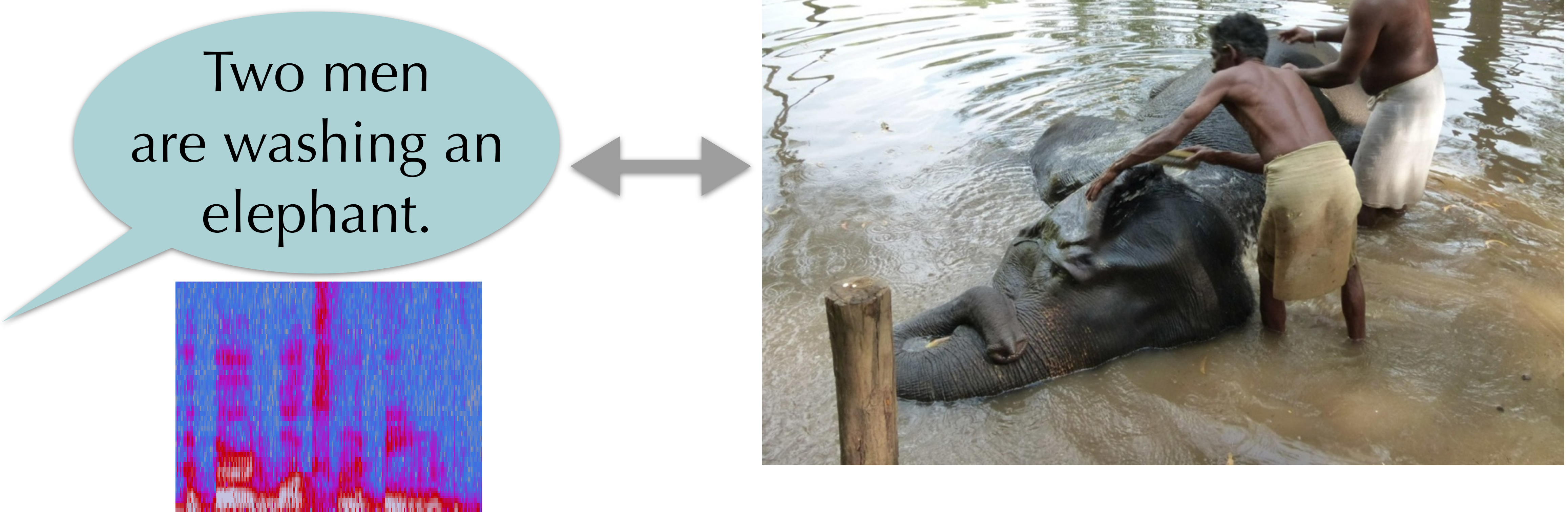
Encoding of Phonology in an RNN model of Grounded Speech

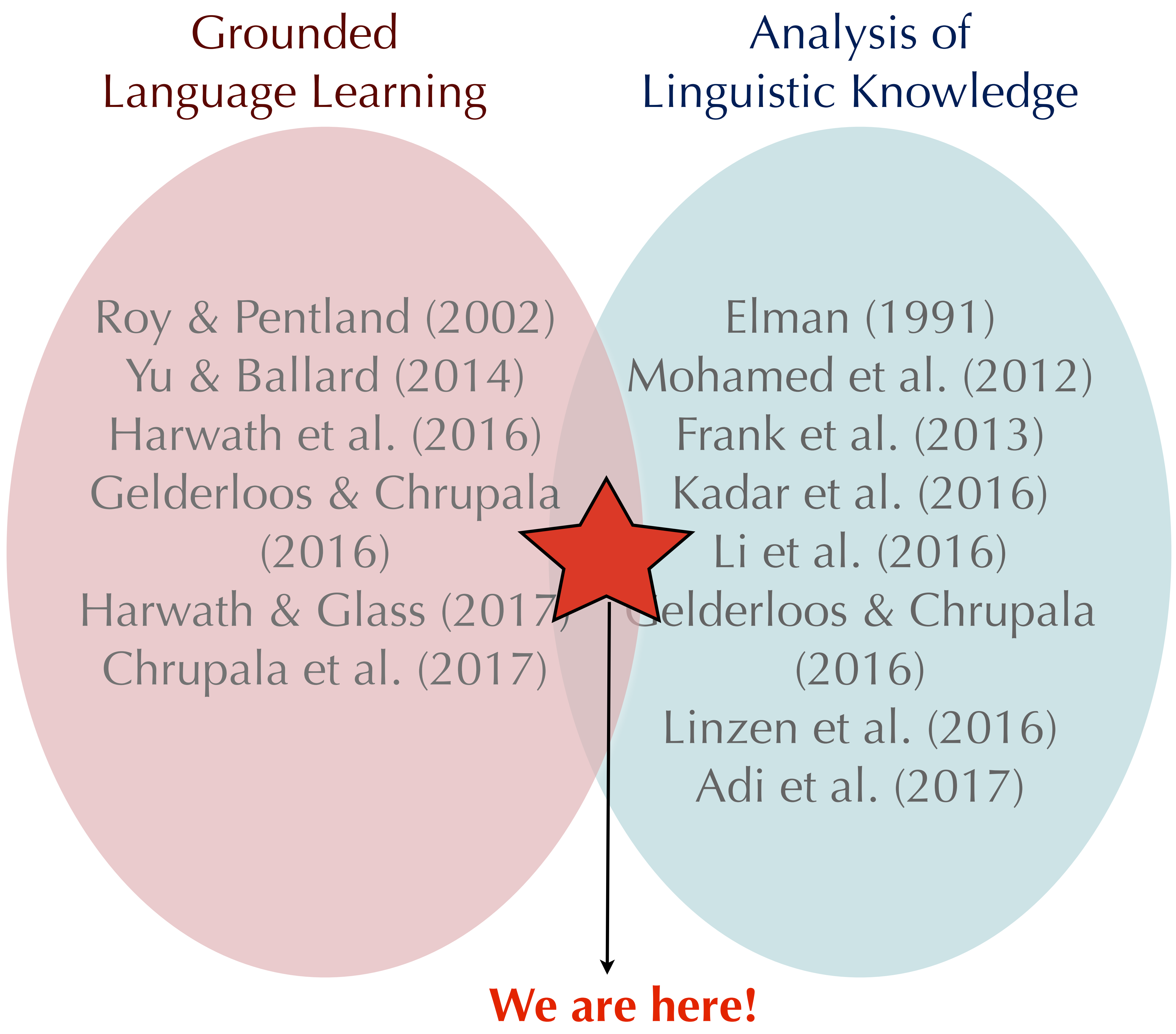
Afra Alishahi, Marie Barking, Grzegorz Chrupała

TiCC Tilburg Centre for Cognition and Communication
Tilburg University

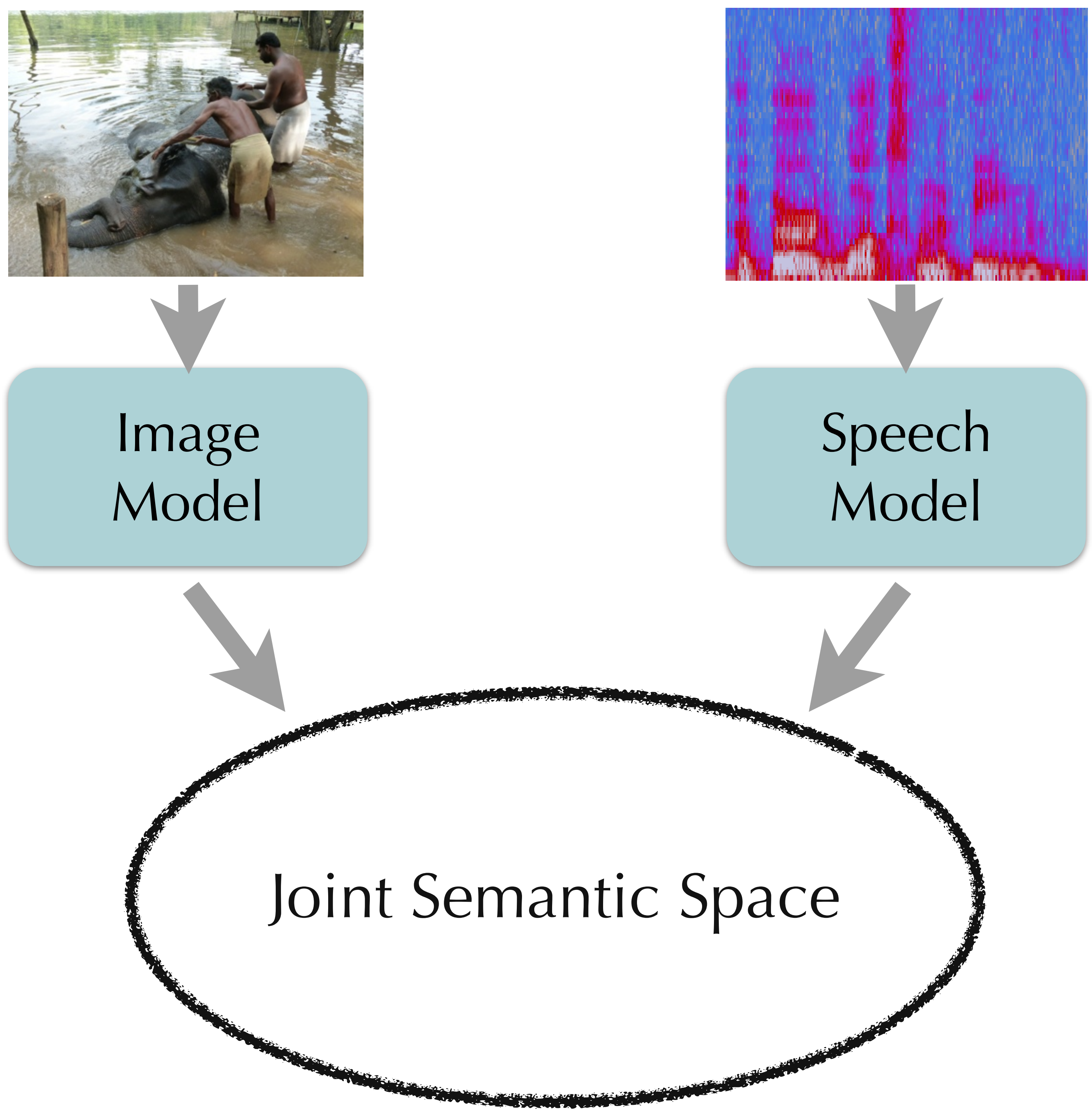


A Realistic Language Learning Scenario




Grounded
Language Learning
Analysis of
Linguistic Knowledge
Roy & Pentland (2002)
Yu & Ballard (2014)
Harwath et al. (2016)
Gelderloos & Chrupala
(2016)
Harwath & Glass (2017)
Chrupala et al. (2017)
Elman (1991)
Mohamed et al. (2012)
Frank et al. (2013)
Kadar et al. (2016)
Li et al. (2016)
Gelderloos & Chrupala
(2016)
Linzen et al. (2016)
Adi et al. (2017)
We are here!

A Model of Grounded Speech Perception



Joint Semantic Space

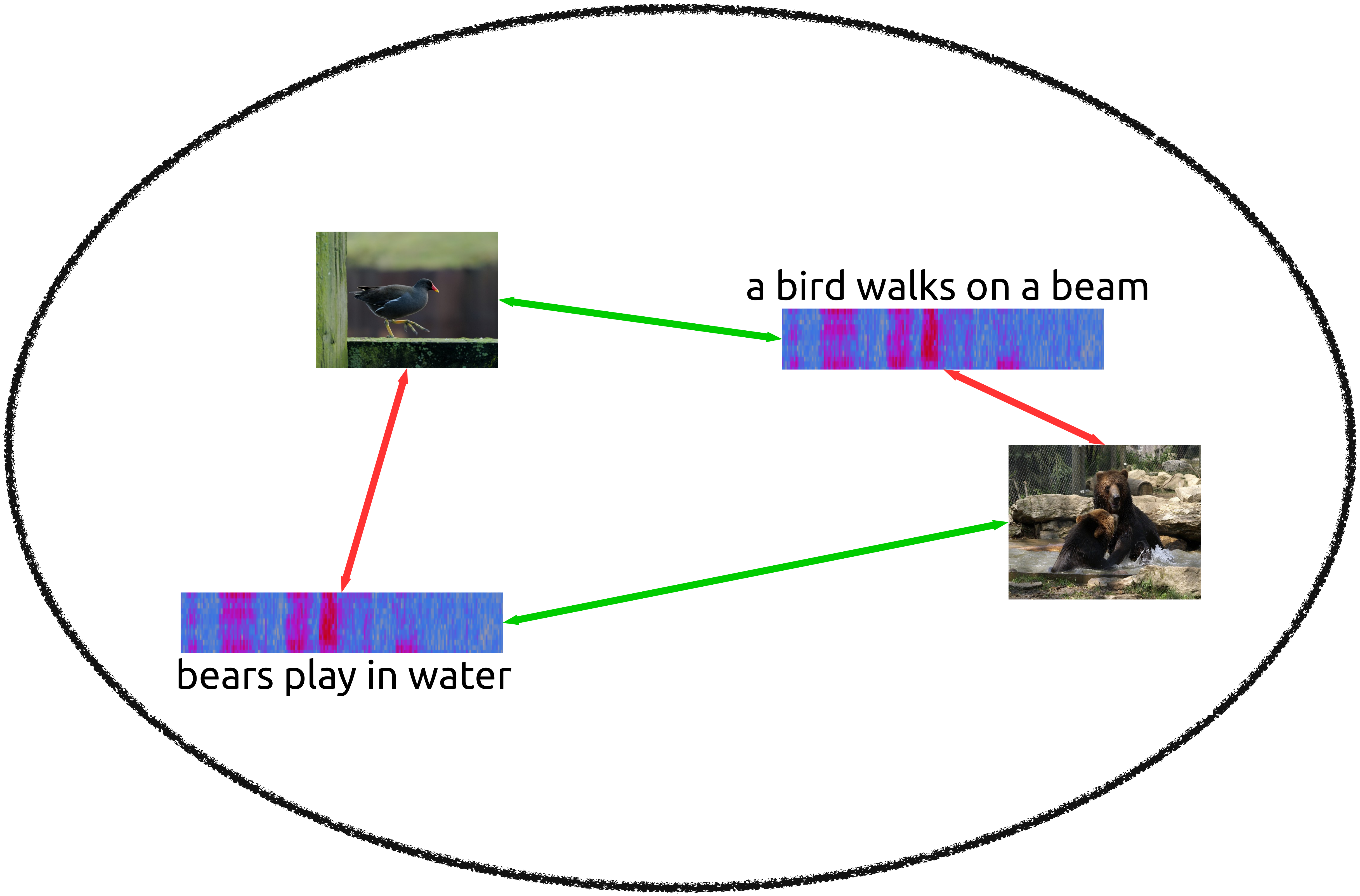
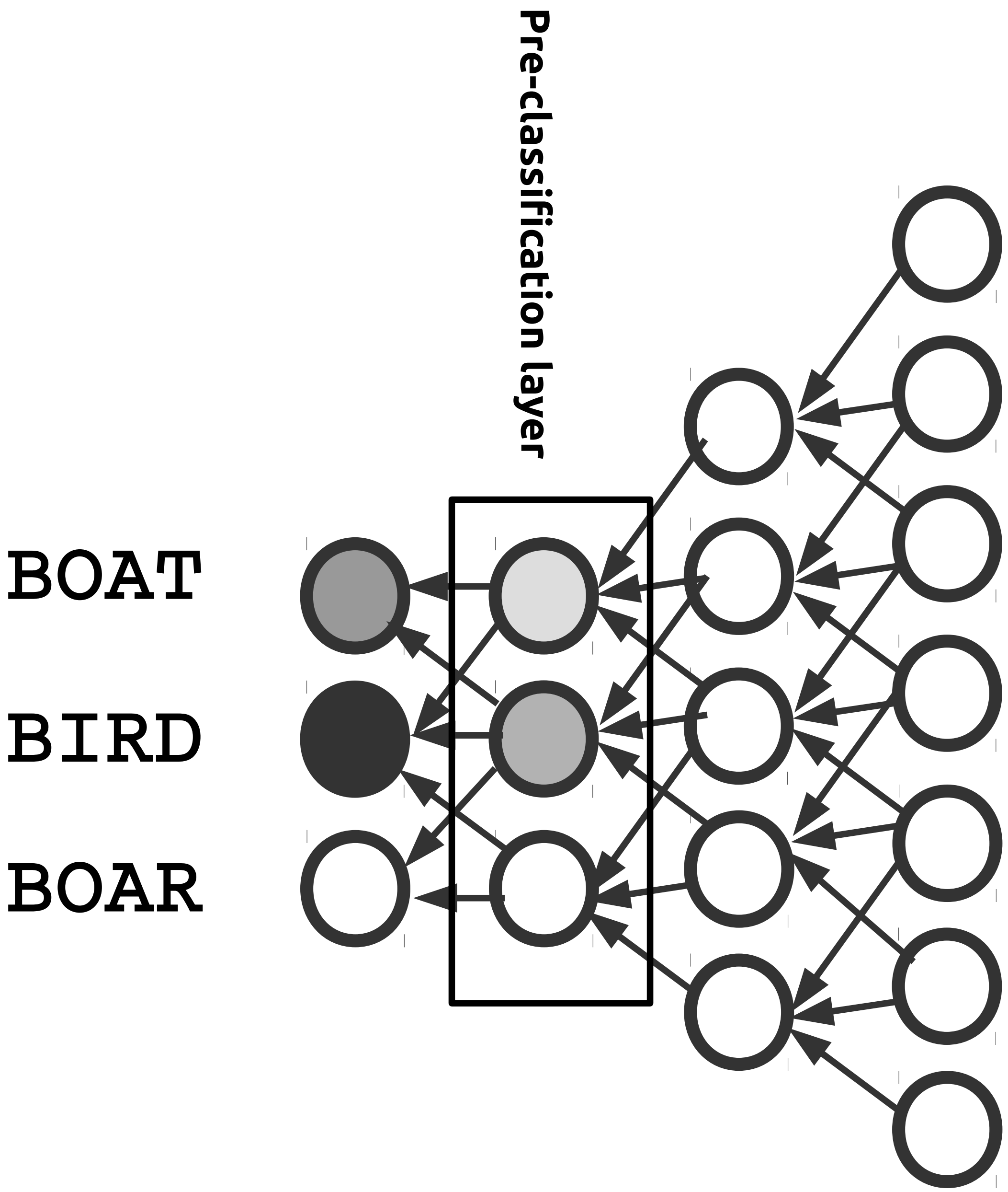
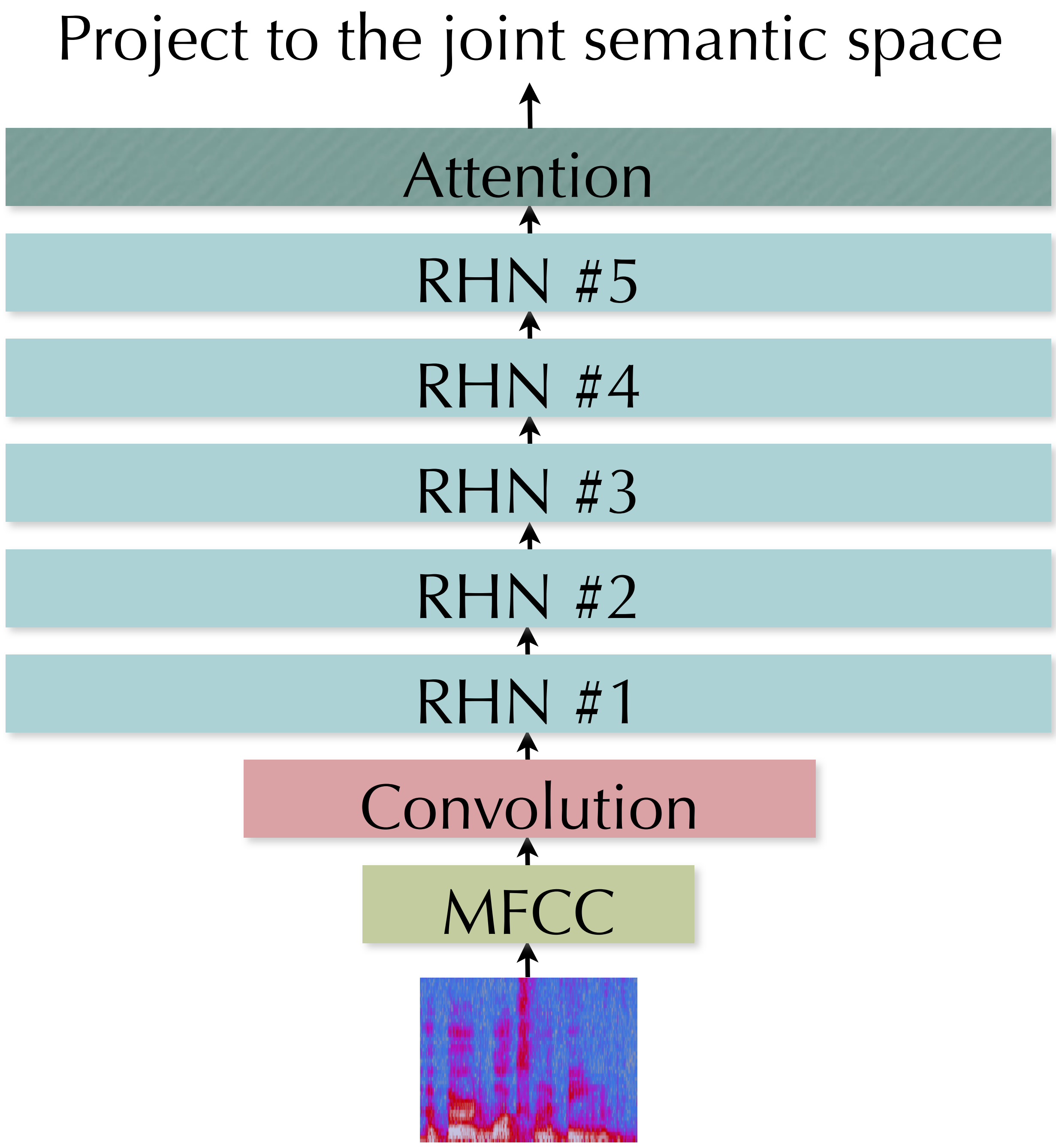


Image Model



VGG-16: Simonyan & Zisserman (2014)

Speech Model



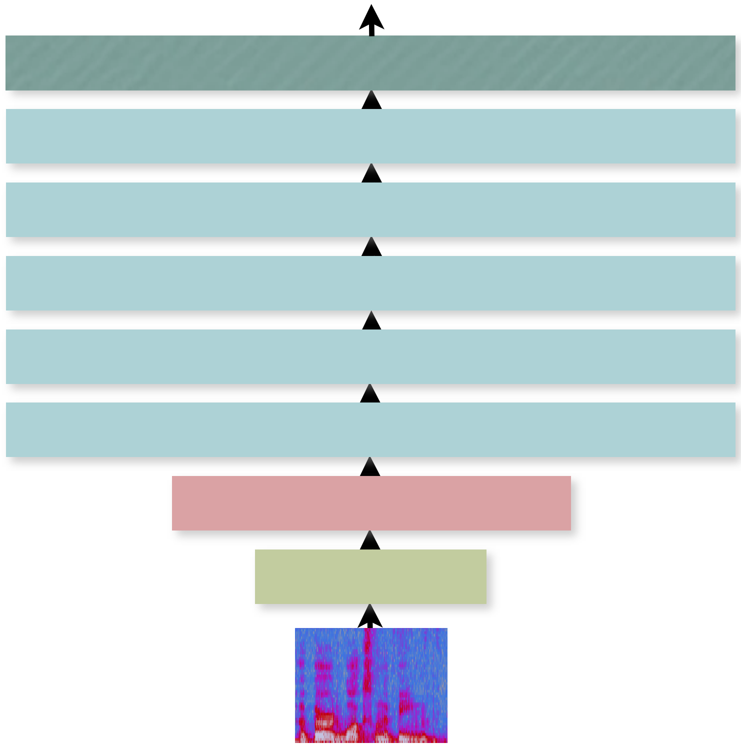
- Attention: weighted sum of last RHN layer units
- RHN: Recurrent Highway Networks (Zilly et al., 2016)
- Convolution: subsampling MFCC vector

Chrupała et al., ACL‘2017

- Representation of language in a model of visually grounded speech signal
    - Using hidden layer activations in a set of auxiliary tasks
    - Predicting utterance length and content, measuring representational similarity and disambiguation of homonyms
- Main findings:
    - Encodings of form and meaning emerge and evolve in hidden layers of stacked RNNs processing grounded speech

Current Study

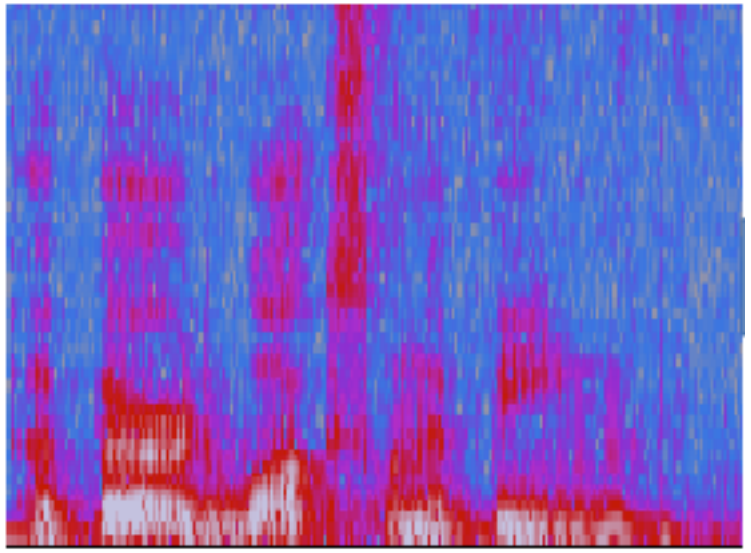
- Questions: how is phonology encoded in
    - MFCC features extracted from speech signal?
    - activations of the layers of the model?
- Data: Synthetically Spoken COCO dataset
- Experiments:
    - Phoneme decoding and clustering
    - Phoneme discrimination
    - Synonym discrimination



Phoneme Decoding

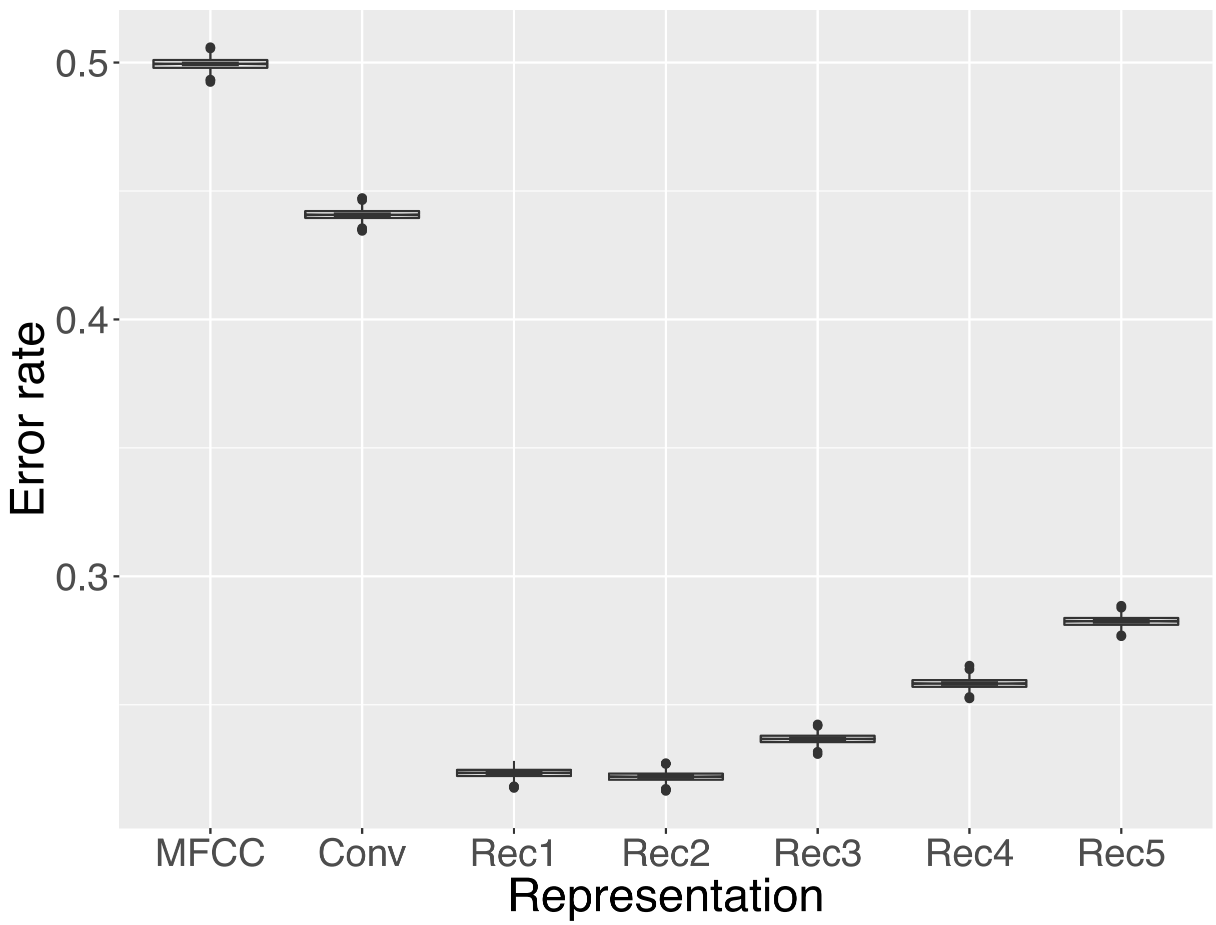
- Identifying phonemes from speech signal/activation patterns: supervised classification of aligned phonemes

- Speech signal was aligned with phonemic transcription using Gentle toolkit (based on Kaldi, Povey et al., 2011)



Phoneme Decoding

- Identifying phonemes from speech signal/activation patterns: supervised classification of aligned phonemes



Phoneme Discrimination

- ABX task (Schatz et al., 2013): discriminate minimal pairs; is X closer to A or to B?

A: `be /bi/` **B:** `me /mi/`

X: `my /maI/`

  - A, B and X are CV syllables
  - (A,B) and (B,X) are minimum pairs, but (A,X) are not (34,288 tuples in total)

Phoneme Discrimination

MFCC	0.71
Convolutional	0.73
Recurrent 1	0.82
Recurrent 2	0.82
Recurrent 3	0.80
Recurrent 4	0.76
Recurrent 5	0.74

Phoneme Discrimination by Class

- The task is most challenging when the target (B) and distractor (A) belong to the same phoneme class

A: be /bi/ **B:** me /mi/

X: my /maI/

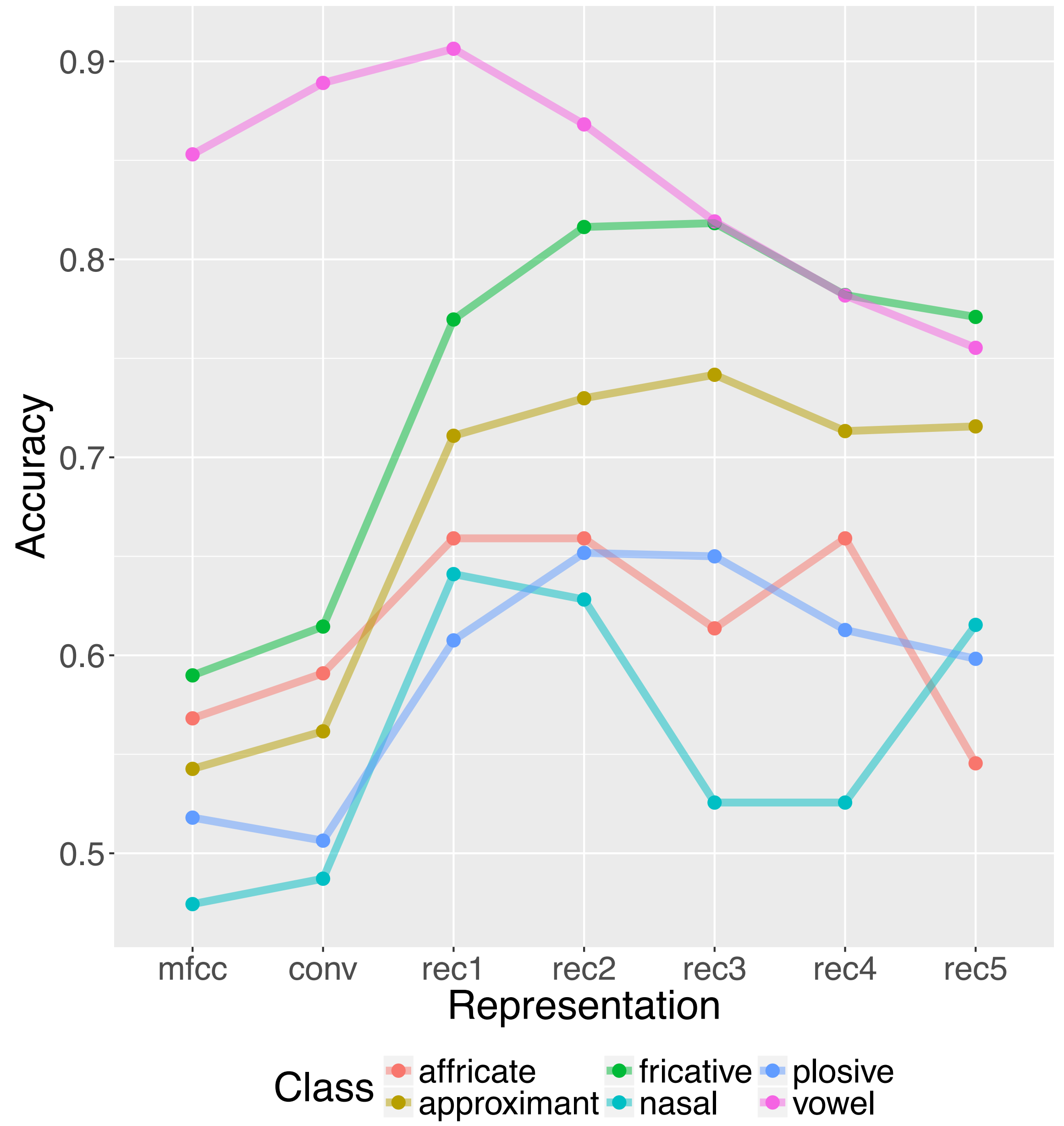
Phoneme Discrimination by Class

- The task is most challenging when the target (B) and distractor (A) belong to the same phoneme class

Vowels	i ɪ ʊ u
	e ɛ ə ɚ ɔɪ ɔ o
	aɪ æ ʌ ɑ aʊ
Approximants	j ɹ l w
Nasals	m n ŋ
Plosives	p b t d k g
Fricatives	f v θ ð s z ʃ ʒ h
Affricates	ʧ ʤ

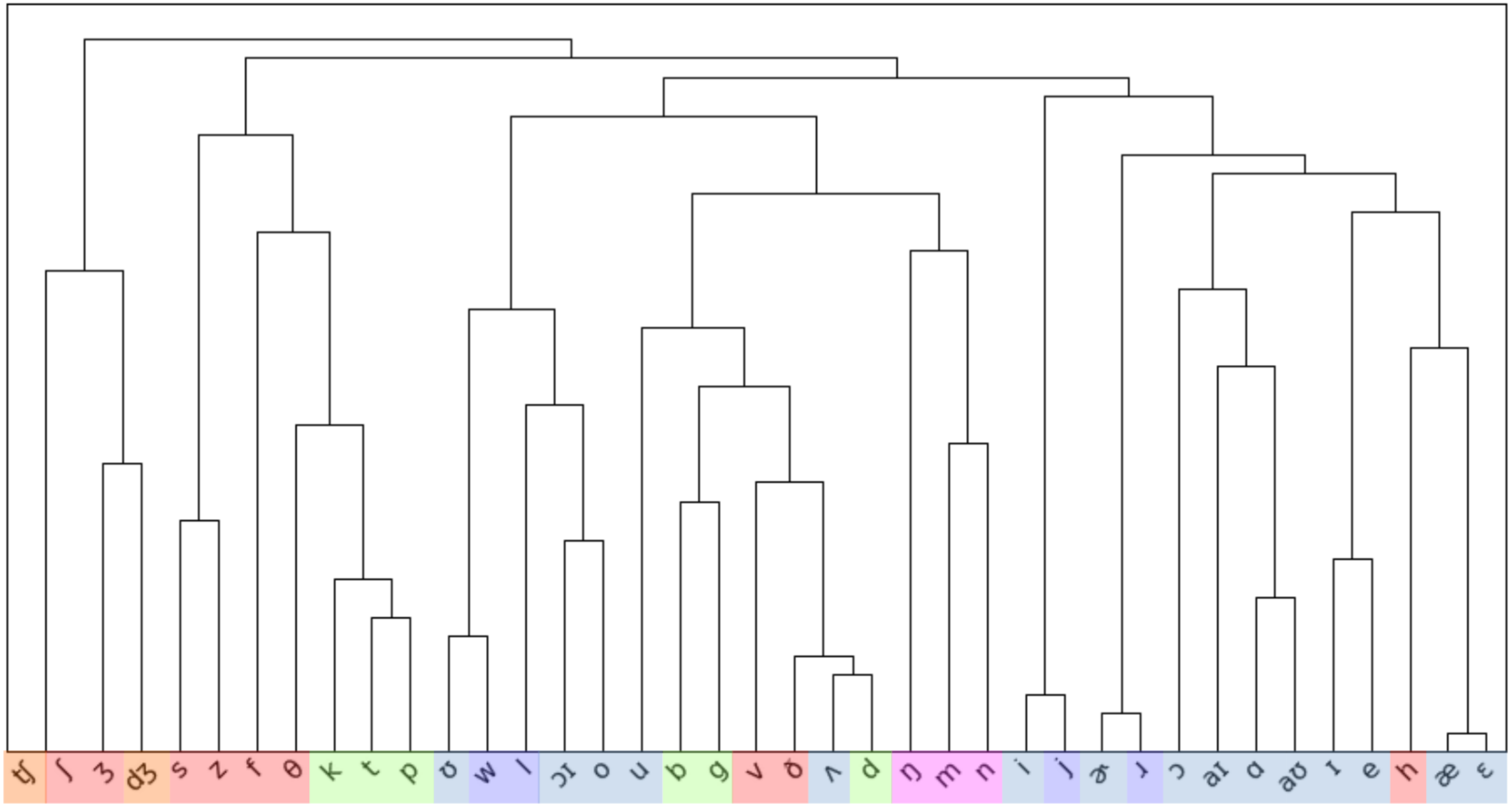
Phoneme Discrimination by Class

- The task is most challenging when the target (B) and distractor (A) belong to the same phoneme class



Organization of Phonemes

- Agglomerative hierarchical clustering of phoneme activation vectors from the first hidden layer:



Synonym Discrimination

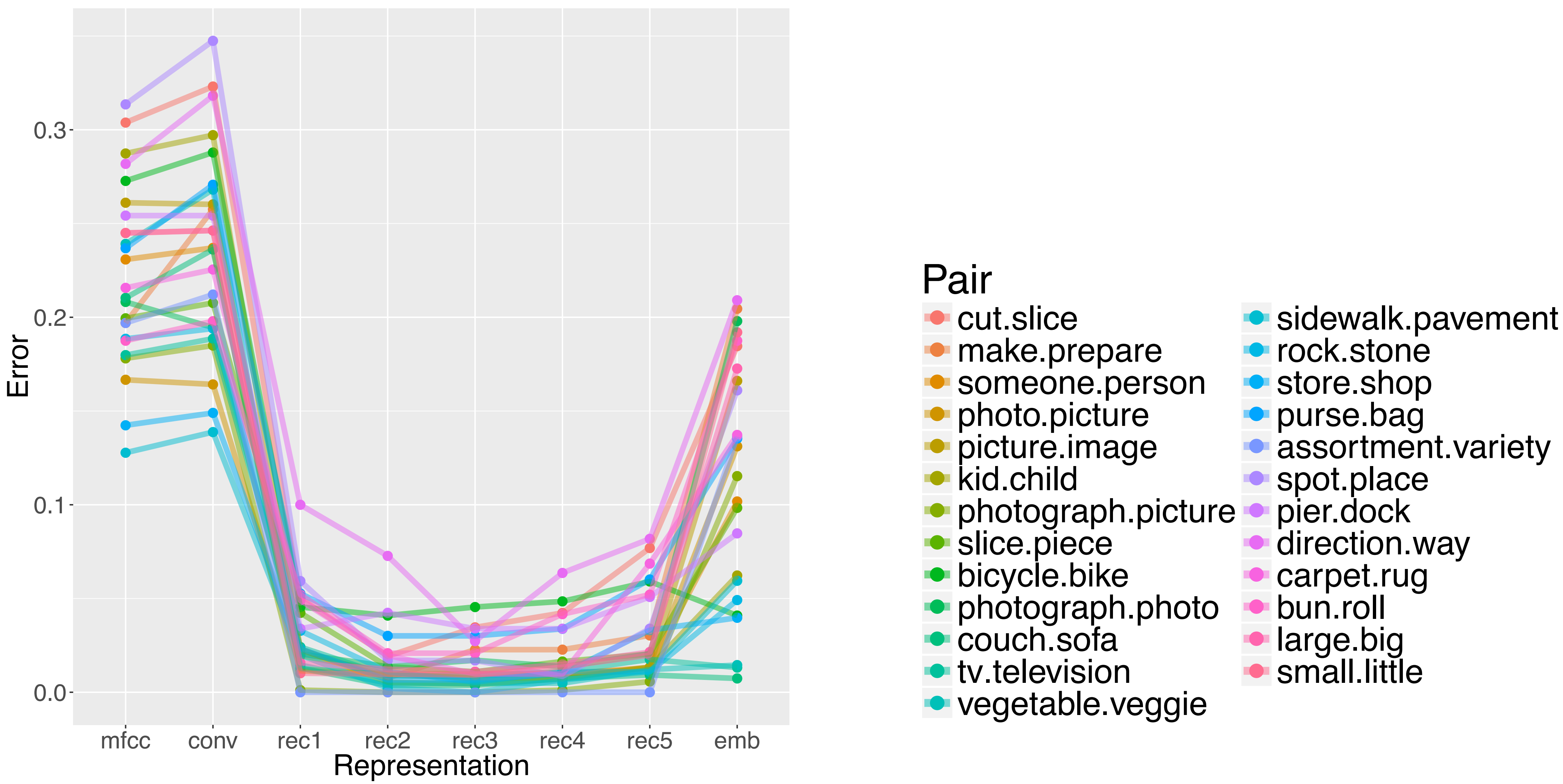
- Distinguishing between synonym pairs in the same context:

  *A girl looking at a **photo***

  *A girl looking at a **picture***

- Synonyms were selected using WordNet synsets:
  - The pair have the same POS tag and are interchangeable
  - The pair clearly differ in form (not *donut/doughnut*)
  - The more frequent token in a pair constitutes less than 95% of the occurrences.

Synonym Discrimination



Conclusion

- Phoneme representations are most salient in lower layers
- Large amount of phonological information persists up to the top recurrent layer
- The attention layer filters out and significantly attenuates encoding of phonology and makes utterance embeddings more invariant to synonymy

`Code:` https://github.com/gchrupala/encoding-of-phonology