Representations of language in a model of visually grounded speech signal

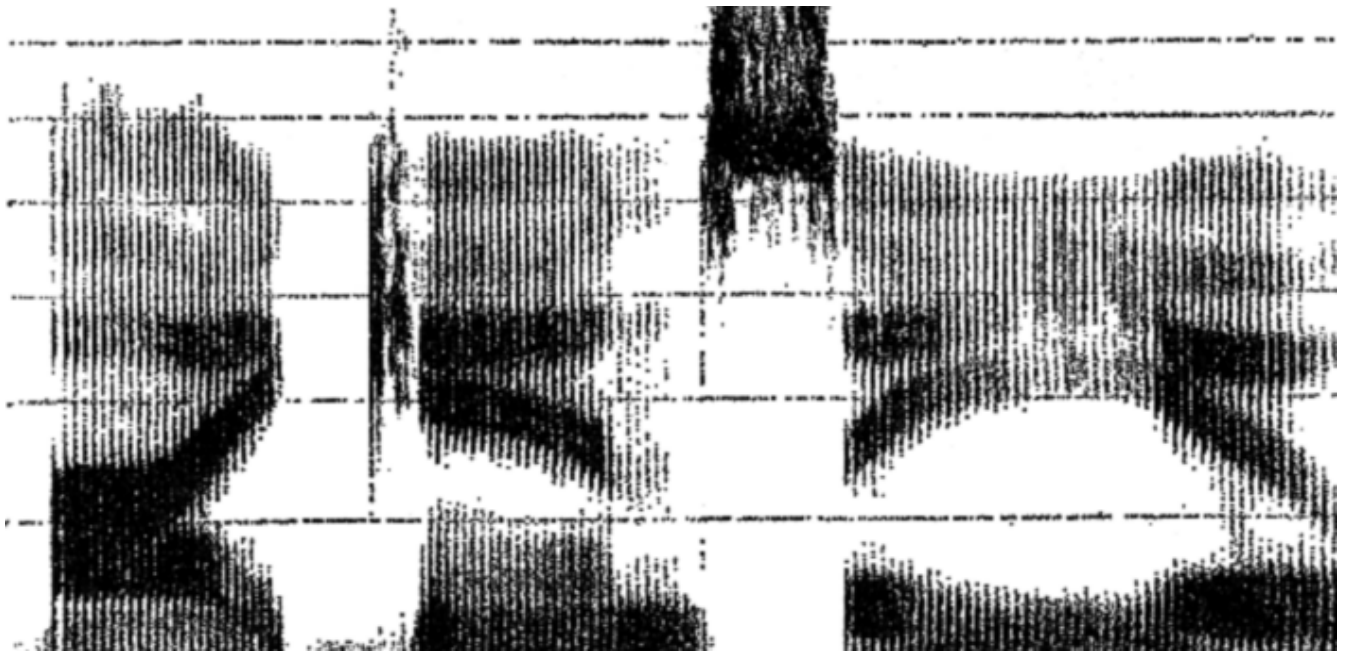
Grzegorz Chrupała Lieke Gelderloos Afra Alishahi

TILBURG UNIVERSITY

Automatic Speech Recognition

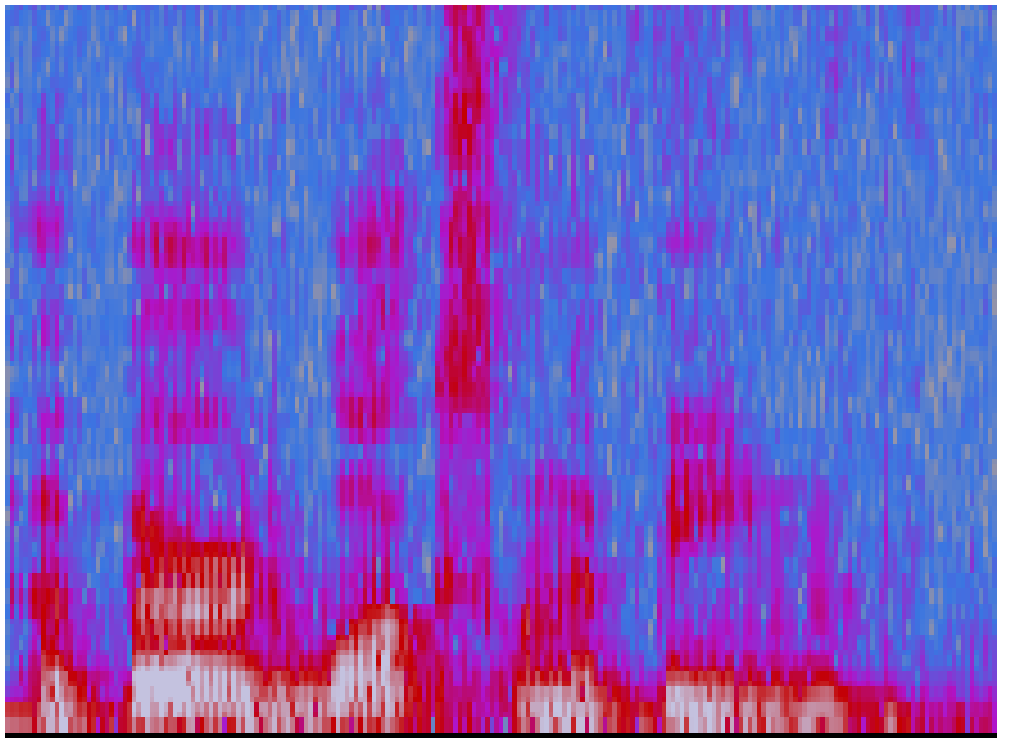
A major commercial success story in Language Technology

Very heavy-handed supervision



I can see you

Grounded speech perception



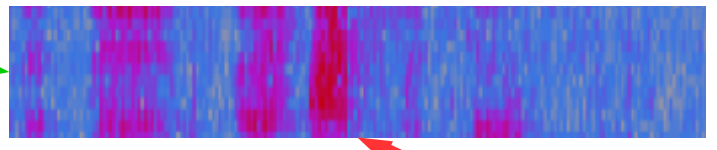
Data

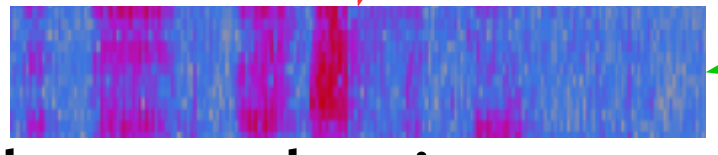
- Flickr8K Audio (Harwath & Glass 2015)
  - 8K images, five audio captions each
- MS COCO Synthetic Spoken Captions

  - 300K images, five synthetically spoken captions each

Project speech and image to joint space

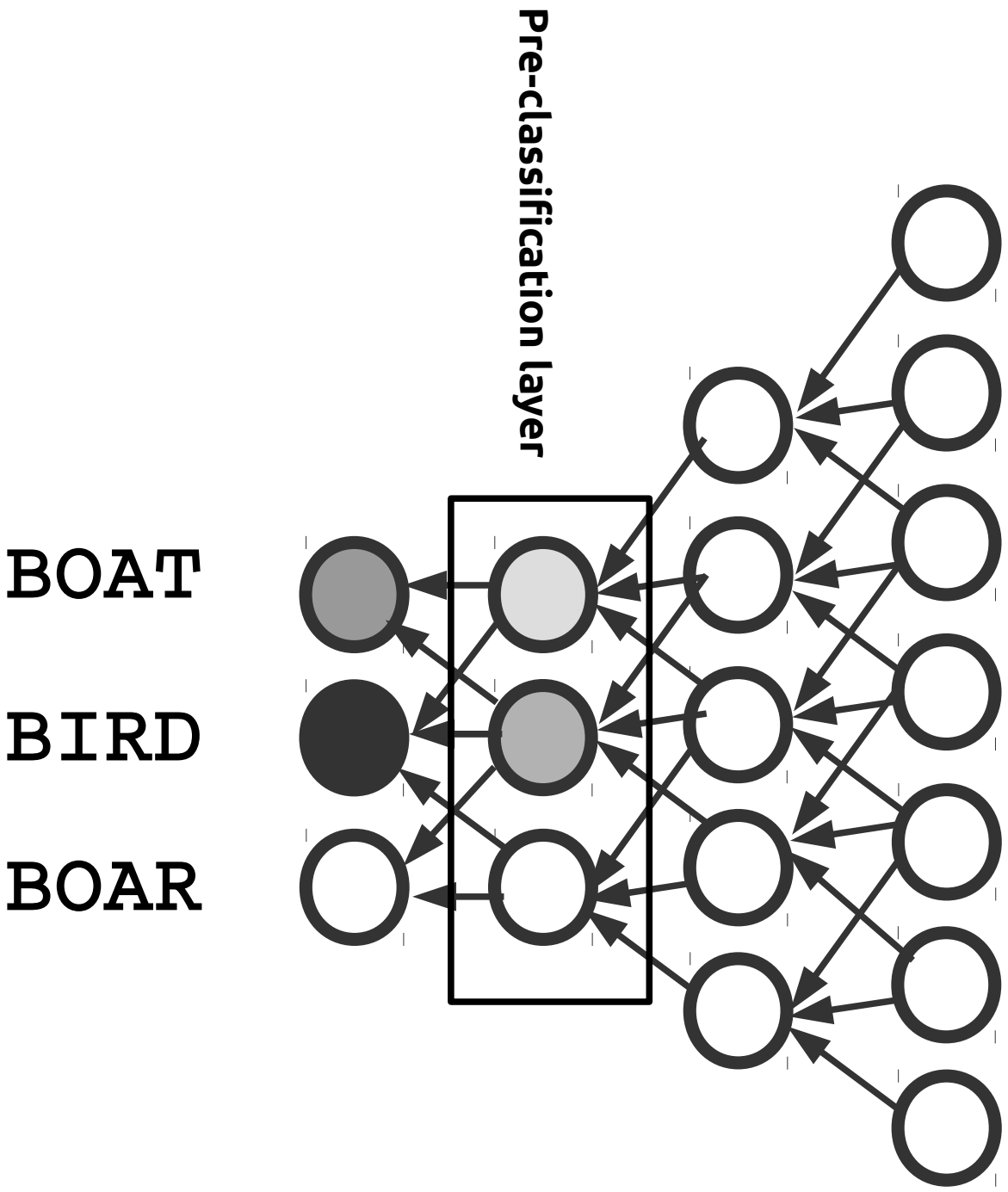
a bird walks on a beam





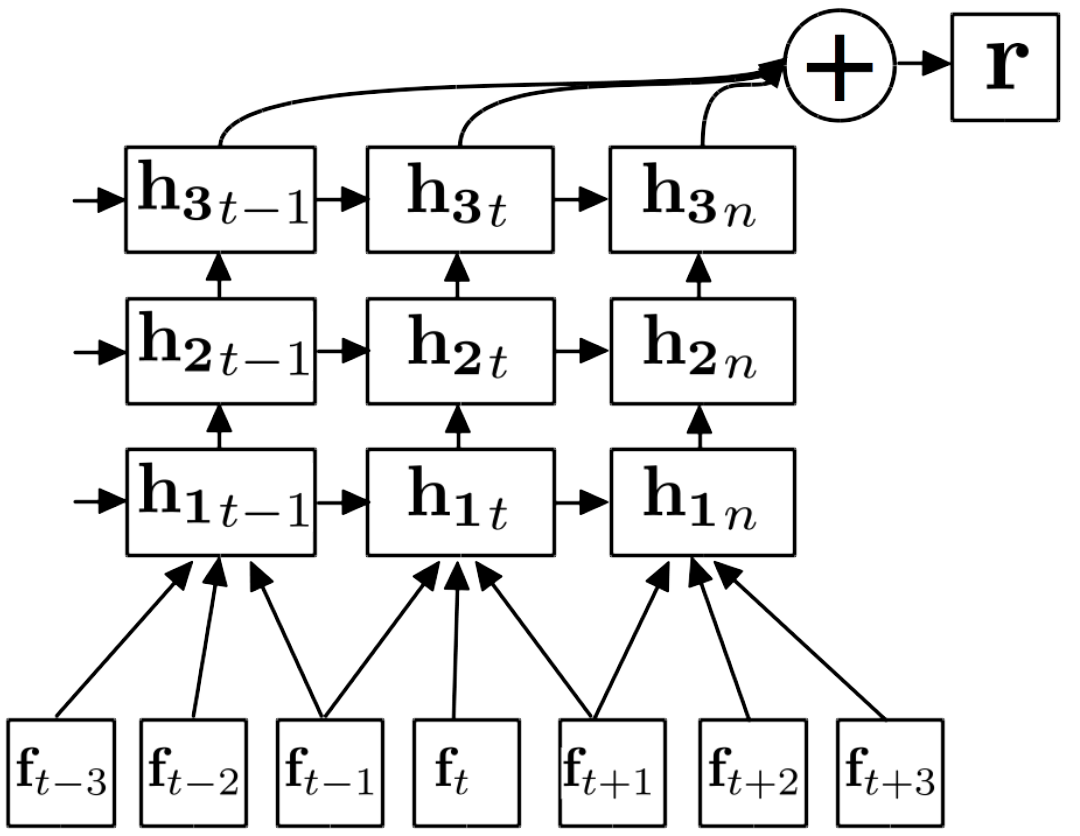
bears play in water

Image model



Speech model

- Input: **MFCC**
- Subsampling CNN
- Recurrent Highway Network (Zilly et al 2016)
- Attention



Model settings

Flickr8K Speech
Attention 128
RHN depth 2, 1024
RHN depth 2, 1024
RHN depth 2, 1024
Conv 6x64, stride 2

COCO Speech
Attention 512
RHN depth 2, 512
RHN depth 2, 512
RHN depth 2, 512
RHN depth 2, 512
Conv 6x64, stride 3

Flickr8K Text
RHN depth 1, 1024
Embedding 300

COCO Text
RHN depth 1, 1024
Embedding 300

Image retrieval

Flickr8K

Model	R@10	$\tilde{r}$
Speech $\text{RHN}_{4,2}$	0.253	48
Harwath & Glass 2015	0.179	-
Text $\text{RHN}_{1,1}$	0.494	11

MSCOCO

Model	R@10	$\tilde{r}$
Speech $\text{RHN}_{5,2}$	0.444	13
Text $\text{RHN}_{1,1}$	0.565	8

Newer CNN architecture: Harwath et al 2016 (NIPS), Harwath and Glass 2017 (ACL)

Levels of representation

- What aspects of sentences are encoded?
- Which layers encode form, which encode meaning?
- Auxiliary tasks (Adi et al 2017)

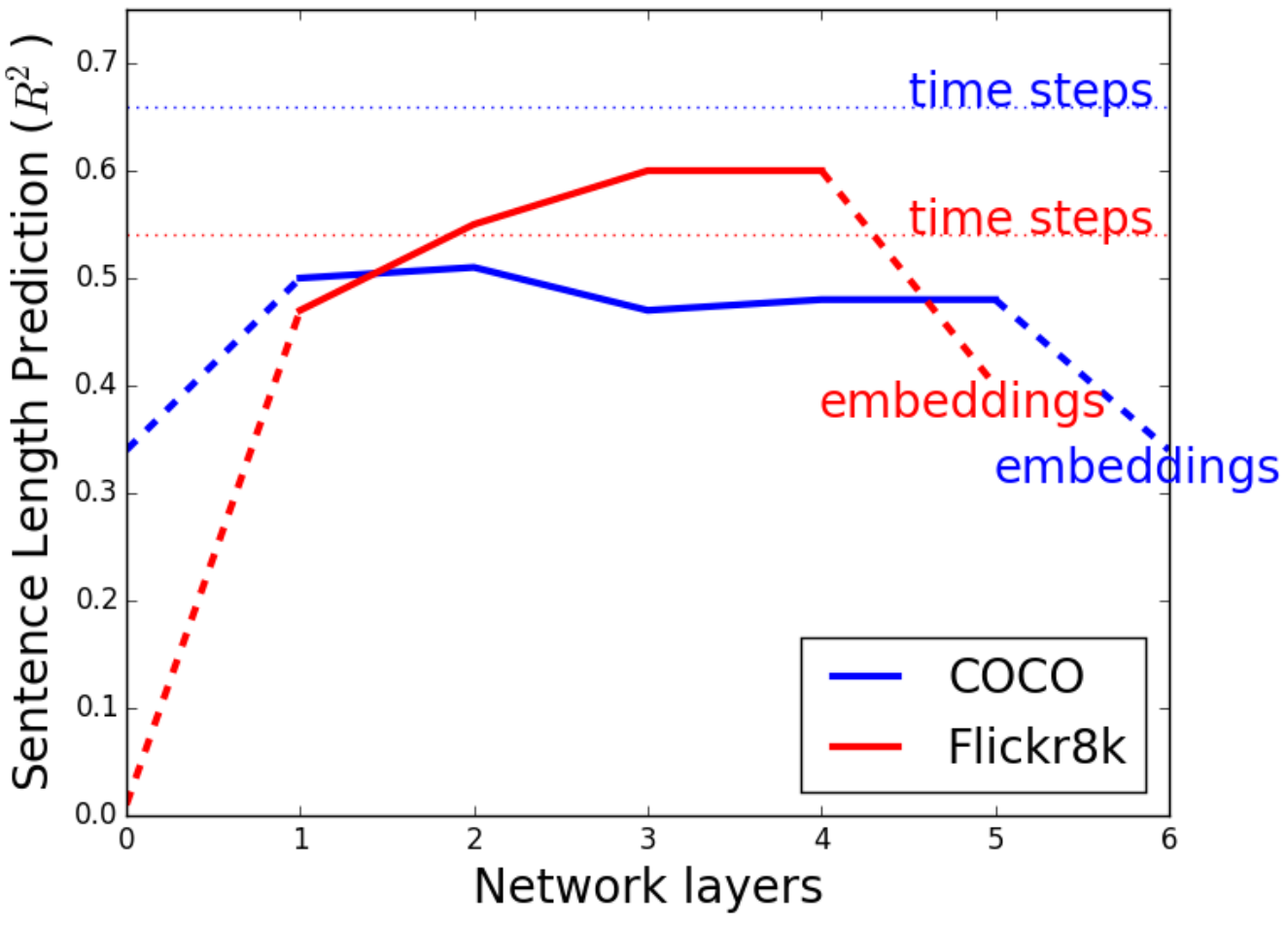
Form-related aspects

Use activation vectors to decode

- Utterance length in words
- Presence of specific words

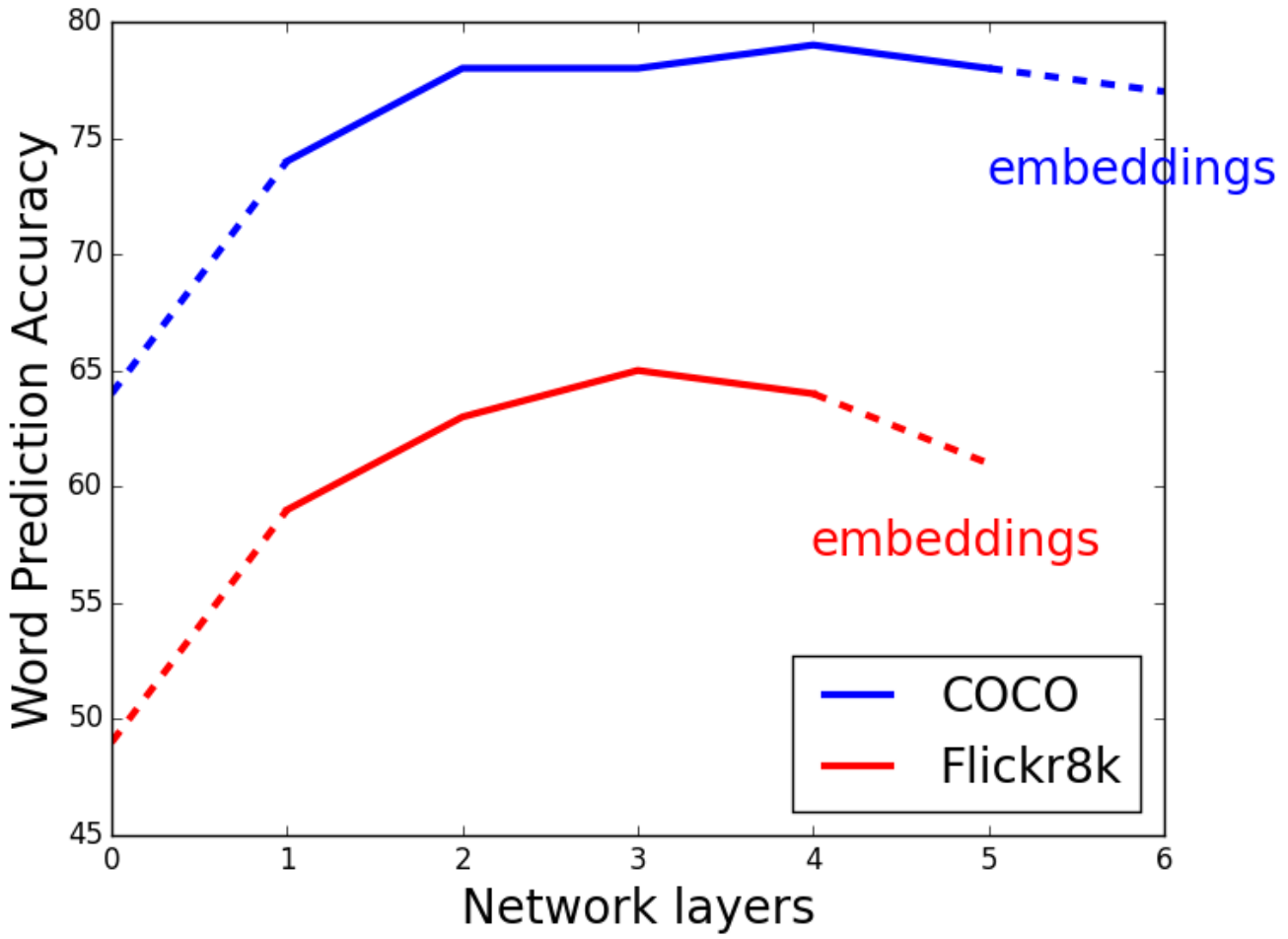
Number of words

- Input
  - Activations for utterance
- Model
  - Linear regression



Word presence

- Input
  - Activations for utterance
    - MFCC for word
- Model
  - MLP



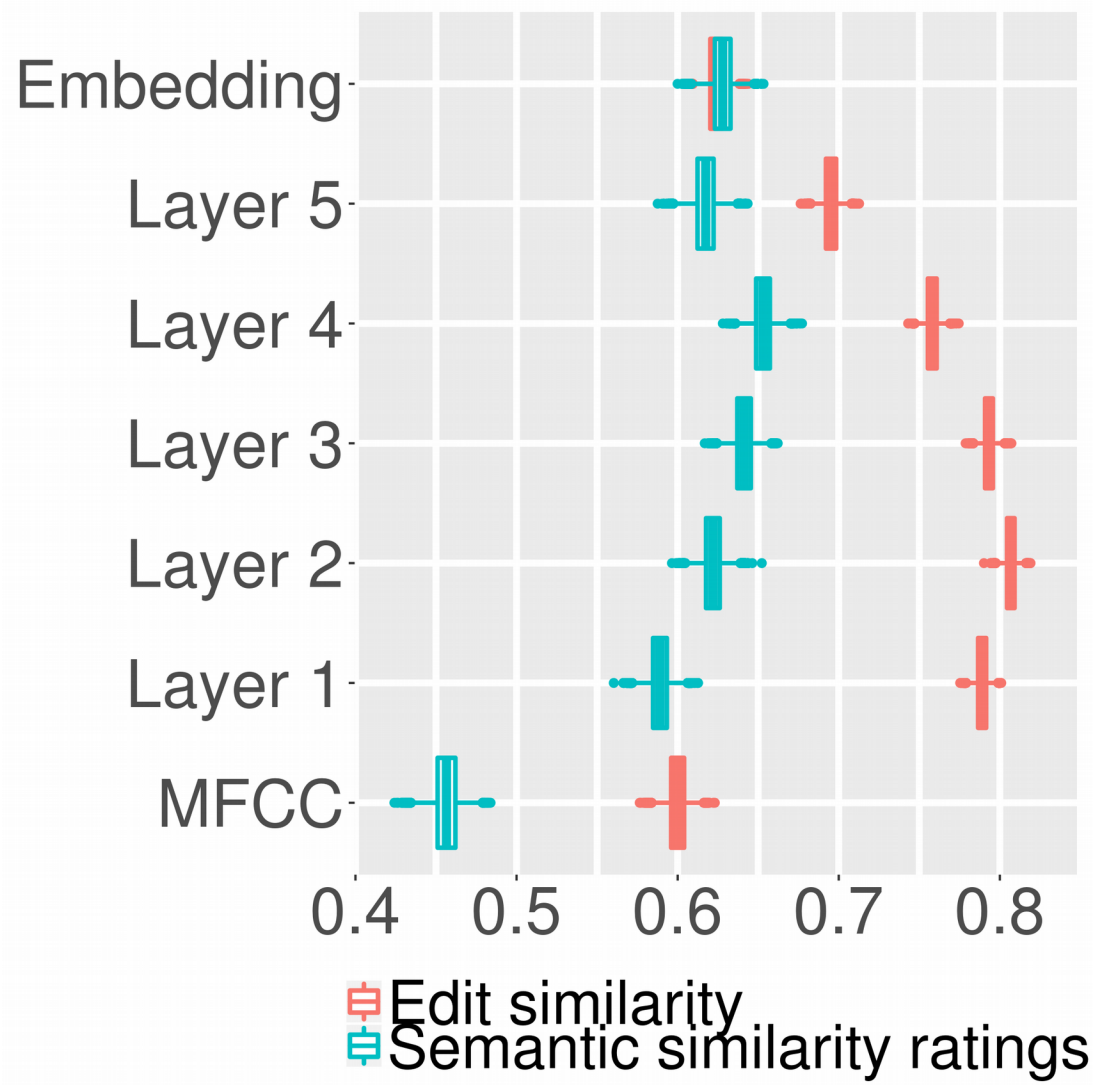
Semantic aspects

Representational Similarity

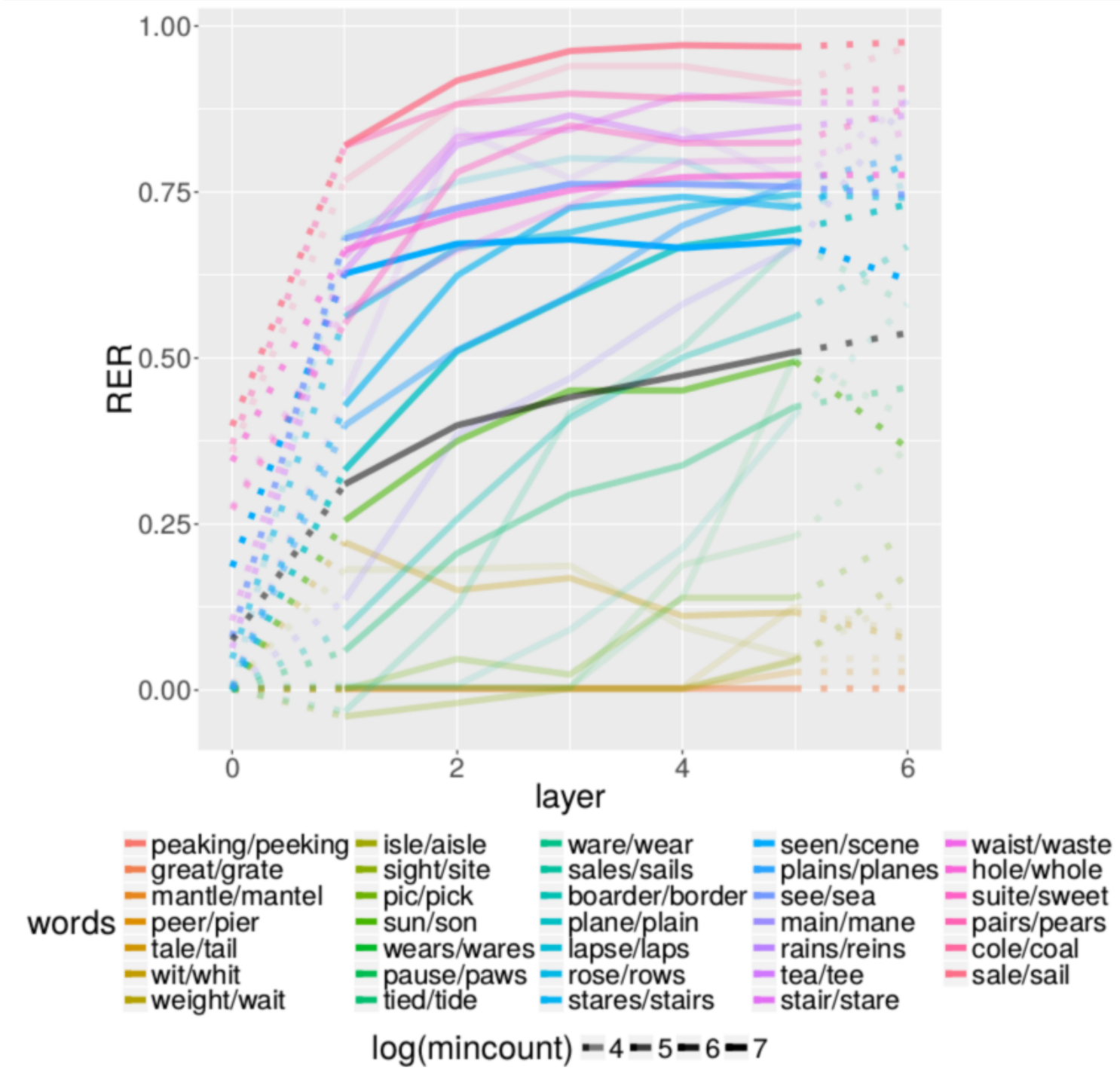
- Correlations between sets of pairwise similarities according to
  - Activations

    AND
  - Edit ops on written sentences
  - Human judgments

(SICK dataset)



Homonym disambiguation



Follow-up work

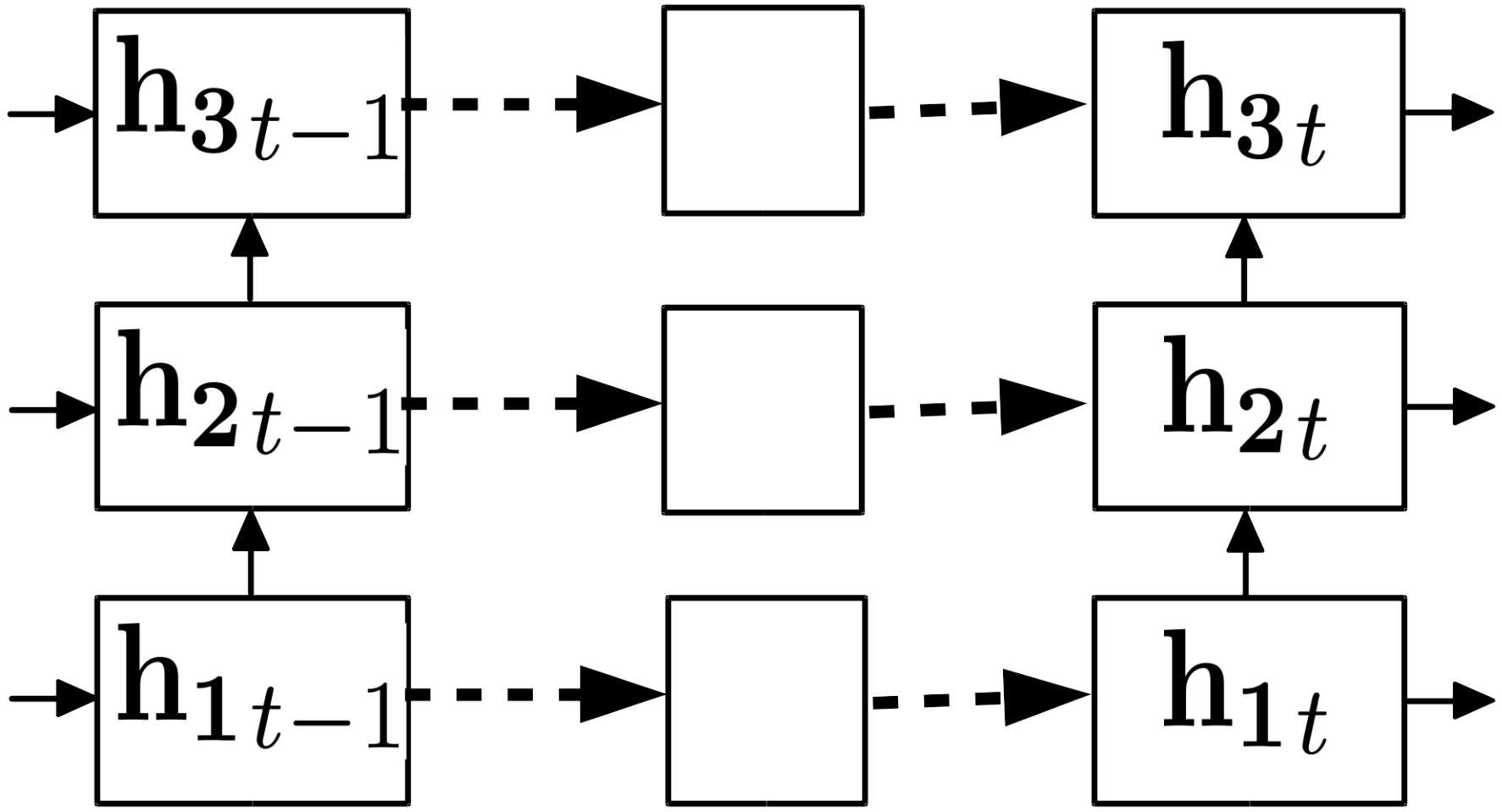
Afra Alishahi, Marie Barking and Grzegorz Chrupała. Encoding of phonology in a recurrent neural model of grounded speech

Friday, session #4 at CoNLL

Conclusion

Encodings of form and meaning emerge and evolve in hidden layers of stacked RHN listening to grounded speech

Code: github.com/gchrupala/visually-grounded-speech
Data: doi.org/10.5281/zenodo.400926


$\mathbf{h}_{3\,t-1}$
$\mathbf{h}_{3\,t}$
$\mathbf{h}_{2\,t-1}$
$\mathbf{h}_{2\,t}$
$\mathbf{h}_{1\,t-1}$
$\mathbf{h}_{1\,t}$

Error analysis

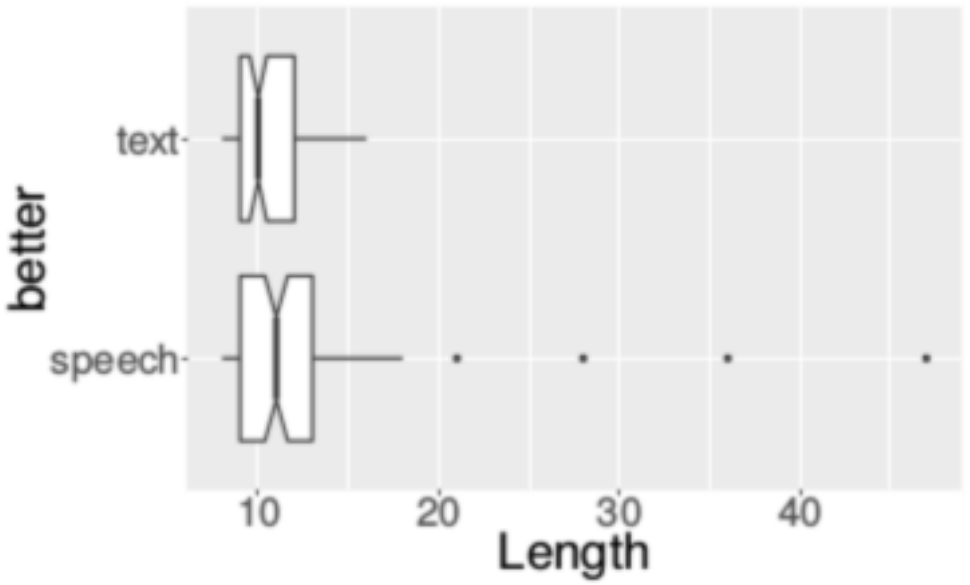
- Text usually better
- Speech better:
  - Long descriptions
  - Misspellings

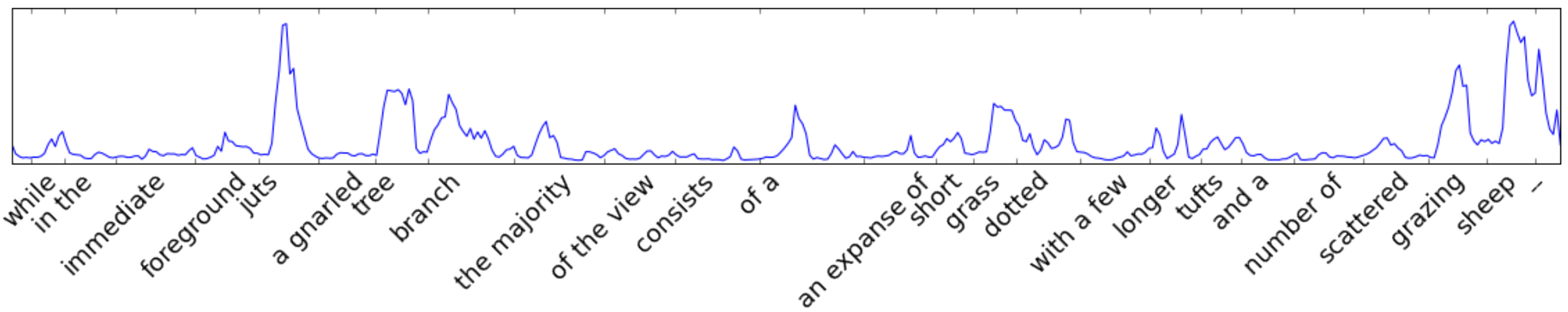
a yellow and white birtd is in flight

Text

Speech

Length


text
speech
10
20
30
40
Length


while
in the
immediate
foreground
juts
a gnarled
tree
branch
the majority
of the view
consists
of a
an expanse of
short
grass
dotted
with a few
longer
tufts
and a
number of
scattered
grazing
sheep
–

Text model

- Convolution → word embedding
- No attention