KU LEUVEN

Ontwerp, analyse en implementatie van een convolutionair neuraal netwerk voor gelijktijdige spraak en beeldherkenning

Overview

1. Problem sketch
2. Current research
3. Neural networks
4. Dataset: TCDTIMIT
5. Objectives
6. Lipreading
7. Speech (audio)
8. Sensor fusion

KU LEUVEN

1. Problem sketch

Speech recognition applications

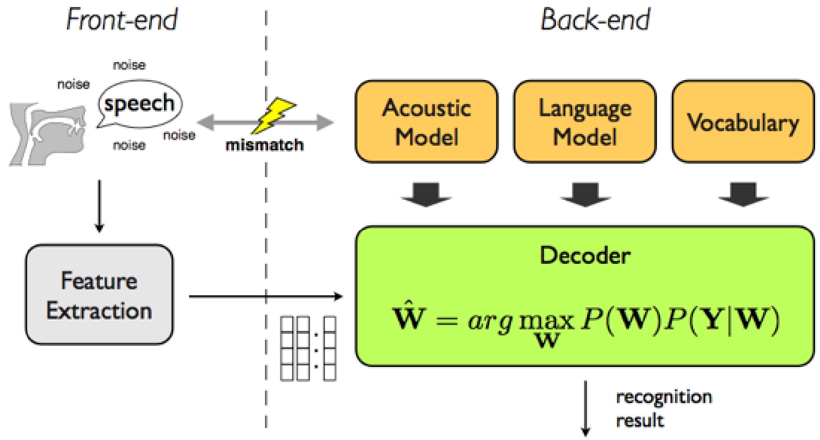
- Automatic subtitles
- Assisting hearing impaired
- Human-computer interation (Siri)
- International meetings (translations)

- Until now: mostly audio
- Use images (lipreading) → robustness, performance

Sony Entertainment

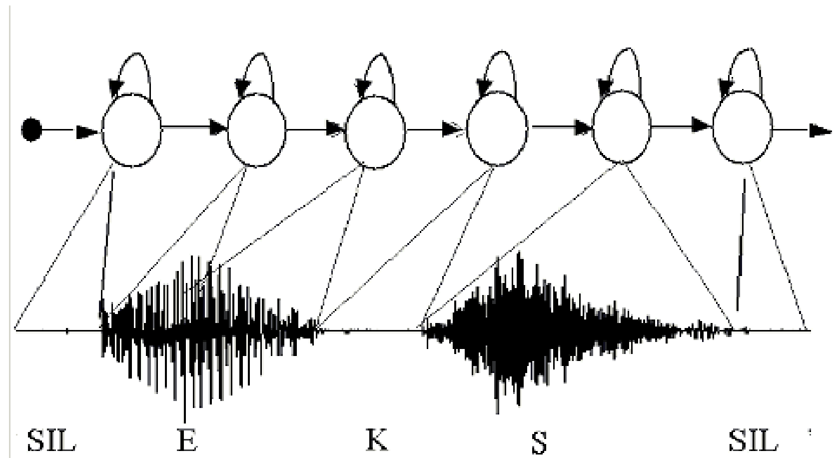
2. Research

- General SR model:



http://www.esat.kuleuven.be/psi/spraak/theses/08-09-en/MDT.php

- In the past:
  - Mostly audio SR
  - Acoustic model: formants, fricatives,...
  - Record sound features;
  - Do statistical correlation
    - → Hidden Markov Models (HMM)
  - Language model on top
  - Often limited in scope (eg. Phone support)



https://www.uea.ac.uk/computing/research-at-the-uea-speech-group

KU LEUVEN

2. Research

- Current: 'Deep learning'
  - Still mostly audio
  - ~~Acoustic model: formants, fricatives,…~~
  - ~~Record sounds, statistical correlation of spectrals~~

    → Convolutional Neural Networks

  - Language model on top (possibly DNN?), or built-in
  - Much broader in scope (Siri, Cortana, SR 'in the wild')

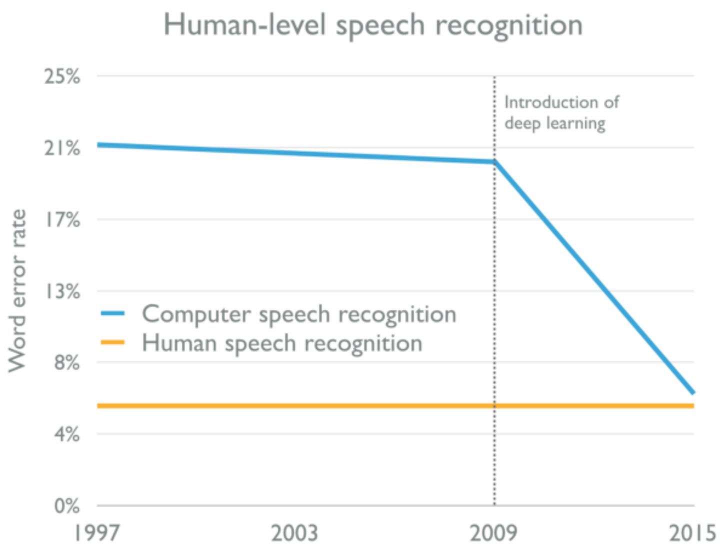
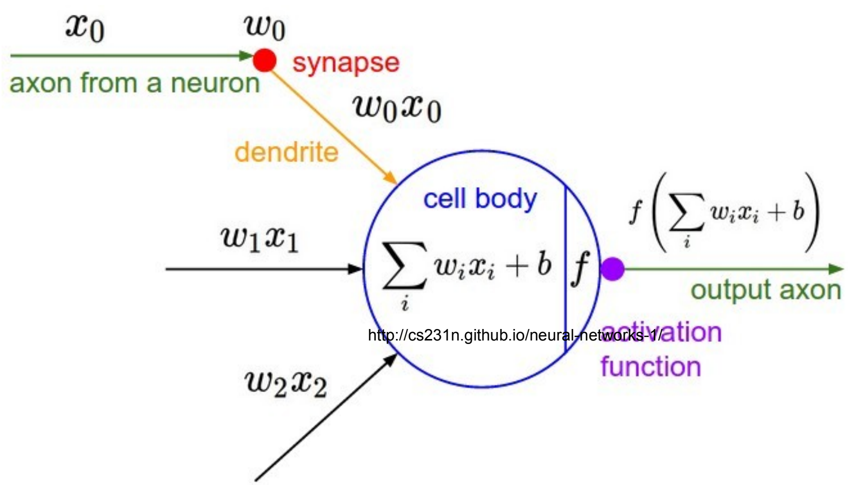


Figure 10

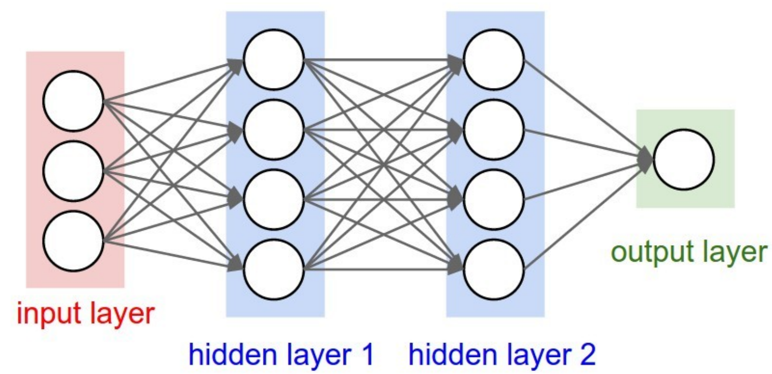
KU LEUVEN

3. Neural Networks

- Simple units with nonlinear output function



http://cs231n.github.io/neural-networks-1/

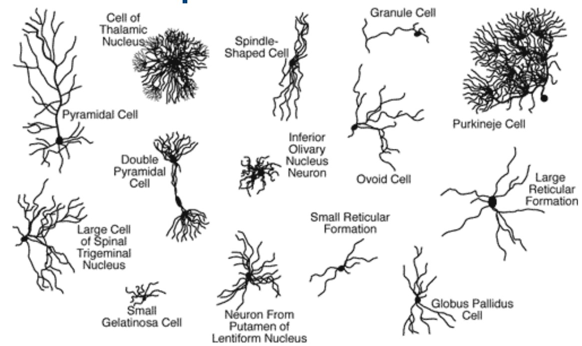


http://cs231n.github.io/neural-networks-1/

KU LEUVEN

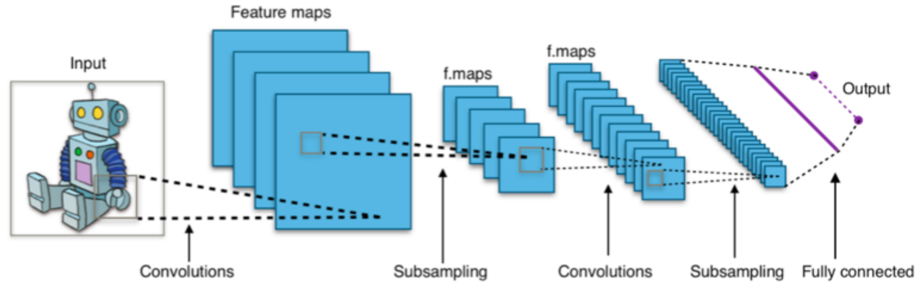
3. Neural Networks: ConvNets

- Goal: Pattern Recognition → high-dimensional input data
- Fully connected Nns don't scale
  - We want to reduce # parameters



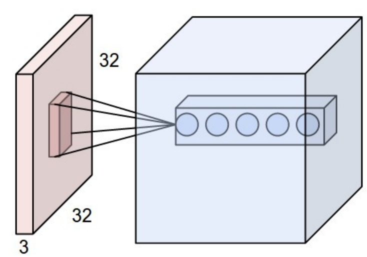
- Brain also uses specialized neurons → **Convolutional Neural Networks**

- Layers in 3D ≈ trainable filters
- parameter sharing + pooling
- Layer types: Conv, ReLu, Pool, FC,...



$$N_{weights} = \sum_{1}^{L} F^2 * C_{i-1} * C_i$$

$$\textit{Activation memory} = \sum_{1}^{L} L_i * W_i * C_i$$

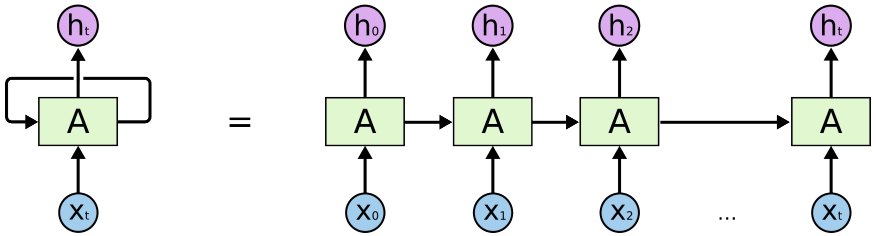



```
INPUT: [224x224x3]        memory:  224*224*3=150K   weights: 0
CONV3-64: [224x224x64]  memory:  224*224*64=3.2M   weights: (3*3*3)*64 = 1,728
CONV3-64: [224x224x64]  memory:  224*224*64=3.2M   weights: (3*3*64)*64 = 36,864
POOL2: [112x112x64]  memory:  112*112*64=800K   weights: 0
```

3. Neural Networks: LSTM

- Goal: add time aspect → memory

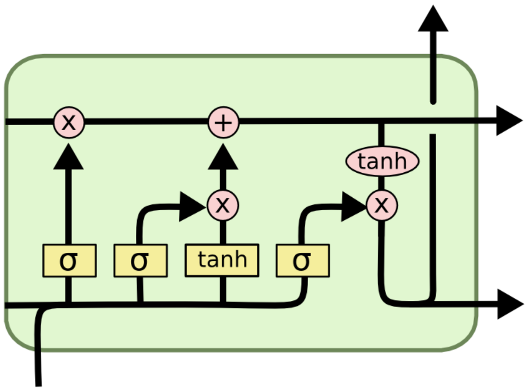
- Add feedback loop
  → Recurrent Neural Networks
- Improved version
  → LSTM Neural Networks



$$W = n_c \times n_c \times 4 + n_i \times n_c \times 4 + n_c \times n_o + n_c \times 3$$

where n_c is the number of memory cells (and number of memory blocks in this case), n_i is the number of input units, and n_o is the number of output units.

Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition
https://arxiv.org/abs/1402.1128



4. Dataset: TCDTIMIT

- Alternatives:
  - GRID: large dataset, but small vocabulary
  - VidTIMIT: small dataset
  - Many non-public databases (Google etc)

- TCDTIMIT:
  - Many speakers, high quality
  - Continuous speech, good coverage of phonemes and visemes. (TIMIT)
  - Available to other researchers.
  - Content:
    - 2255 sentences from TIMIT
    - 59 volunteers (98 sentences each)
    - 3 professional lipspeakers (377 sentences each)
    - ~25 phonemes/sentence
    - Total: 235k phoneme examples; ~ 6k each

Harte, N.; Gillen, E., "TCD-TIMIT: An Audio-Visual Corpus of Continuous Speech," Multimedia, IEEE Transactions on , vol.17, no.5, pp.603,615, May 2015 doi: 10.1109/TMM.2015.2407694

4. Dataset: TCDTIMIT

- Issues downloading & extracting
- Lacking documentation
- Very little support
- Files missing
- After processing:
    - time mismatch phoneme- frame
    - frames missing
    - Other issues

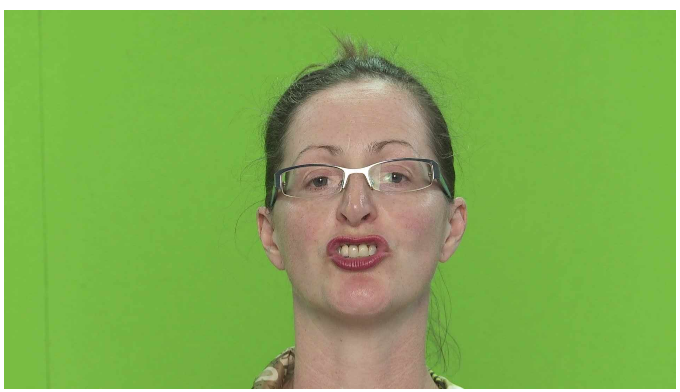
→ write own software to extract data from videos
→ open source for other researchers

KU LEUVEN

4. Dataset: TCDTIMIT

- Goal: labeled frames of phoneme pronounciation
- SW pipeline:
  - Extract phoneme time information
  - Extract frames
  - Remove invalid frames
  - Extract faces, mouths
  - Grayscale and compress
  - Pickle for simple loading in Python

Frame	Phoneme
16	sil
34	sh
37	iy
40	hh
44	ae
45	d
47	y
49	uh
...	...



Sa1.mp4 (60MB)

38 x
sa1_34_sh.jpg (2KB)

4. Dataset: TCDTIMIT

- Train/test/validation set splits:
- For each speaker:
  - 80% training set
  - 10% validation set
  - 10% test set

- Baseline results from database paper (using HMM)

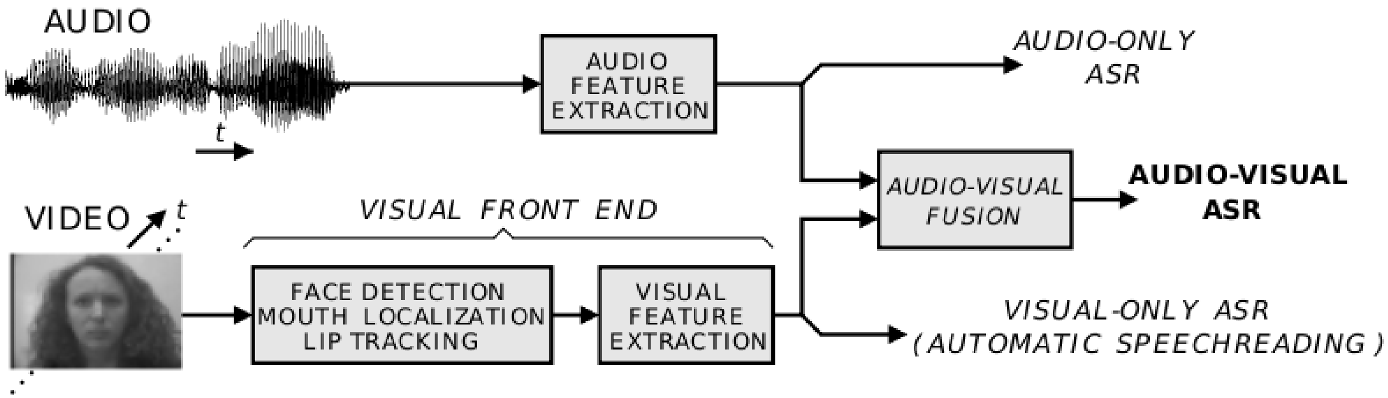
	Split 1 (Table 4.1)		Split 2	
	Train set	Test set	Train set	Test set
%correct	42.33	46.78	41.18	46.97
%accuracy	36.50	34.77	35.53	35.61

Harte, N.; Gillen, E., "TCD-TIMIT: An Audio-Visual Corpus of Continuous Speech," Multimedia, IEEE Transactions on , vol.17, no.5, pp.603,615, May 2015 doi: 10.1109/TMM.2015.2407694

5. Objectives

- Combine lipreading and audio to achieve:
  - Better performance (we use more information)
  - Better robustness (low quality recording, background noise,…)
    - → use best information source available

- Work on phonemes, not words or sentences
  - Simpler; also smaller networks needed
  - Language independent
  - Modularity



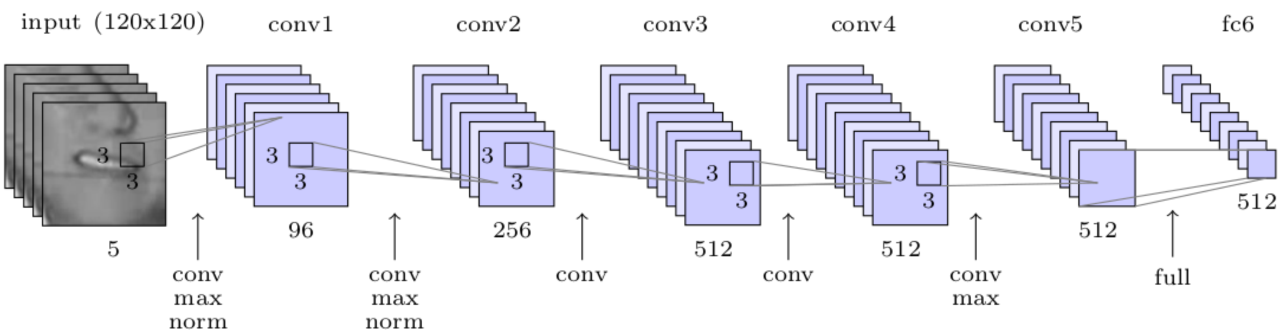
http://ml.sun.ac.za/people/helge-reikeras/

KU LEUVEN

6. Lipreading

- CNN for pattern recognition, then FC

- Networks tested: 1) CIFAR 10, 8 layer network
  2) ResNet 50 layerscifar
  3) Google DeepMind network
- No time-aspect (yet)



Google DeepMind network

6. Lipreading: Phoneme – Viseme map

- Limited correllation lips ↔ sound (aspirated or not,…) → map to visemes
  -> Classification problem: 39 phonemes or 12 visemes
- Networks are trained on phonemes:
  - More general; possibly the CNN can discover more information

Viseme	TIMIT Phonemes	Description	Visibility Rank	Occurrence [%]
/A	/f/ /v/	Lip to Teeth	1	3.15
/B	/er/ /ow/ /r/ /q/ /w/ /uh/ /uw/ /axr/ /ux/	Lips Puckered	2	15.49
/C	/b/ /p/ /m/ /em/	Lip Together	3	5.88
/D	/aw/	Lips Relaxed-Moderate Opening to Lips Puckered-Narrow	4	0.7
/E	/dh/ /th/	Tongue Between Teeth	5	2.9
/F	/ch/ /jh/ /sh/ /zh/	Lips Forward	6	1.2
/G	/oy/ /ao/	Lips Rounded	7	1.81
/H	/s/ /z/	Teeth Approximated	8	4.36
/I	/aa/ /ae/ /ah/ /ay/ /ey/ /ih/ /iy/ /y/ /eh/ /ax-h/ /ax/ /ix/	Lips Relaxed Narrow Opening	9	31.46
/J	/d/ /l/ /n/ /t/ /el/ /nx/ /en/ /dx/	Tongue Up or Down	10	21.1
/K	/g/ /k/ /ng/ /eng/	Tongue Back	11	4.84
/S	/sil/ /pcl/ /tcl/ /kcl/ /bcl/ /dcl/ /gcl/ /h#/ /#h/ /pau/ /epi/	Silence	-	-

6. Lipreading: Google WLAS

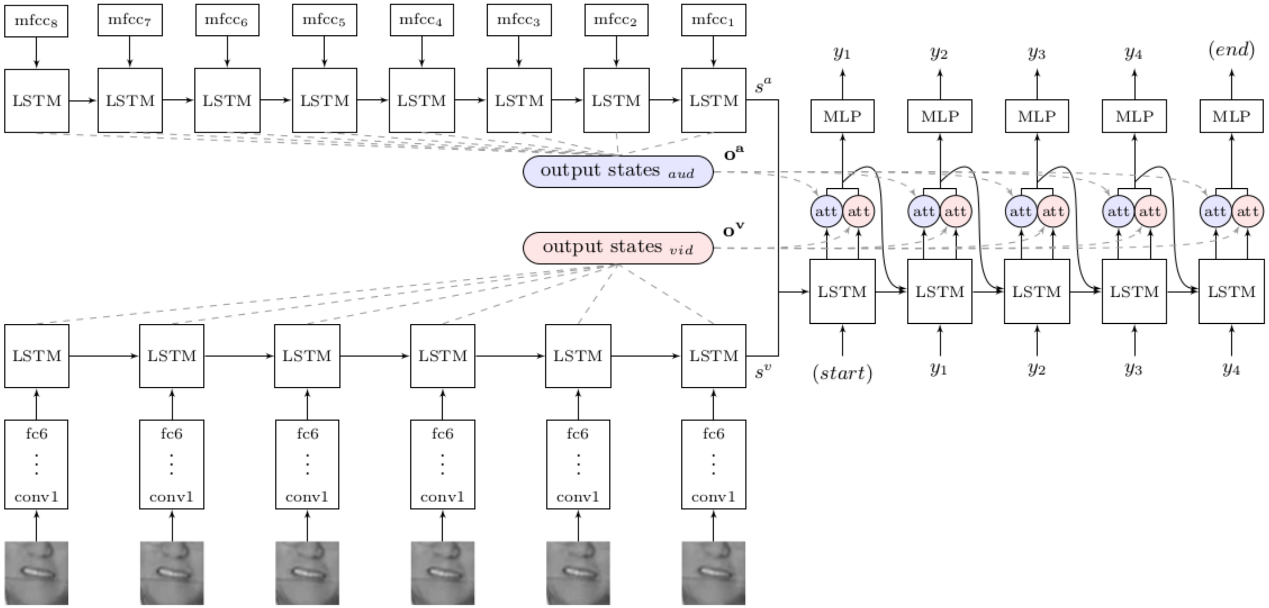
- Nov 2016
- Goal: transcribe videos of mouth motion to characters
- Beats a professional lip reader on videos from BBC television
- Audio and visual parts merged with alignment mechanism

Method	SNR	CER	WER
Lips only			
Professional‡	-	58.7%	73.8%
WAS	-	59.9%	76.5%
WAS+CL	-	47.1%	61.1%
WAS+CL+SS	-	44.2%	59.2%
WAS+CL+SS+BS	-	42.1%	53.2%
Audio only			
LAS+CL+SS+BS	clean	16.2%	26.9%
LAS+CL+SS+BS	10dB	33.7%	48.8%
LAS+CL+SS+BS	0dB	59.0%	74.5%
Audio and lips			
WLAS+CL+SS+BS	clean	13.3%	22.8%
WLAS+CL+SS+BS	10dB	22.8%	35.1%
WLAS+CL+SS+BS	0dB	35.8%	50.8%

KU LEUVEN

6. Lipreading: Google WLAS

$$s^v, \mathbf{o}^v = \texttt{Watch}(\mathbf{x}^v)$$
$$s^a, \mathbf{o}^a = \texttt{Listen}(\mathbf{x}^a)$$
$$P(\mathbf{y}|\mathbf{x}^v, \mathbf{x}^a) = \texttt{Spell}(s^v, s^a, \mathbf{o}^v, \mathbf{o}^a)$$



6. Lipreading:
Phoneme results DeepMind network

- Train and test on lipspeakers:

validation error rate:	57.58%
test error rate:	56.68%

- Train on lipspeakers, test on volunteers:

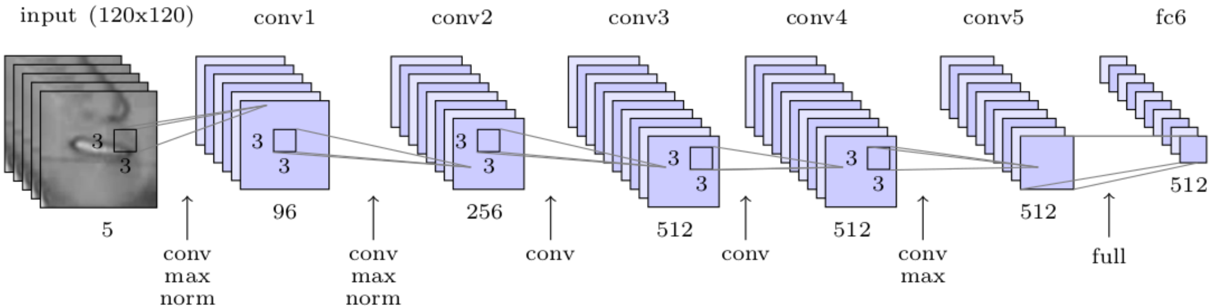
validation error rate:	64.6%
test error rate:	91.6%

- Trained and test on volunteers:

validation error rate:	74.48%
test error rate:	73.53%

- Train on volunteers, test on lipspeakers:

validation error rate:	76.58%
test error rate:	92.68%



- Only 1 (small) FC layer -> not many weights (about 7M)
- Relative performance lipspeaker/volunteer similar to TCDTIMIT paper

6. Lipreading:
Phoneme results CIFAR10

- Train and test on lipspeakers:

validation error rate:	57.40%
test error rate:	58.62%

- Trained and test on volunteers:

validation error rate:	74.48%
test error rate:	72.76%

- Training takes about 2x longer than on Google network (350s/epoch)
- Performance not better
- Some more layers, more parameters -> decent for lipreading

http://www.cs.toronto.edu/~kriz/cifar.html

6. Lipreading:

Phoneme results ResNet50

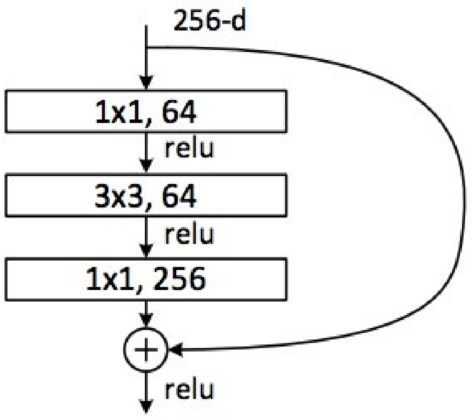
- Train and test on lipspeakers:

validation error rate:	61.95%
test error rate:	62.45%

- Trained and test on volunteers:

validation error rate:	74.48%
test error rate:	72.76%

- Training takes about 5x longer than on Google network (500s/epoch)
- Performance not better
- Many more layers, more complex architecture with more parameters -> not well suited for lipreading



https://github.com/KaimingHe/deep-residual-networks

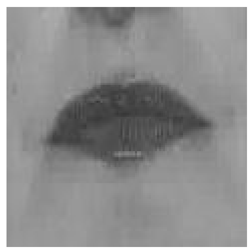
KU LEUVEN

6. Lipreading: example

- Take picture
- Extract face, mouth, convert to grayscale and resize to 120x120x1
- Reshape image for evaluation
- Evaluate, print phoneme predictions (takes 0.2s on laptop)
- Phonemes need extra processing (mapping)

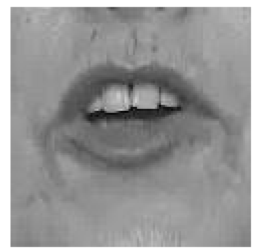
sa1_123_w

```
['w', 0.97530985]
['uw', 0.013956073]
['aa', 0.0087260948]
['r', 0.00060936378]
['ow', 0.00048886862]
['l', 0.00036438892]
['t', 0.0001477778]
['p', 0.00010646239]
['ah', 8.7175431e-05]
```



sa1_179_sil

```
['sil', 0.9990834]
['ah', 0.00034327869]
['v', 0.00027455814]
['m', 0.00016214803]
['f', 5.49916e-05]
['ih', 2.7548622e-05]
['l', 2.5572908e-05]
```



sa1_120_aa

```
['ah', 0.65317434]
['aa', 0.21935698]
['er', 0.027025498]
['sil', 0.021670166]
['r', 0.017895222]
['ow', 0.011971737]
['l', 0.0048446977]
['ay', 0.0012279666]
```

Viseme	TIMIT Phonemes
/A	/f/ /v/
/B	/er/ /ow/ /r/ /q/ /w/ /uh/ /uw/ /axr/ /ux/
/C	/b/ /p/ /m/ /em/
/D	/aw/
/E	/dh/ /th/
/F	/ch/ /jh/ /sh/ /zh/
/G	/oy/ /ao/
/H	/s/ /z/
/I	/aa/ /ae/ /ah/ /ay/ /ey/ /ih/ /iy/ /y/ /eh/ /ax-h/ /ax/ /ix/
/J	/d/ /l/ /n/ /t/ /el/ /nx/ /en/ /dx/
/K	/g/ /k/ /ng/ /eng/
/S	/sil/ /pcl/ /tcl/ /kcl/ /bcl/ /dcl/ /gcl/ /h#/ /#h/ /pau/ /epi/

KU LEUVEN

6. Lipreading: Phoneme – Viseme map

- Limited correllation lips ↔ sound (aspirated or not,…) → map to visemes
  -> Classification problem: 39 phonemes or 12 visemes

- Training DeepMind network with viseme labels -> much improved scores

On lipspeakers

validation error rate:	36.01%
test error rate:	37.41%

On volunteers

validation error rate:	50.01%
test error rate:	52.41%

- Compared to previous result (phonemes):

validation error rate:	57.58%
test error rate:	56.68%

validation error rate:	74.48%
test error rate:	73.53%

- Remark:
  Phoneme result will be improved
  after processing/combining with speech (todo)

KU LEUVEN

Lipreading: binaryNets

- BinaryNet uses binary (+/- 1) weights
  -> efficient HW implementation possible
- Training has to happen with full precision for gradients

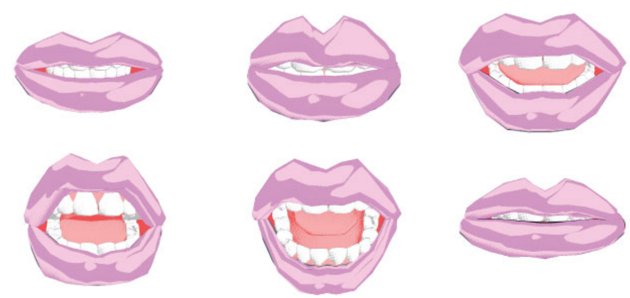
```
Epoch 63 of 500 took 168.815496922s
  LR:                           0.000671220654262
  training loss:                12411.3772716
  validation loss:              24672.7545517
  validation error rate:        87.2159090909%
  best epoch:                   63
  best validation error rate:   87.2159090909%
  test loss:                    23885.5305231
  test error rate:              87.7840909091%
```

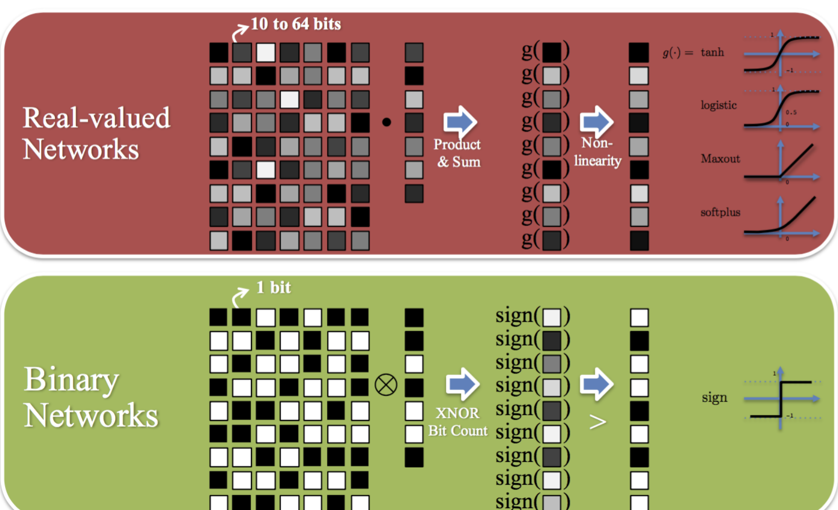
validation error rate:	87.22%
test error rate:	87.78%

- Very high error rates -> more training needed

Next semester

- Goal: robust phoneme recognition using both image and sound
- Audio SR
- Sensor fusion
- HW efficiency -> train networks with:
  - Binary weights
  - Fixed- point weights

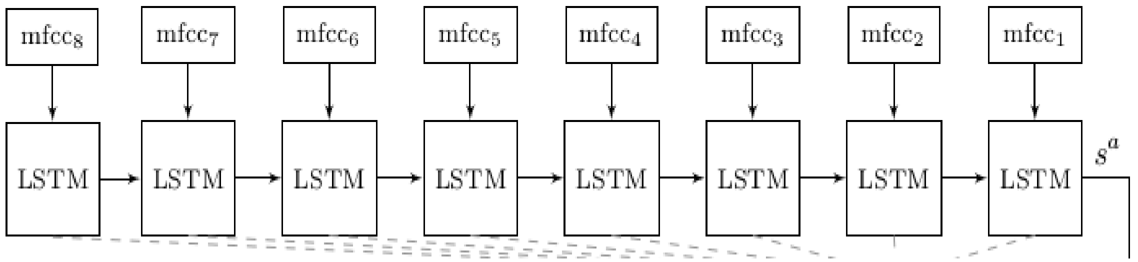


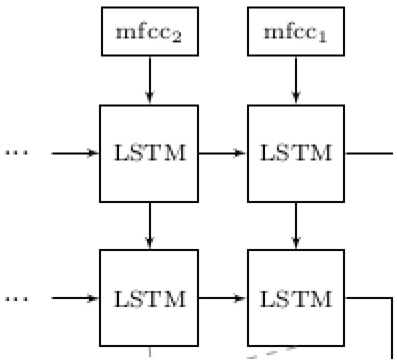


KU LEUVEN

7. Audio SR

- Two-layer LSTM architecture, MFCC as input
- Train with noise to make more robust
- Two layer LSTM





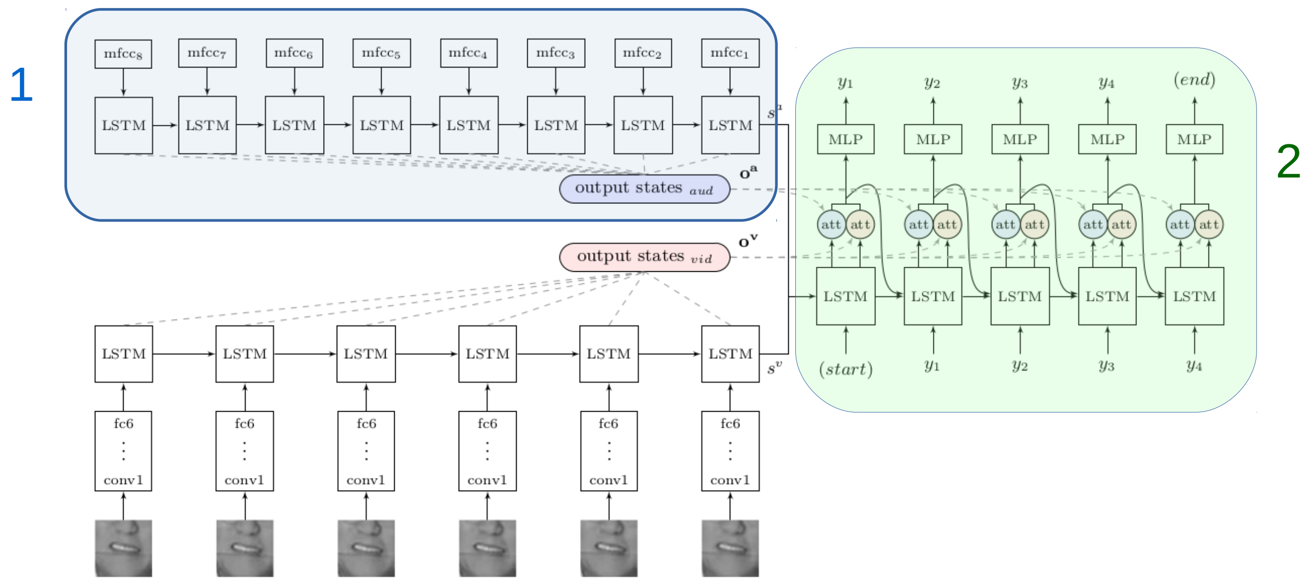
KU LEUVEN

8. Fusing audio and visual

- SR: inherent time aspect
- Lipreading: mostly time-independent, could benefit from limited time aspect
- Audio and video synchronized thanks to labeled dataset

  -> possible to combine feature vectors

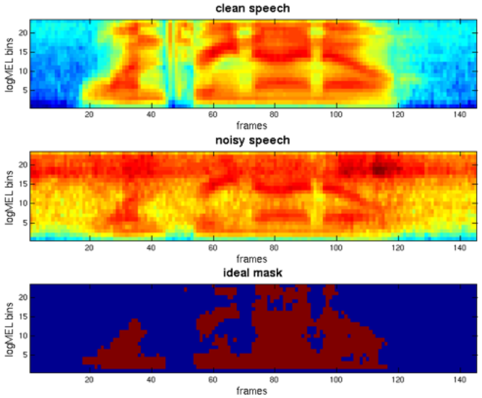
  -> LSTM for audio (1), LSTM for feature fusing (2)



KU LEUVEN

8. Fusing audio and visual

- 'Late fusion': combine output sequences (weighting)
- Weighting determined by:
  - performance of seperate models
  - S/N of audio
  - Quality of video/image
  - ...



http://www.esat.kuleuven.be/psi/spraak/theses/08-09-en/MDT.php

- Analyse performance:
  - Different amounts of audio and/or image noise
  - Comparison audio only/visual only/ audio-visual

KU LEUVEN

Questions