



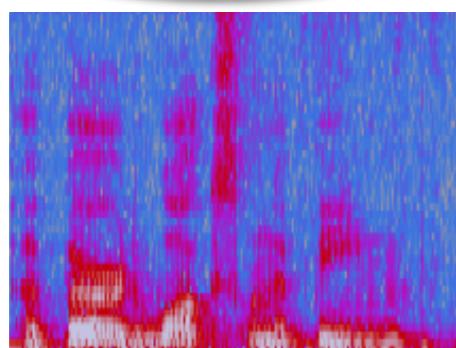
Encoding of Phonology in an RNN model of Grounded Speech

Afra Alishahi, Marie Barking, Grzegorz Chrupała



A Realistic Language Learning Scenario

Two men
are washing an
elephant.



Grounded Language Learning

- Roy & Pentland (2002)
- Yu & Ballard (2014)
- Harwath et al. (2016)
- Gelderloos & Chrupala (2016)
- Harwath & Glass (2017)
- Chrupala et al. (2017)

Analysis of Linguistic Knowledge

- 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)

Grounded Language Learning

- Roy & Pentland (2002)
- Yu & Ballard (2014)
- Harwath et al. (2016)
- Gelderloos & Chrupala (2016)
- Harwath & Glass (2017)
- Chrupala et al. (2017)

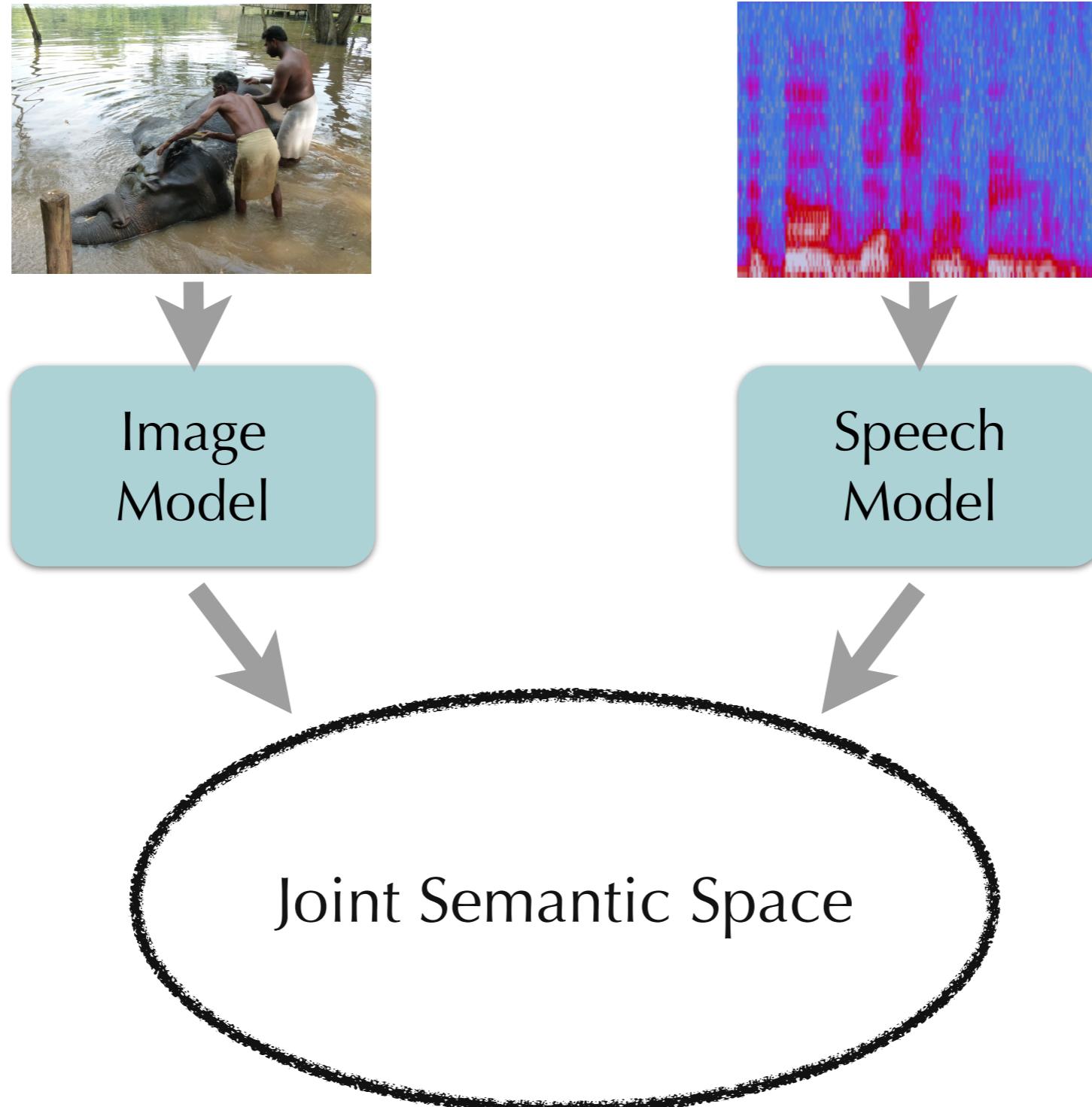
Analysis of Linguistic Knowledge

- 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

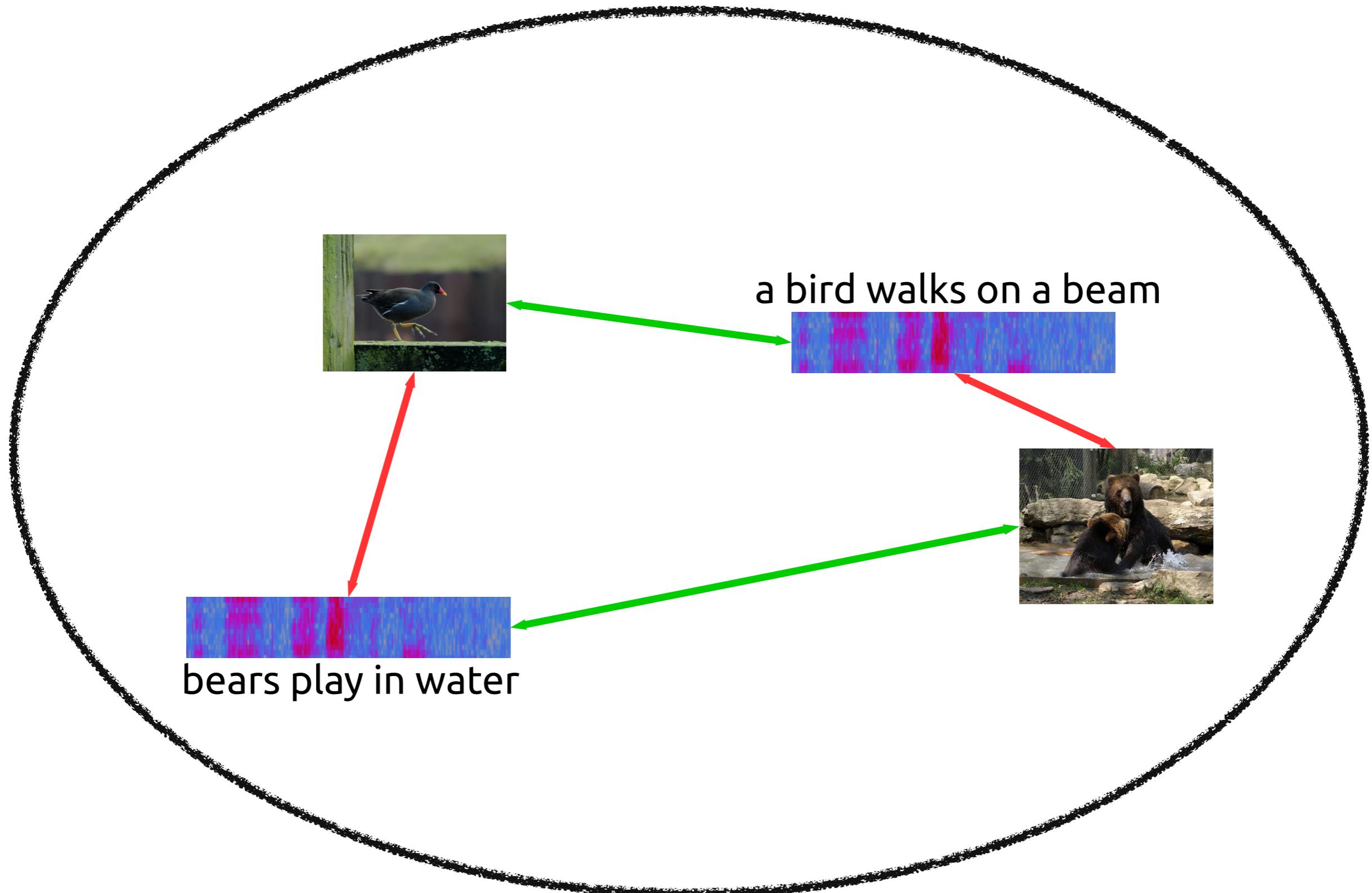
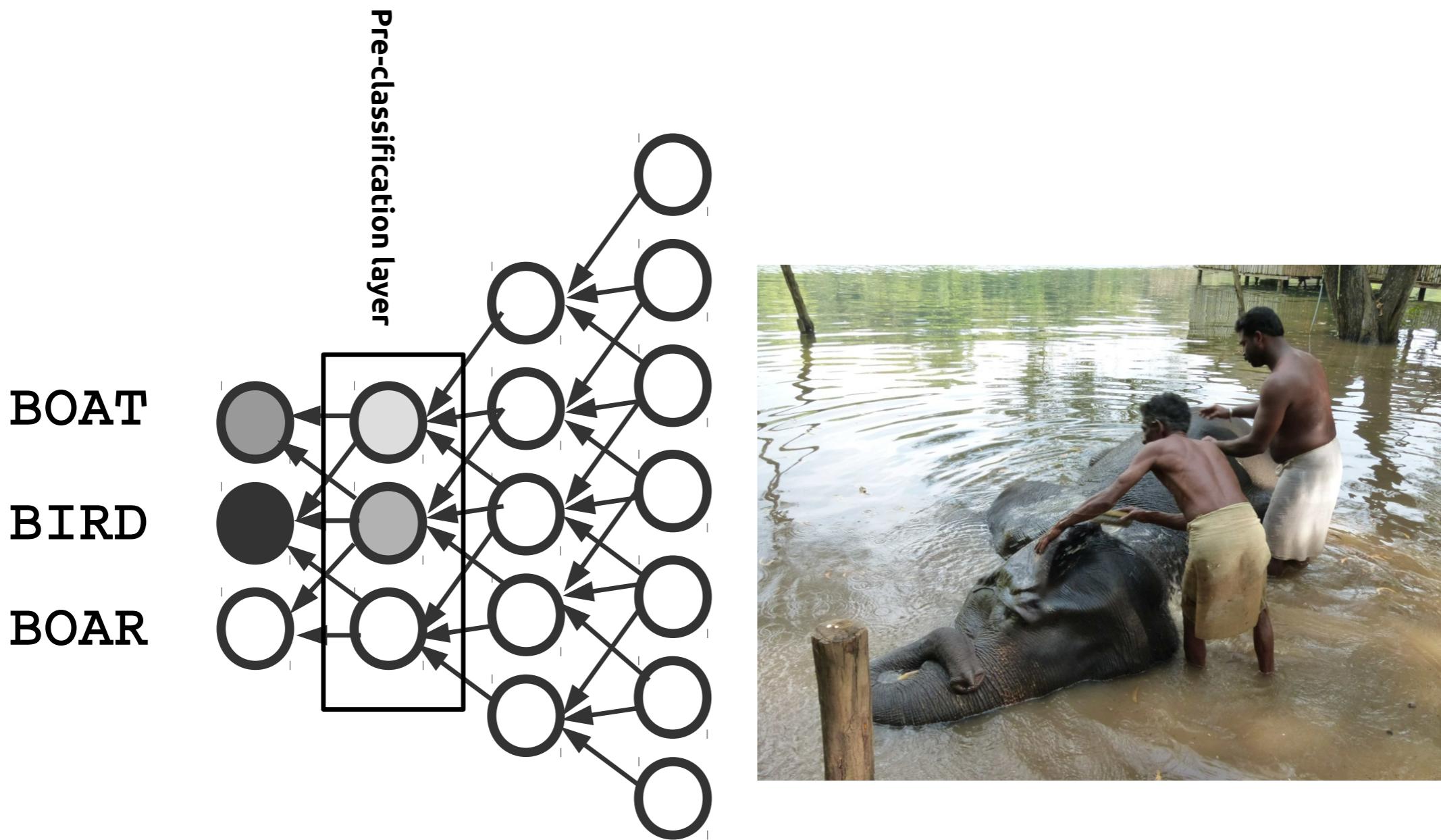


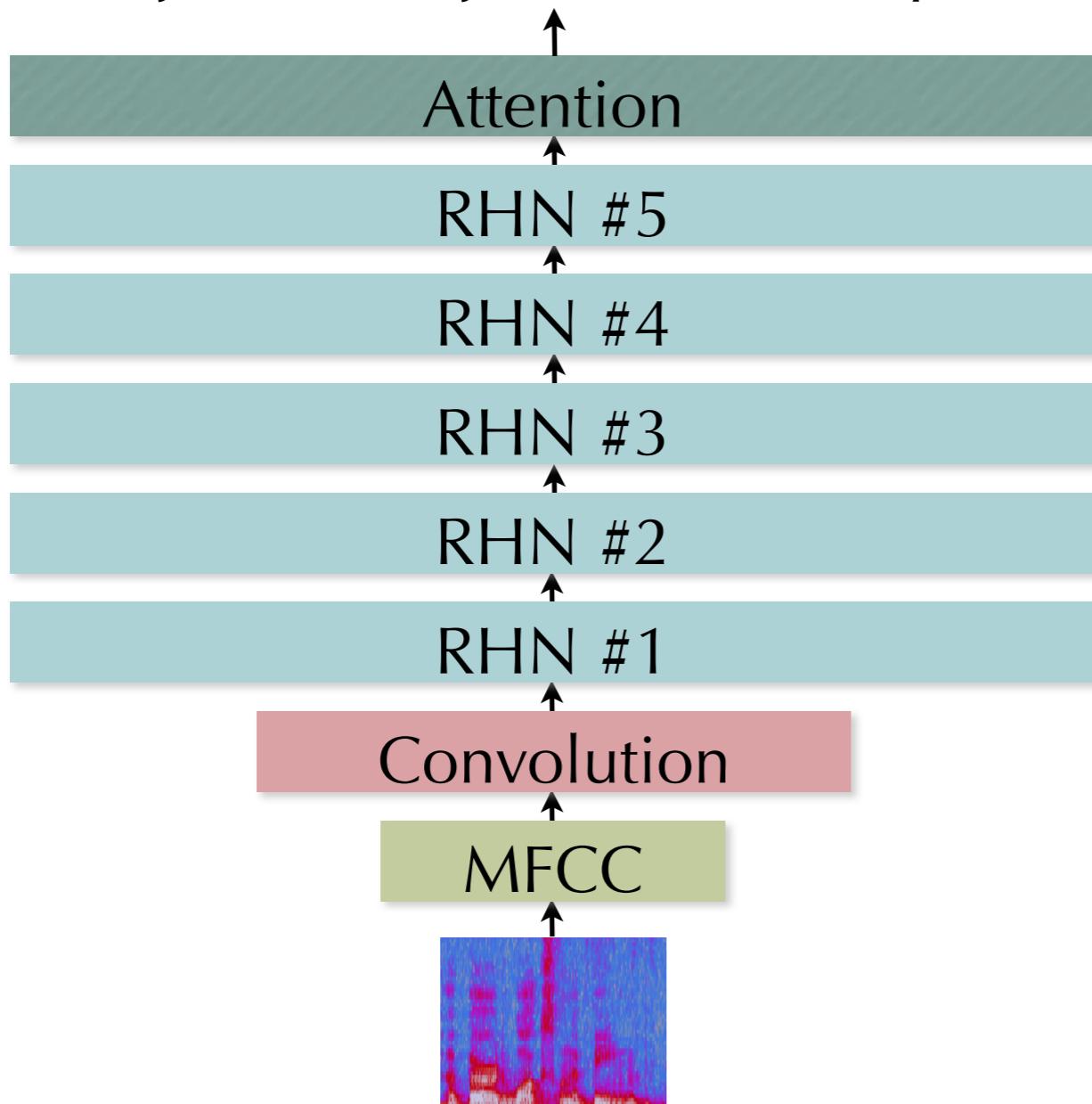
Image Model



VGG-16: Simonyan & Zisserman (2014)

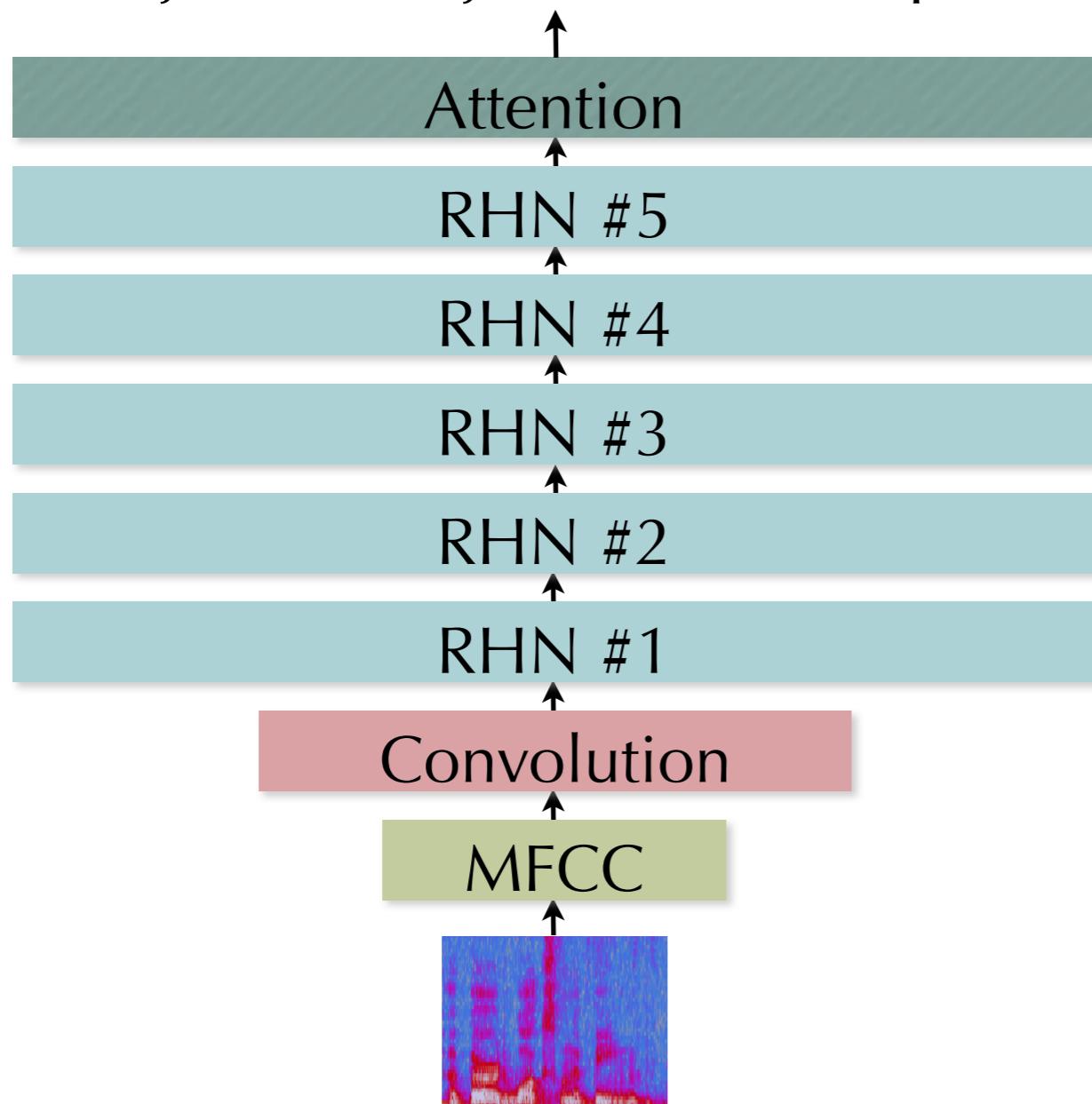
Speech Model

Project to the joint semantic space



Speech Model

Project to the joint semantic space



- 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

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

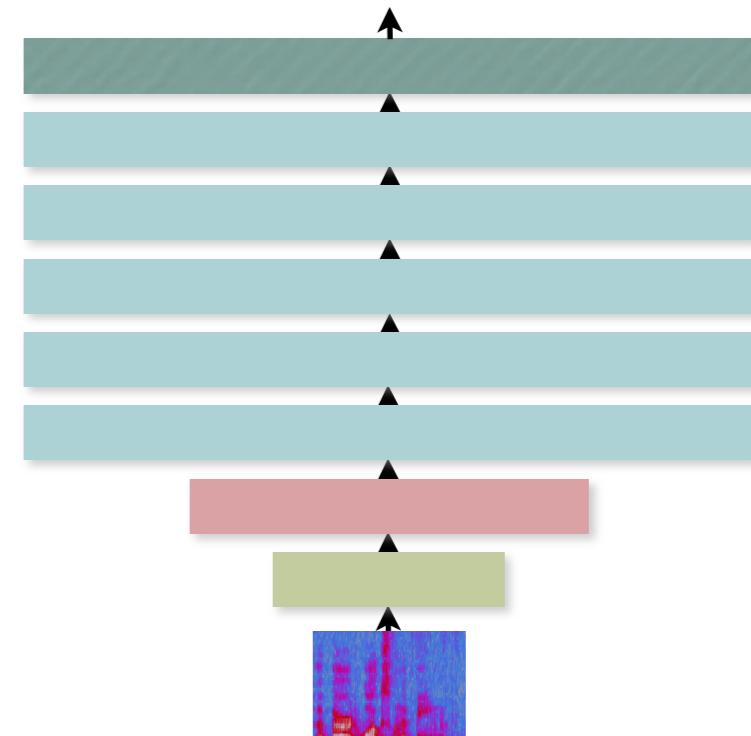
- Questions: how is phonology encoded in
 - MFCC features extracted from speech signal?
 - activations of the layers of the model?

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

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

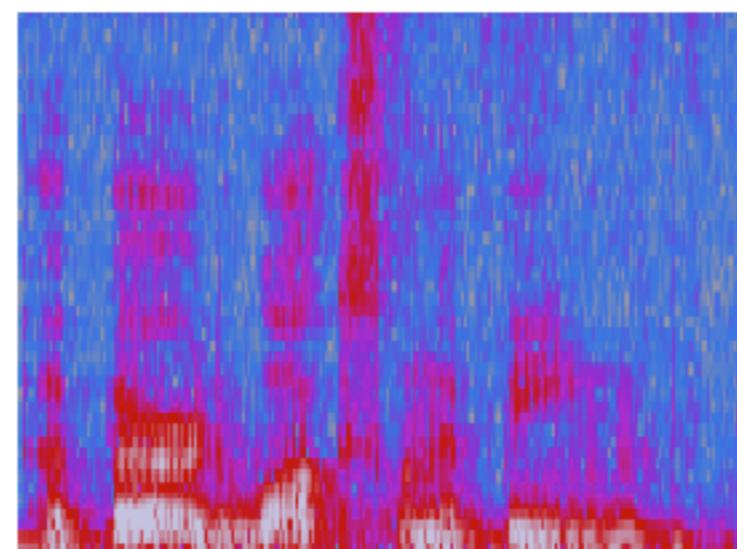


Phoneme Decoding

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

Phoneme Decoding

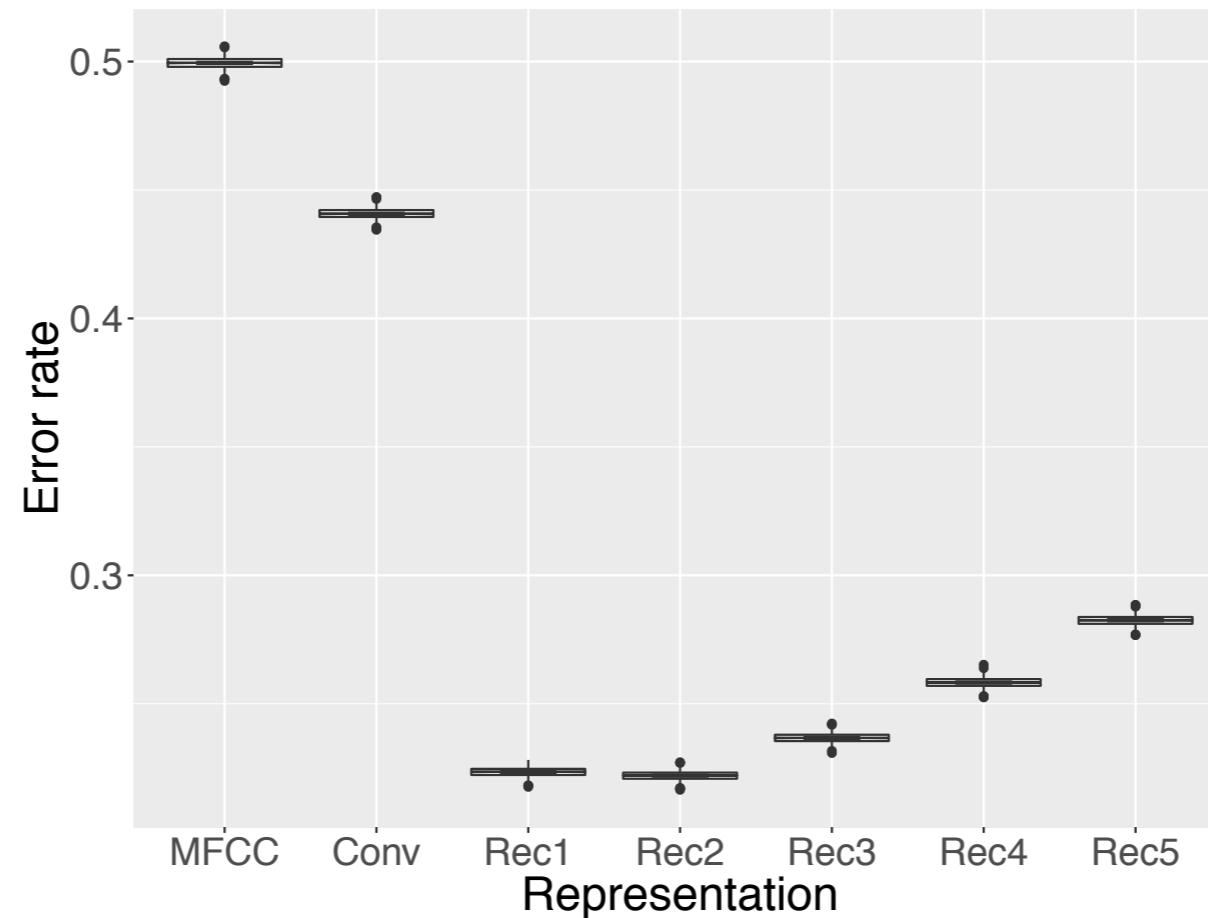
- Identifying phonemes from speech signal/activation patterns: supervised classification of aligned phonemes
- Speech signal was aligned with phonemic transcription using Gentle toolkit (based on Kaldi, Povey et al., 2011)



a b'ɔ:k'swɪm

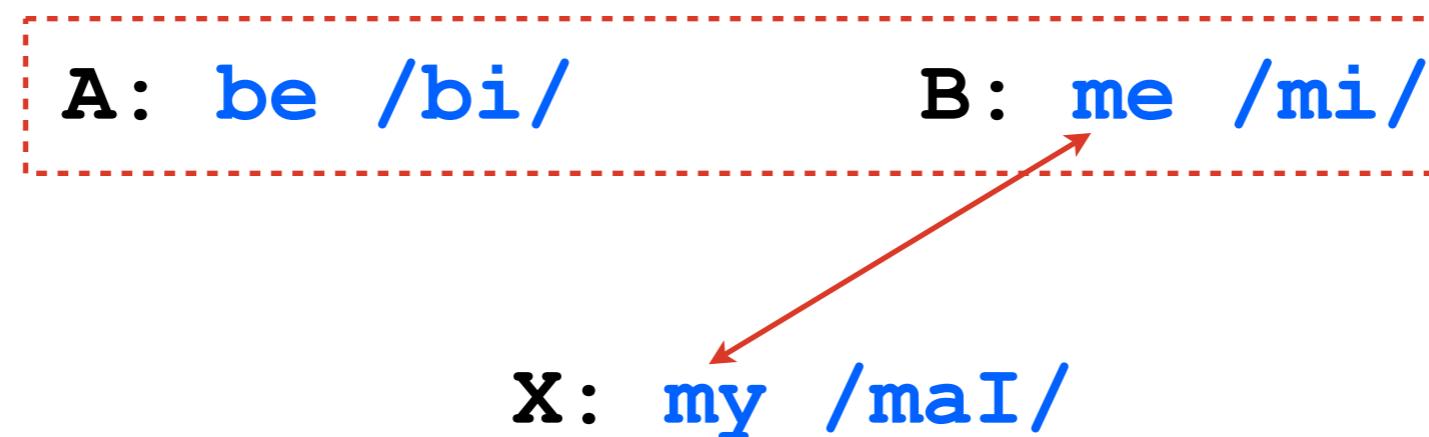
Phoneme Decoding

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



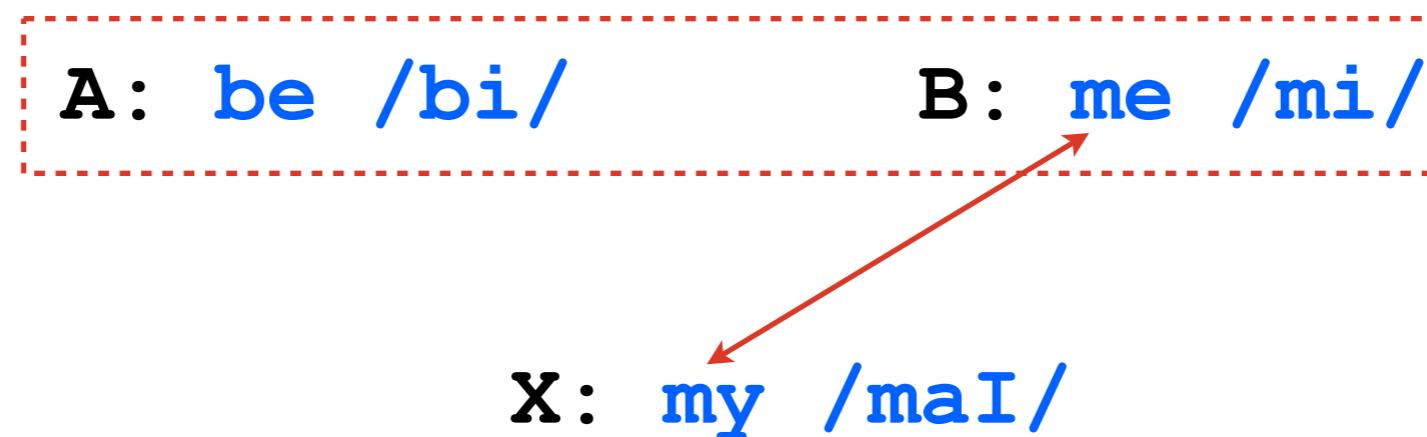
Phoneme Discrimination

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



Phoneme Discrimination

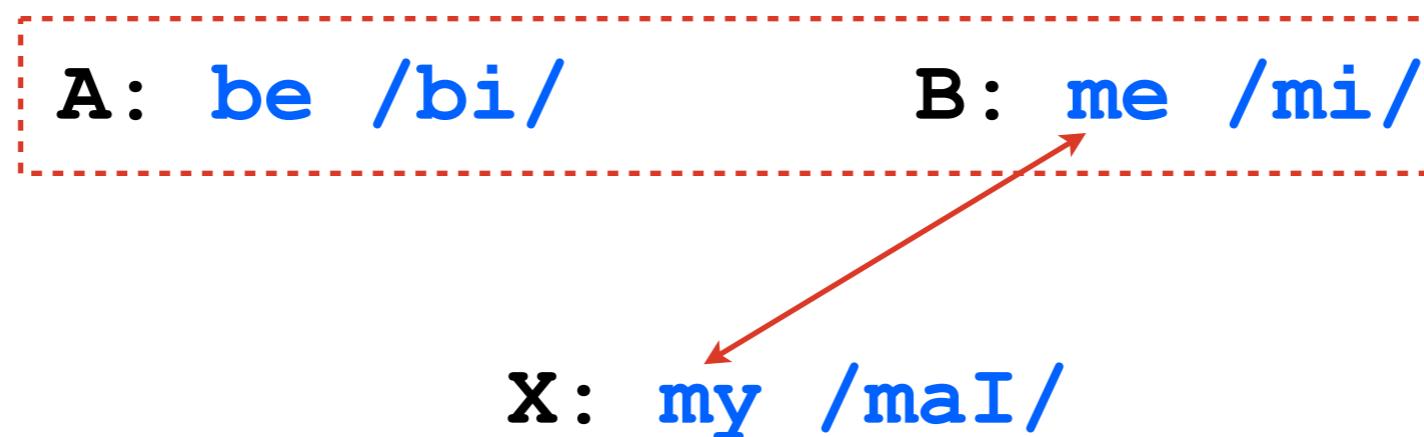
- ABX task (Schatz et al., 2013): discriminate minimal pairs; is X closer to A or to B?



- A, B and X are CV syllables

Phoneme Discrimination

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



- 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

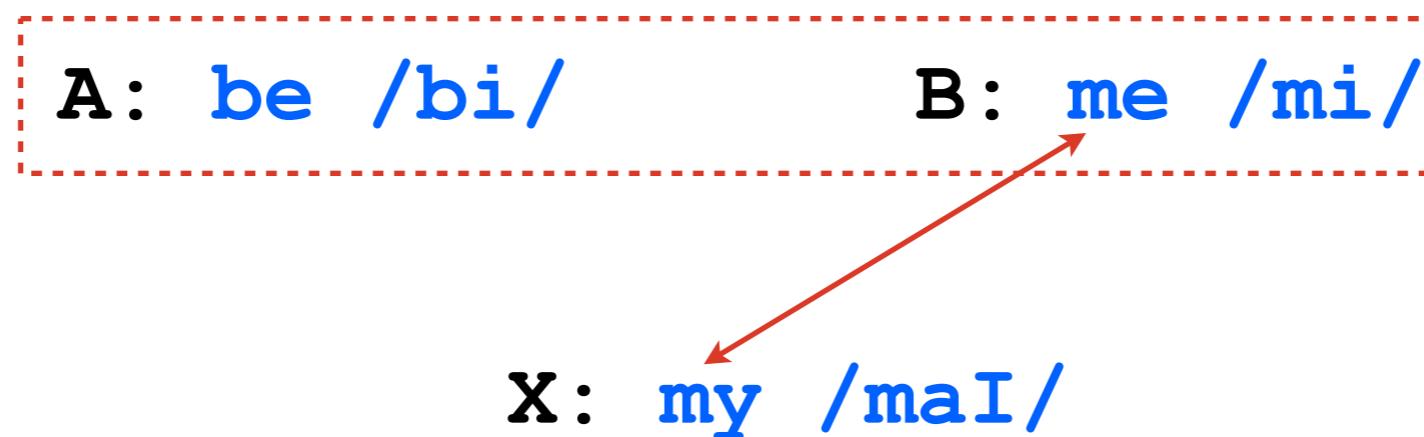
MFCC	0.71
Convolutional	0.73
Recurrent 1	0.82
Recurrent 2	0.82
Recurrent 3	0.80
Recurrent 4	0.76
Recurrent 5	0.74

Phoneme Discrimination

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



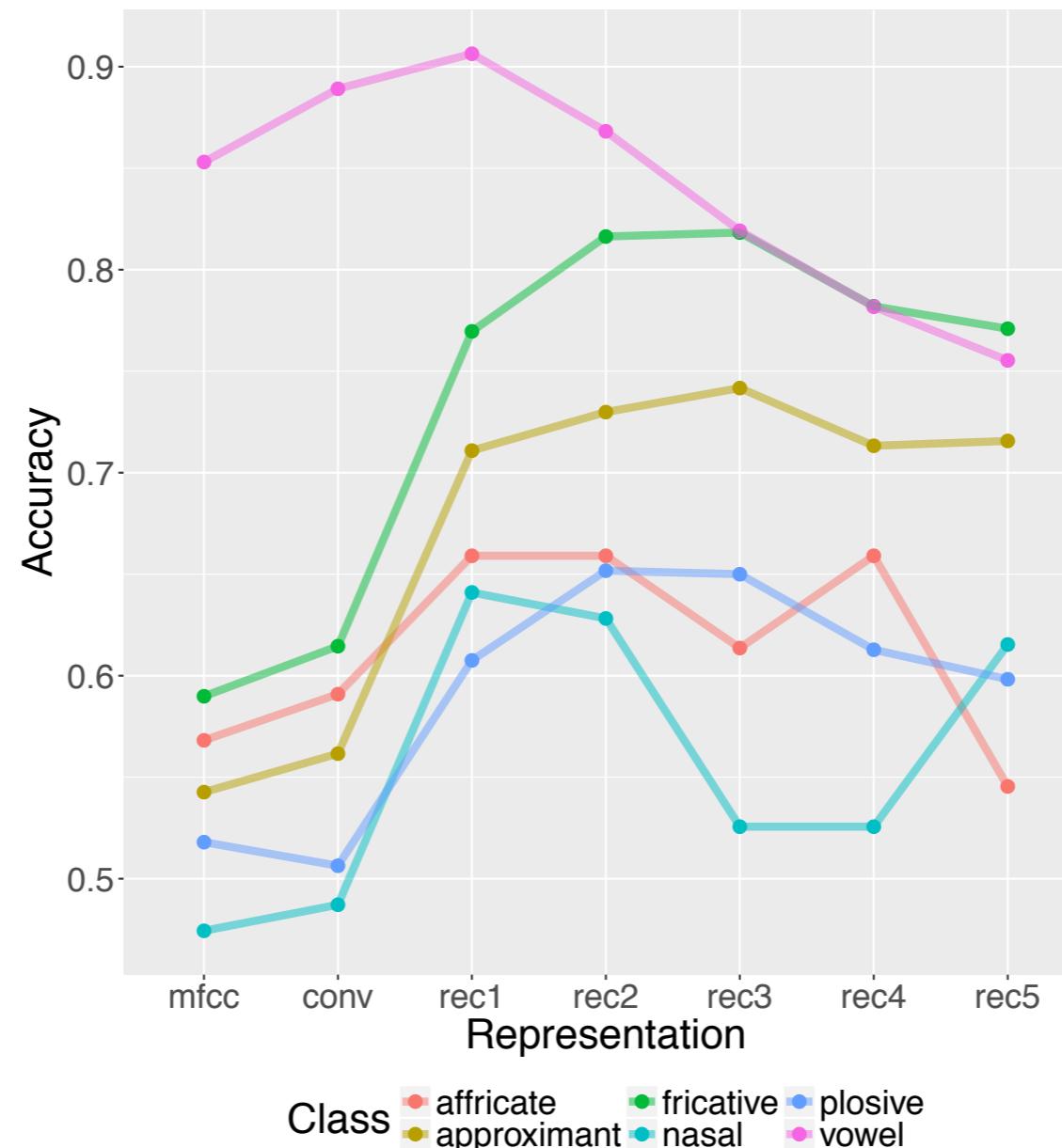
Phoneme Discrimination by Class

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

Vowels	i ɪ ʊ u ə ɛ ə ɔ ɪ ɔ ə aɪ æ ʌ ɑ əʊ
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	tʃ ɸ

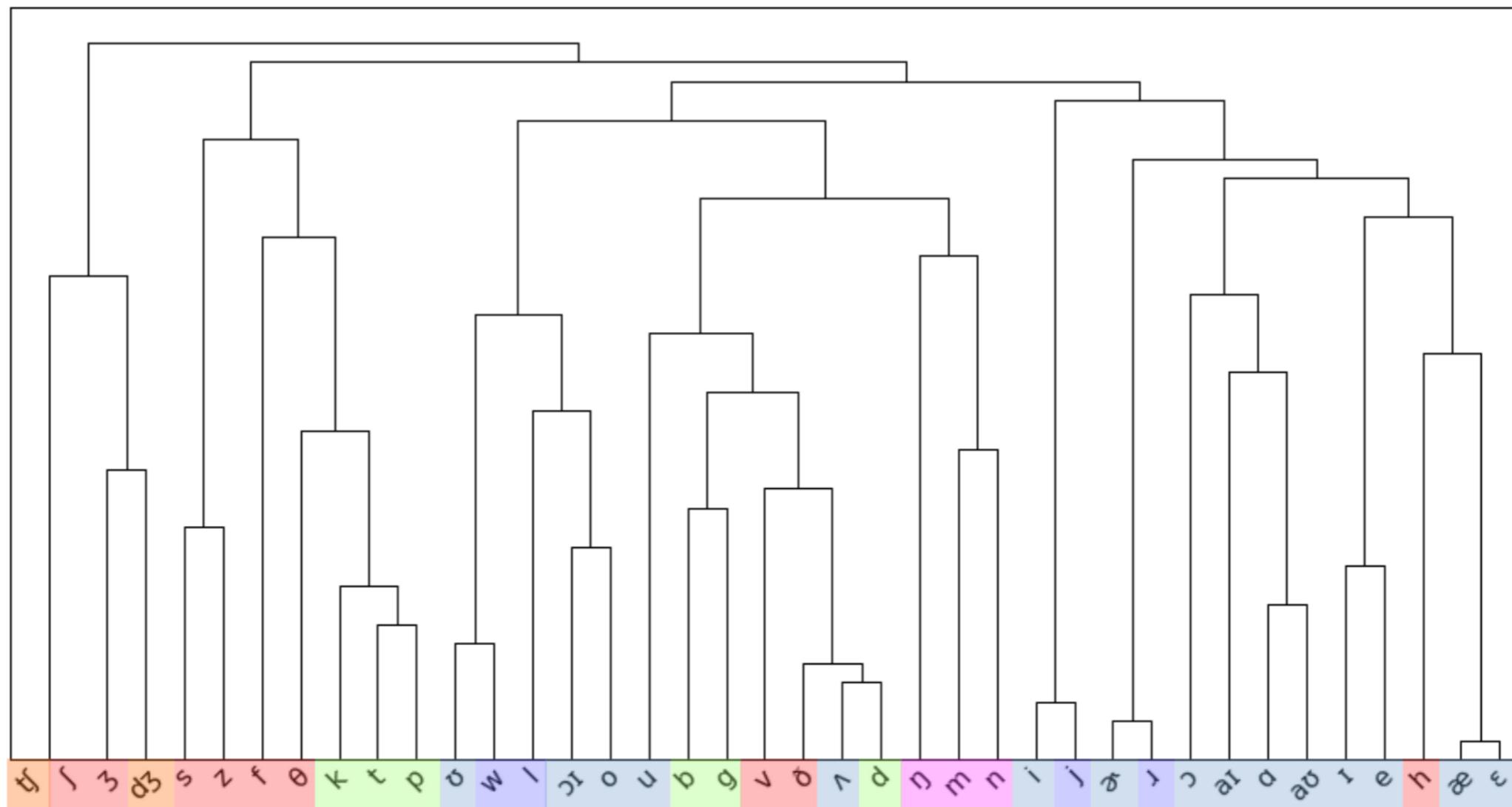
Phoneme Discrimination by Class

- 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***

Synonym Discrimination

- Distinguishing between synonym pairs in the same context:

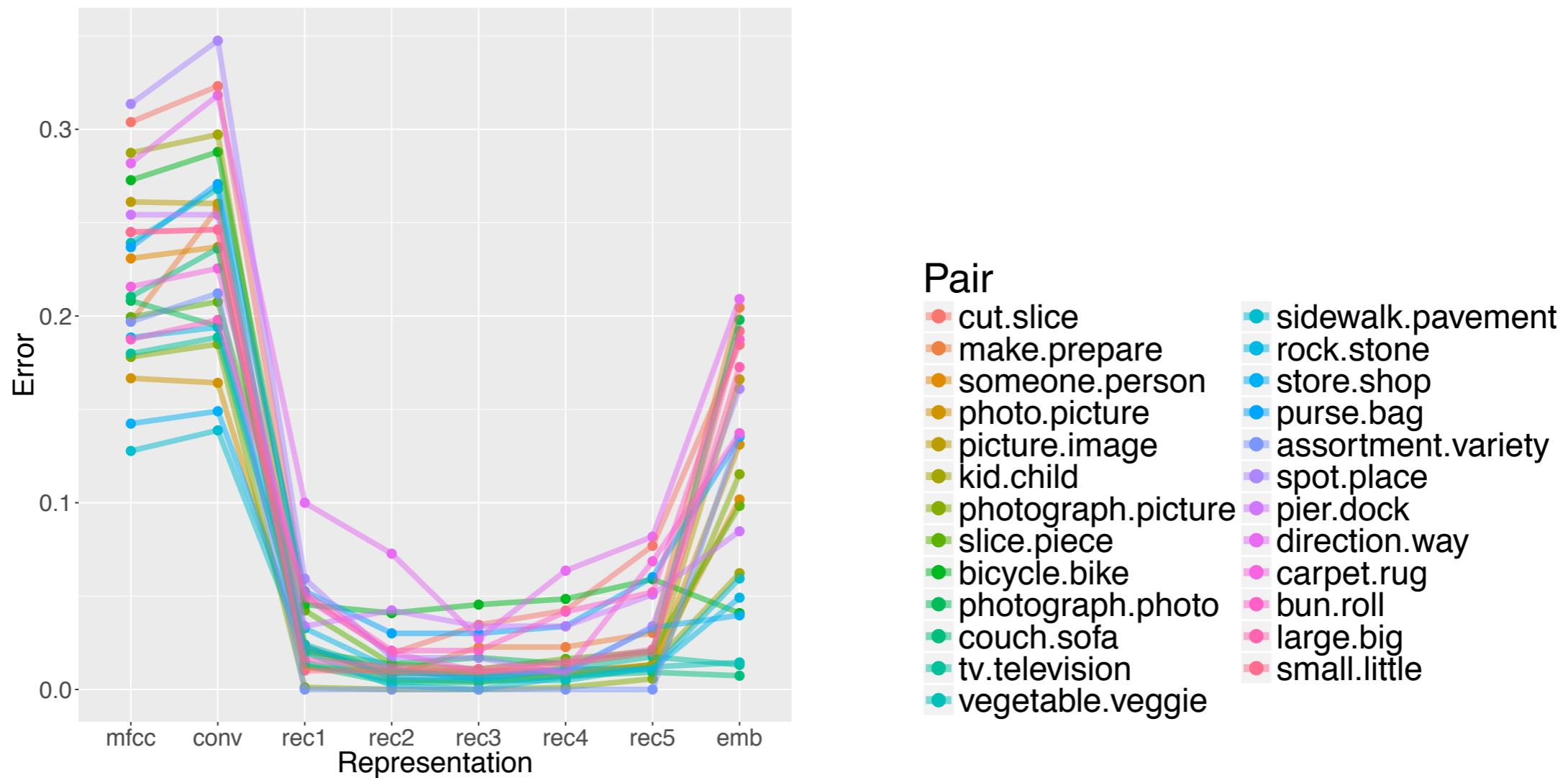
A girl looking at a photo

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

Conclusion

- Phoneme representations are most salient in lower layers

Conclusion

- Phoneme representations are most salient in lower layers
- Large amount of phonological information persists up to the top recurrent layer

Conclusion

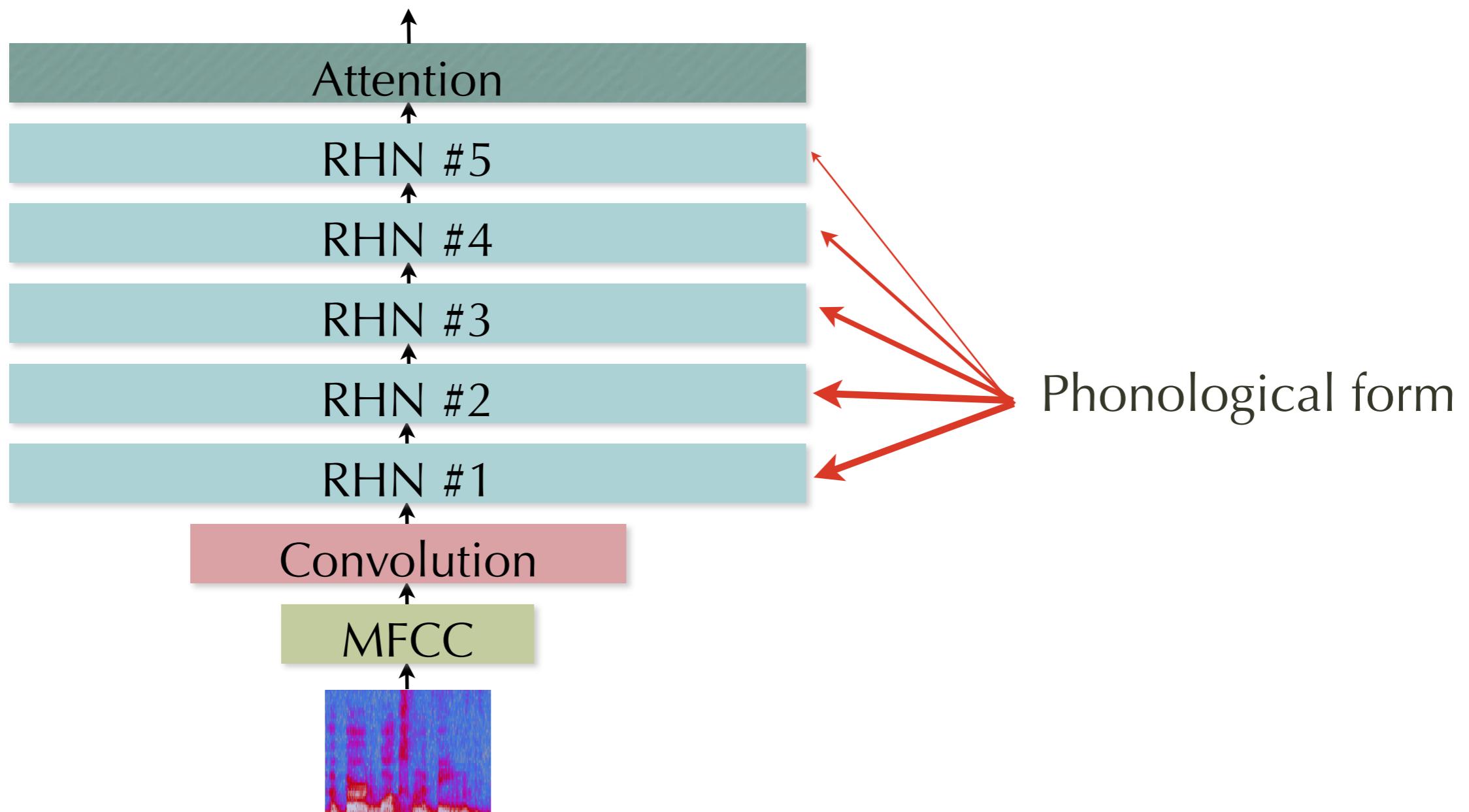
- Phoneme representations are most salient in lower layers
- Large amount of phonological information persists up to the top recurrent layer
- The attention layer filters out and significantly attenuates encoding of phonology and makes utterance embeddings more invariant to synonymy

Conclusion

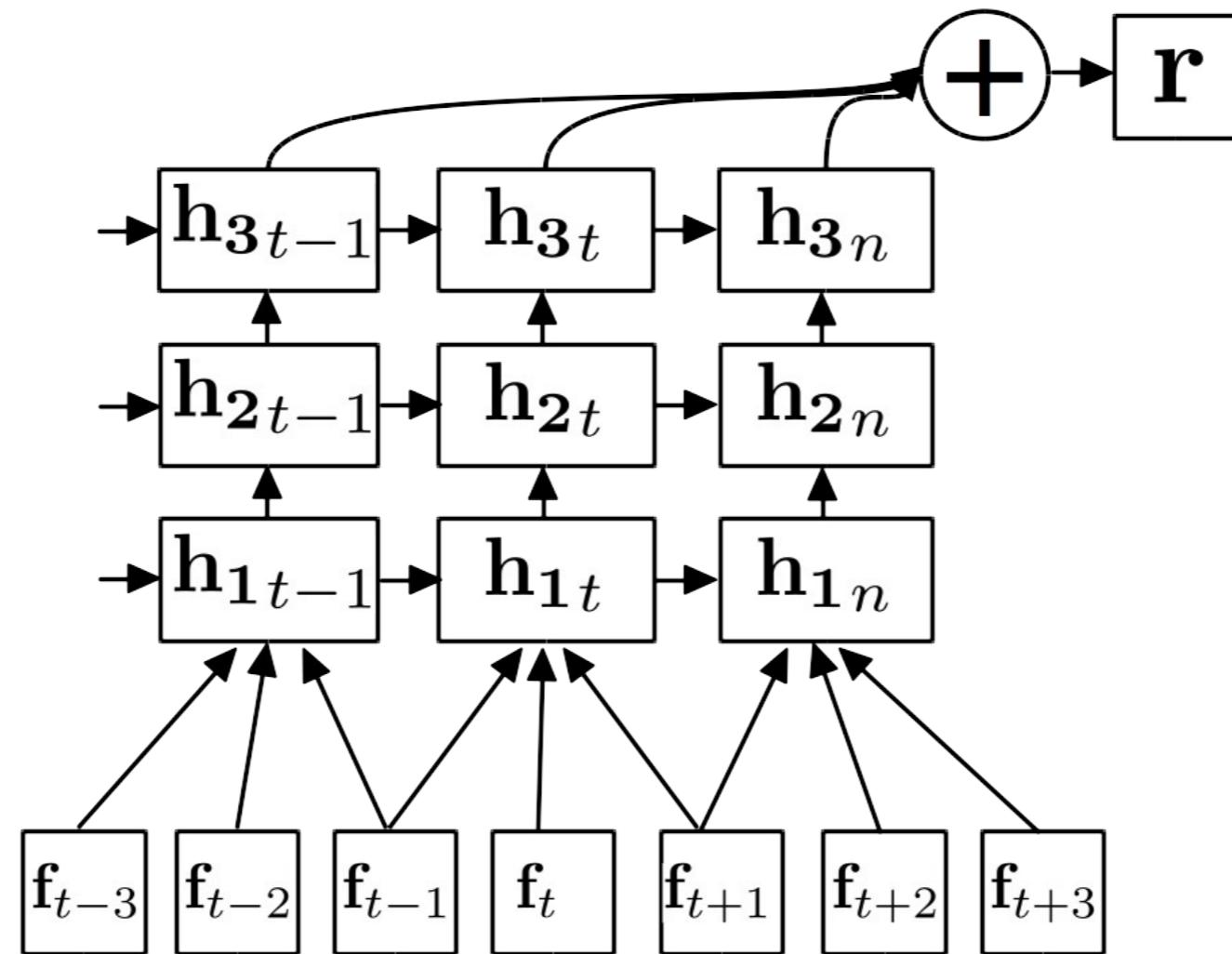
- Phoneme representations are most salient in lower layers
- Large amount of phonological information persists up to the top recurrent layer
- The attention layer filters out and significantly attenuates encoding of phonology and makes utterance embeddings more invariant to synonymy

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

Speech Model



Speech Model



Input: MFCC features extracted from speech signal

Objective Function

- Minimize distance between all corresponding utterance-image (u,i) pairs, and maximize the distance between non-corresponding ones:

$$\sum_{u,i} \left(\sum_{u'} \max[0, \alpha + d(u, i) - d(u', i)] + \sum_{i'} \max[0, \alpha + d(u, i) - d(u, i')] \right)$$

Experimental Setup

- Model specifications:

Attention: size 512
Recurrent 5: size 512
Recurrent 4: size 512
Recurrent 3: size 512
Recurrent 2: size 512
Recurrent 1: size 512
Convolutional: size 64, length 6, stride 3
Input MFCC: size 13

Organization of Phonemes

- Adjusted Rand Index for the comparison of the phoneme type hierarchy induced from representations against phoneme classes:

