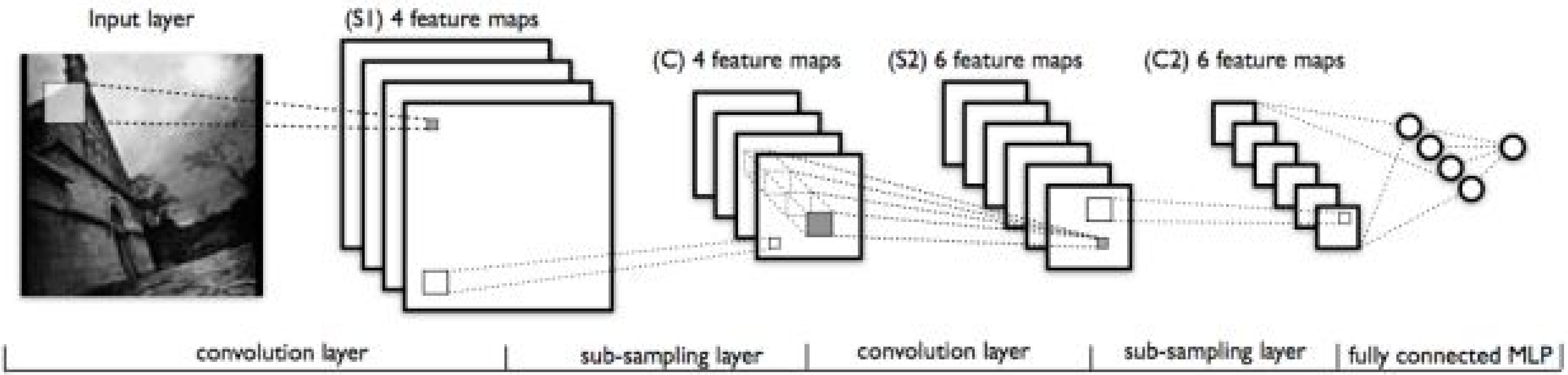


1. Adjust parameters to maximise probability of data with gradient descent



Convolutional Neural Networks

- Model: Deep stacks of local operations.
- Good for: relationships over **space (2D)**:

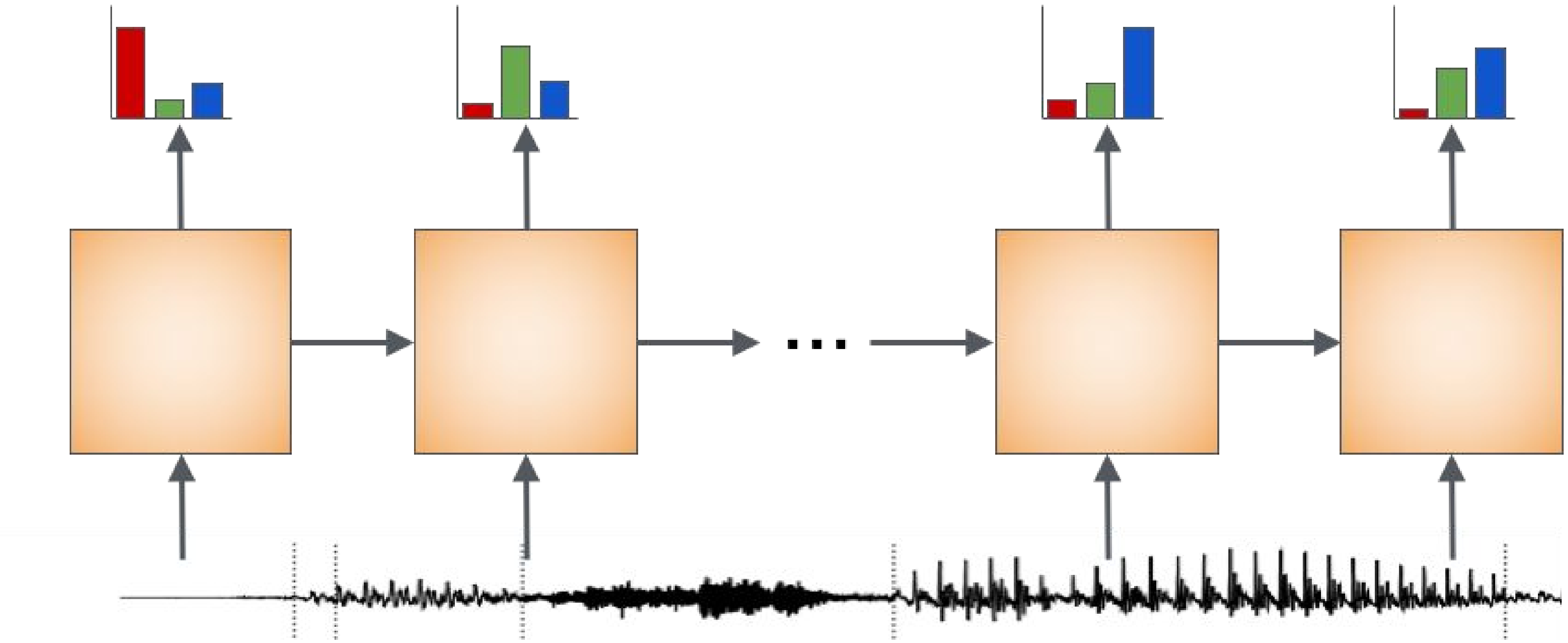


deeplearning.net

- Also good for **time (1D)**
- Or in our case, **space & time (3D)**: every layer can model either or both. Lets the optimisation decide what's best.

Recurrent Neural Networks

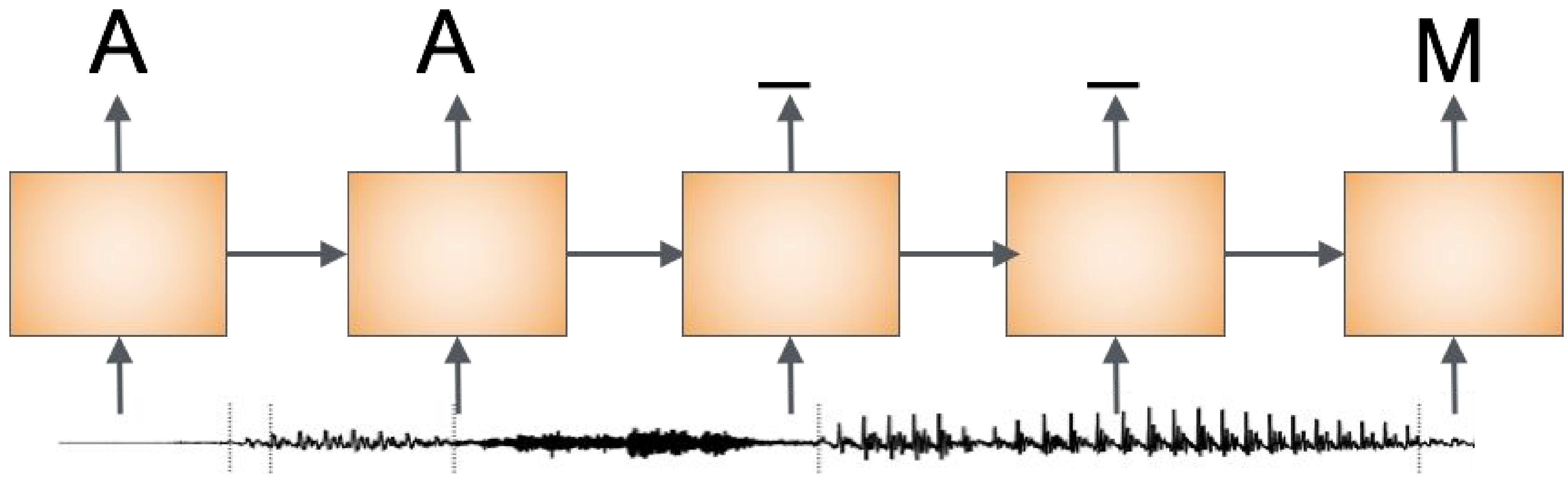
- Model: carry information over time using a state
- Good for: **sequences**



- Often used to predict classes at each timestep
- But what if inputs/outputs are unequal length, or aren't aligned?

Recurrent Neural Networks

- If inputs/outputs aren't aligned, CTC (Graves 2006) efficiently marginalises over all alignments
- To do this, let the RNN output **blanks** or **duplicates**:



- Sum over every way to output the same sequence:
  p(**am**) = p(aam) + p(amm) + p(_am) + p(a_m) + p(am_)

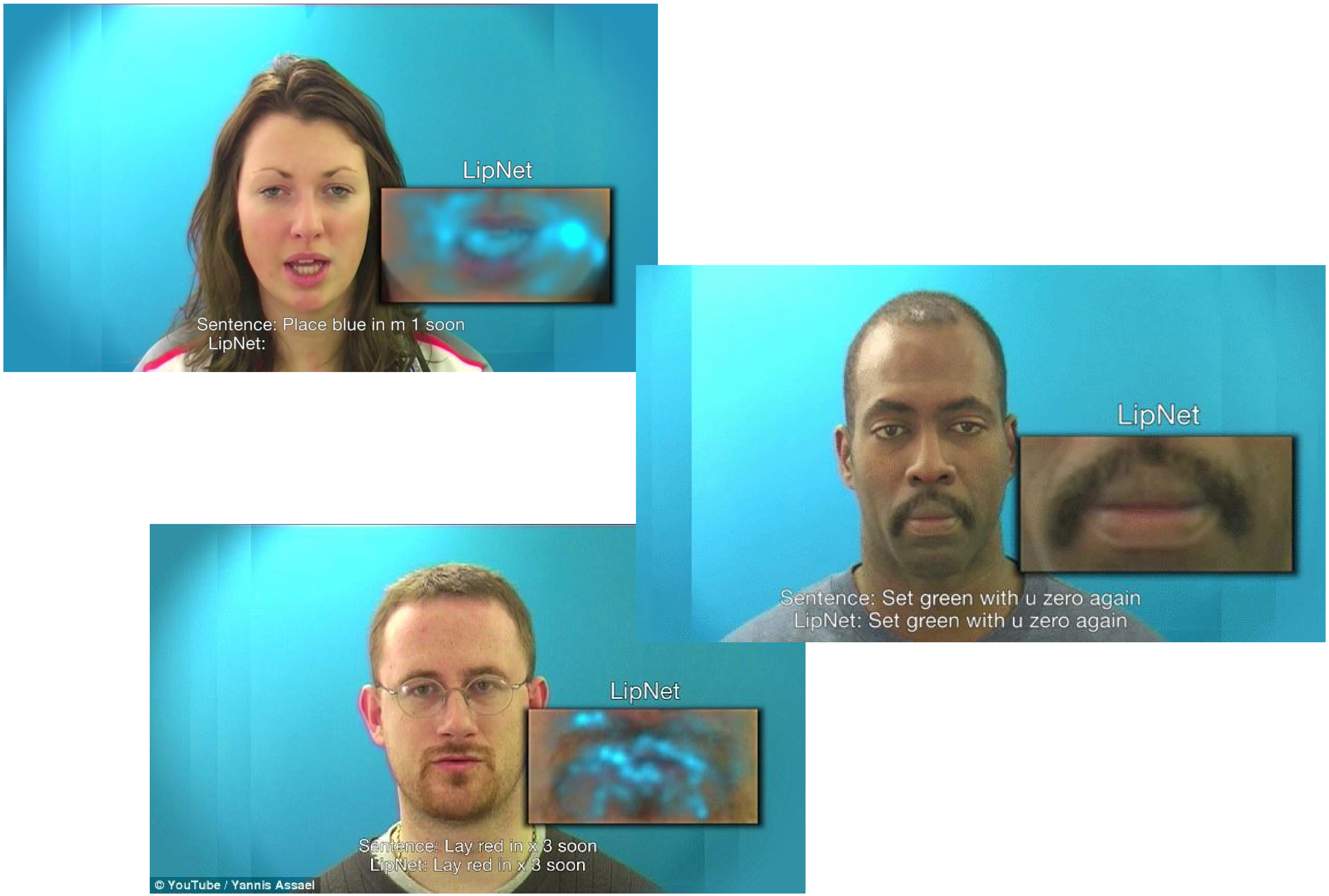
LipNet

- Monosyllabic vs Compound words (Easton & Basala, 1982)
- Spatiotemporal features
- End-to-end, sentence-level
- GRID corpus 33000 sentences

TABLE I. Sentence structure for the Grid corpus. Keywords are identified with asterisks.

command	color*	preposition	letter*	digit*	adverb
bin	blue	at	A–Z	1–9, zero	again
lay	green	by	excluding W		now
place	red	in			please
set	white	with			soon

GRID corpus

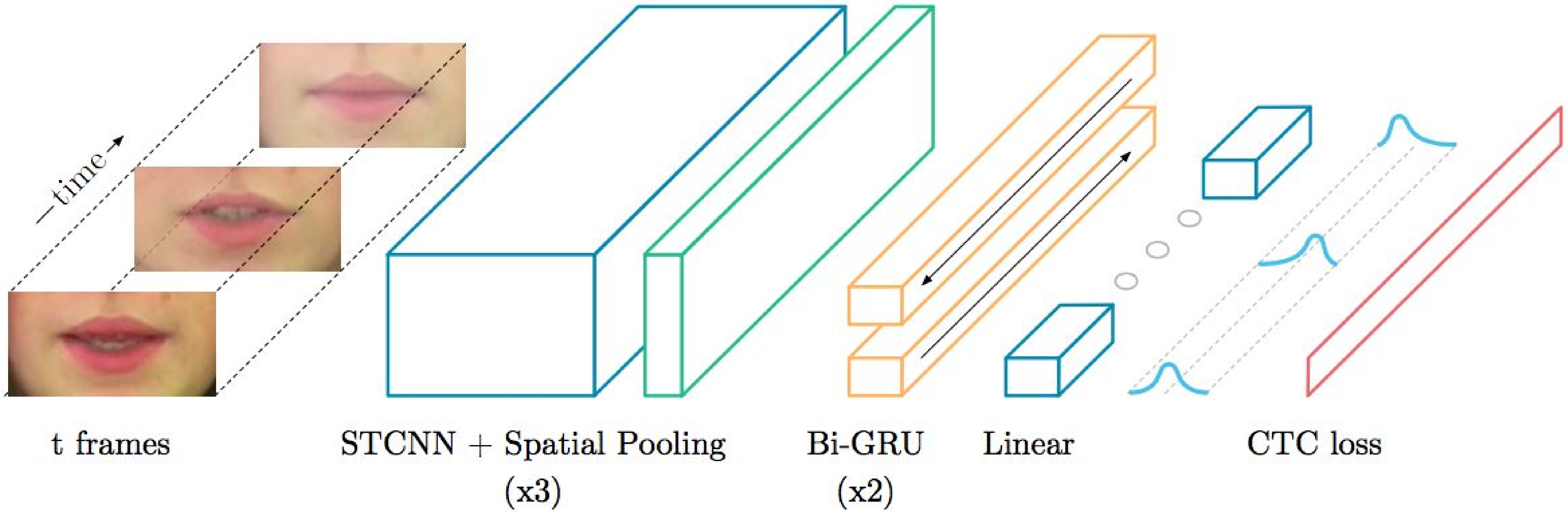


UNIVERSITY OF OXFORD

Preprocessing

- Facial Landmarks
- Crop the mouth
- Affine transform the frames
- Smoothen using Kalman filter
- Temporal augmentation

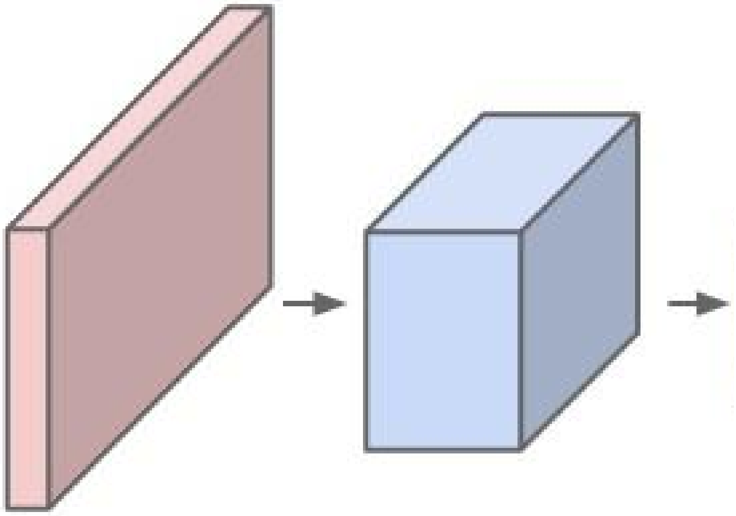
Model Architecture



UNIVERSITY OF OXFORD

Baselines

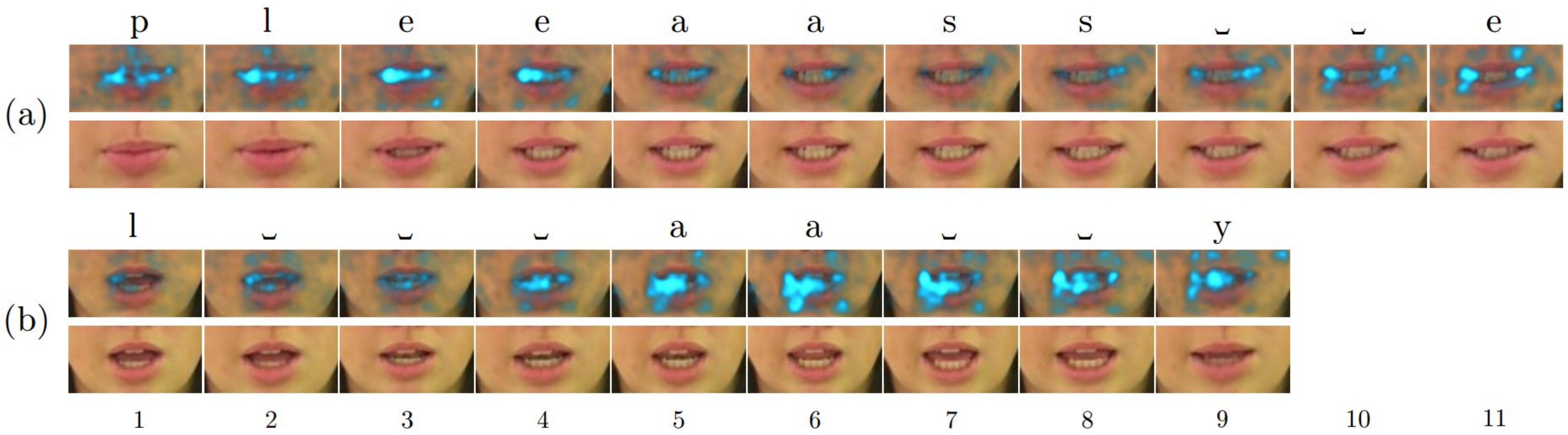
- Hearing-Impaired People
  3 students from the Oxford Students' Disability Community
- Baseline-LSTM
  Replicate previous state-of-the-art architecture by (Wand et al., 2016)
- Baseline-2D
  Spatial-only convolutions
- Baseline-NoLM
  Language model disabled



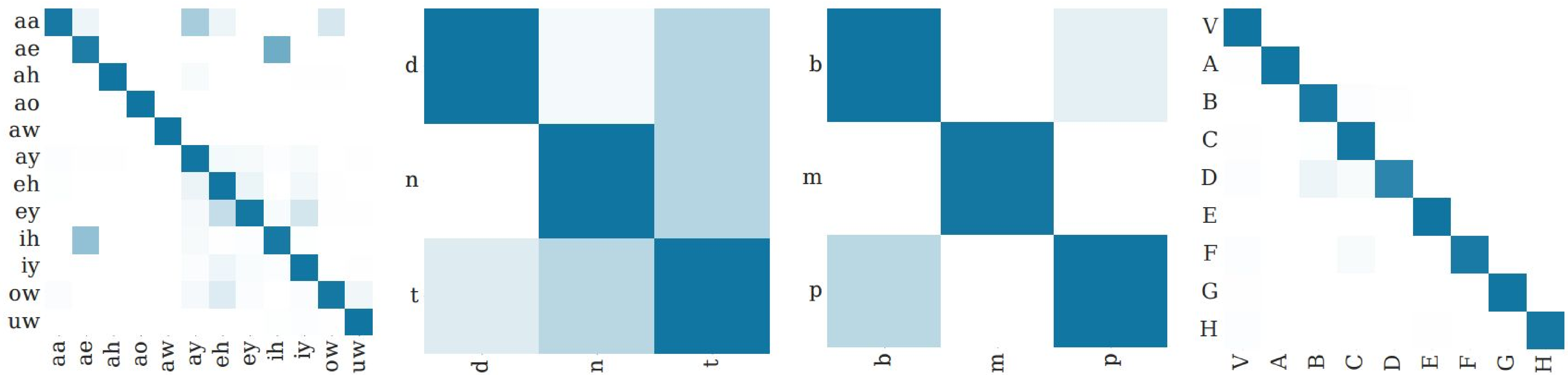
Lipreading Performance

	Unseen Speakers		Overlapped Speakers	
	CER	WER	CER	WER
Hearing Impaired		47.7%		
Baseline-LSTM	38.4%	52.8%	15.2%	26.3%
Baseline-2D	16.2%	26.7%	4.3%	11.6%
Baseline-NoLM	6.7%	13.6%	2.0%	5.6%
LipNet	**6.4%**	**11.4%**	**1.9%**	**4.8%**

Learned Representations



Viseme Confusions



(a) Lip-rounding vowels (b) Alveolar (c) Bilabial (d) Viseme Categories

Thank you!

Thank you NVIDIA!