The representation of speech in the human and artificial brain

Dr. Odette Scharenborg

Associate Professor & Delft Technology Fellow
Delft University of Technology
The Netherlands
o.e.scharenborg@tudelft.nl

TU Delft

The representation of speech in the human and deep neural networks

Dr. Odette Scharenborg

Associate Professor & Delft Technology Fellow
Delft University of Technology
The Netherlands
o.e.scharenborg@tudelft.nl

Acknowledgements

- Polina Drozdova
- Roeland van Hout

- Sebastian Tiesmeyer
- Nikki van der Gouw
- Martha Larson

- Najim Dehak
- Junrui Ni
- Mark Hasegawa-Johnson

TU Delft

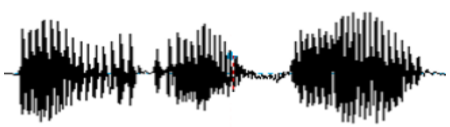
Ultimate goal

Automatic speech recognition (ASR) for all the world's languages & all types of speech

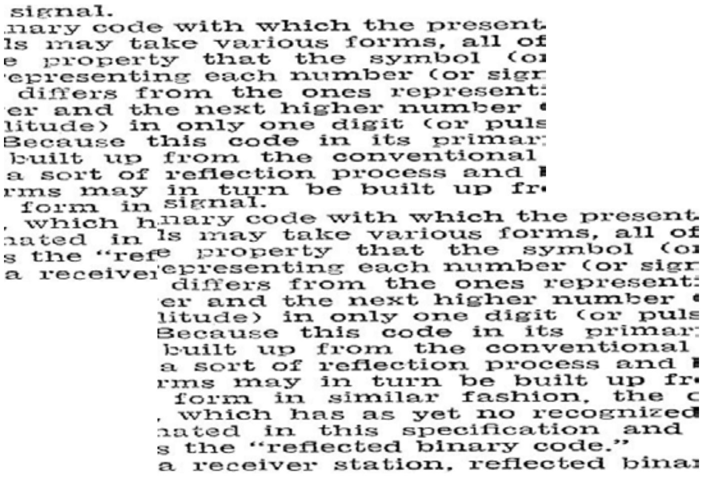
Problem

Transcription 1

Transcription 2



Transcription 3



Possible solution

Create a flexible ASR that is

- trained on language/speech type X and
- mapped to language/speech type Y

What we need

1. Invariant units of speech which transfer easily and accurately to new languages & types of speech

2. ASR system that can flexibly adapt to new languages & types of speech

3. ASR system that can *decide* when to create a new sound category

4. ASR system that can *create* a new sound category

What we need

1. Invariant units of speech which transfer easily and accurately to new languages & types of speech

2. ASR system that can flexibly adapt to new languages & types of speech

3. ASR system that can *decide* when to create a new sound category

4. ASR system that can *create* a new sound category

Human listeners

1. Adapt to all types of pronunciations/speakers

2. Create new sound categories when learning a new language

Comparing humans and machines

- Different hardware:  vs. 

- Same task: recognition of words from speech

⇨ Comparing humans and machines [Scharenborg, 2007]:

  – provide insights into human speech processing (computational modelling)

  – improve ASR technology (MFCCs, PLPs, template-based ASR)

Ultimate question

- What is the optimal representation of speech for a DNN-based ASR to be able to map one language/type of speech to another?

This talk

- How do humans adapt? ➔ Perceptual learning in humans
- Perceptual learning in DNNs
- Perceptual learning in DNNs over time
- Representation of speech in DNNs

Perceptual learning

"[…] relatively long-lasting changes to an organism's perceptual system that improve its ability to respond to its environment and are caused by this environment" [Goldstone, 1998, p. 586]

[Drozdova, van Hout, Scharenborg, 2018; Scharenborg & Janse, 2013]

Perceptual learning

"[…] relatively **long-lasting changes** to an organism's perceptual system that improve its ability to **respond to its environment** and are caused by this environment" [Goldstone, 1998, p. 586]

[Drozdova, van Hout, Scharenborg, 2018; Scharenborg & Janse, 2013]

Lexically-guided perceptual learning

Disambiguating context

[l/ɹ]

Ambiguous input

exposure

appe[l/ɹ] (= *apple*)

wekke[l/ɹ] (= *alarm clock*)

test

mee[l/ɹ]

Adjustment of the phoneme category boundary

(= *flour*)

(= *lake*)

[Drozdova, van Hout & Scharenborg, 2015, 2016; Norris, McQueen & Cutler, 2003]

General procedure

Exposure: lexical decision task

Amb-l group:

- 40 ambiguous /l/ words
- 40 natural /r/ words

Amb-r group

- 40 ambiguous /r/ words
- 40 natural /l/ words

+ filler words and nonwords

Test: phonetic categorisation task

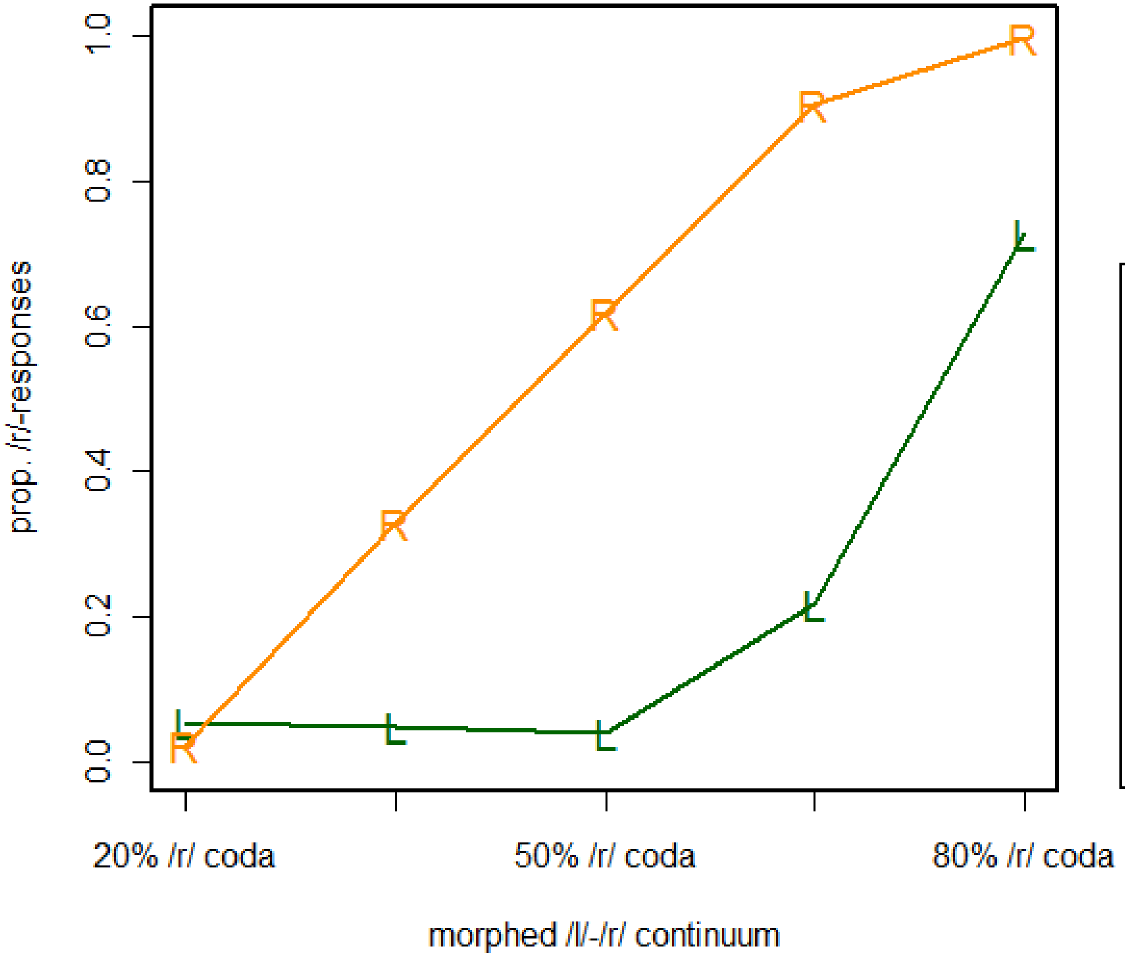
Ambiguous [l/r] test items

20% 40% 50% 60% 80% /r/

5 steps, each 18 times

TUDelft

Phonetic categorisation results



R: listeners who learned to map [l/ɹ] onto [ɹ]

L: listeners who learned to map [l/ɹ] onto [l]

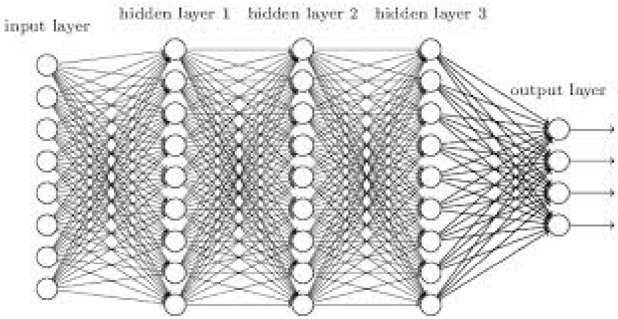
[Scharenborg, Mitterer & McQueen, Interspeech, 2011]

Lexically-guided perceptual learning

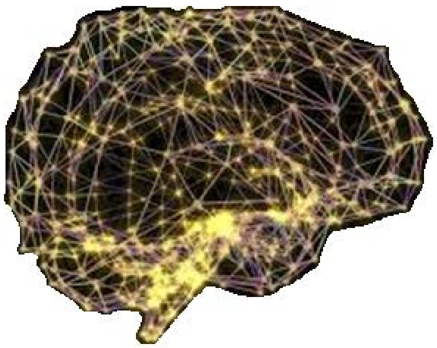
- Causes a temporary change in phonetic category boundaries
- Needs lexical or phonotactic knowledge
- Generalises to words that have not been presented earlier
- Speaker dependent
- Fast: only 10-15 ambiguous items needed
- Long-lasting effect

TU Delft [Clarke-Davidson et al., 2008; Eisner & McQueen, 2005, 2006; Kraljic & Samuel, 2005, 2007; McQueen et al., 2006; Scharenborg & Janse, 2013; Drozdova, van Hout, & Scharenborg, 2016]

Perceptual learning in DNNs



inspired by



… but can they flexibly adapt to different speech types like human listeners can?

TUDelft [Scharenborg, Tiesmeyer, Hasegawa-Johnson, Dehak, 2018]

If so, **visualize** the mechanism by which the DNNs perform this adaptation → **intermediate representations** that might have **correlates in human perceptual adaptation**

Lexical retuning

1. Train baseline DNN on Dutch read speech ~64h:

2. Retrain the baseline model using the acoustic stimuli of the human experiment [Scharenborg & Janse, 2013]:

Baseline	AmbR	AmbL
• 120 natural words • 40 [ɹ]-final words • 40 [l]-final words	• 120 natural words • 40 [l]-final words • 40 [ɹ]-final words: [ɹ] replaced by [l/ɹ]	• 120 natural words • 40 [ɹ]-final words • 40 [l]-final words: [l] replaced by [l/ɹ]

Findings

More [ɹ] responses when exposed to wekke[l/ɹ]

→ Learn to interpret [l/ɹ] as [ɹ]

More [l] responses when exposed to appe[l/ɹ]

→ Learn to interpret [l/ɹ] as [l]

Expectations

Differences particularly between AmbL and AmbR models

Results

- Lexical retuning results
- Inter-segment distances
- Visualisations of the clusters in the hidden layers

Lexical retuning results

Test set = train set

Sound presented	Sound(s) classified (%)
Baseline, retrained model	
[l]	**l(97.6)**, m(3.4)
[ɹ]	**ɹ(95.0)**, sil(5.0)
[l/ɹ]	**l(46.9)**, sil(23.5), ə(19.8), **ɹ(8.6)**, ɛi (1.2)
AmbL model	
[l]	o(78.0), ɔ(10.4), sil(2.4), e(2.4), ɛi (2.4), øː(2.4)
[ɹ]	**ɹ(87.5)**, e(7.5), ʌu(2.5), ɛi(2.5)
[l/ɹ]	**l(81.5)**, ə(12.3), ə(2.5), ɛi(2.5), t(1.2)
AmbR model	
[l]	**l(97.6)**, sil(3.4)
[ɹ]	**ɹ(72.5)**, ə(15.0), sil(10.0), t(2.5)
[l/ɹ]	**ɹ(88.9)**, sil(6.2), ə(4.9)

Inter-segment distances for each layer

For higher layers, distance

→ [l]-[l/ɹ] decreases for AmbL model

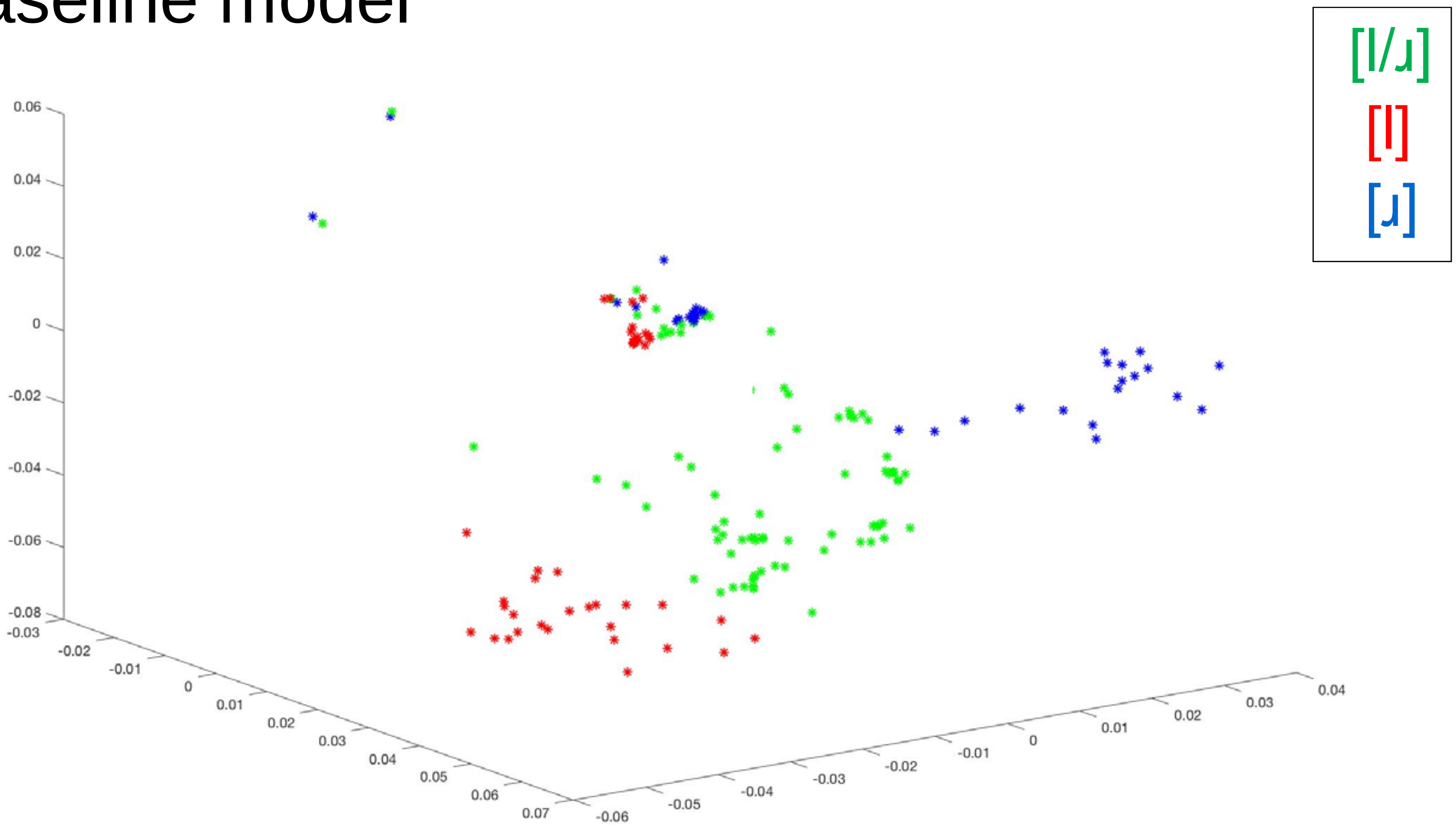
→ [ɹ]-[l/ɹ] decreases for AmbR model

TU Delft

Visualisations of the learned clusters in the hidden layers

PCA visualisations of the 4th hidden layers

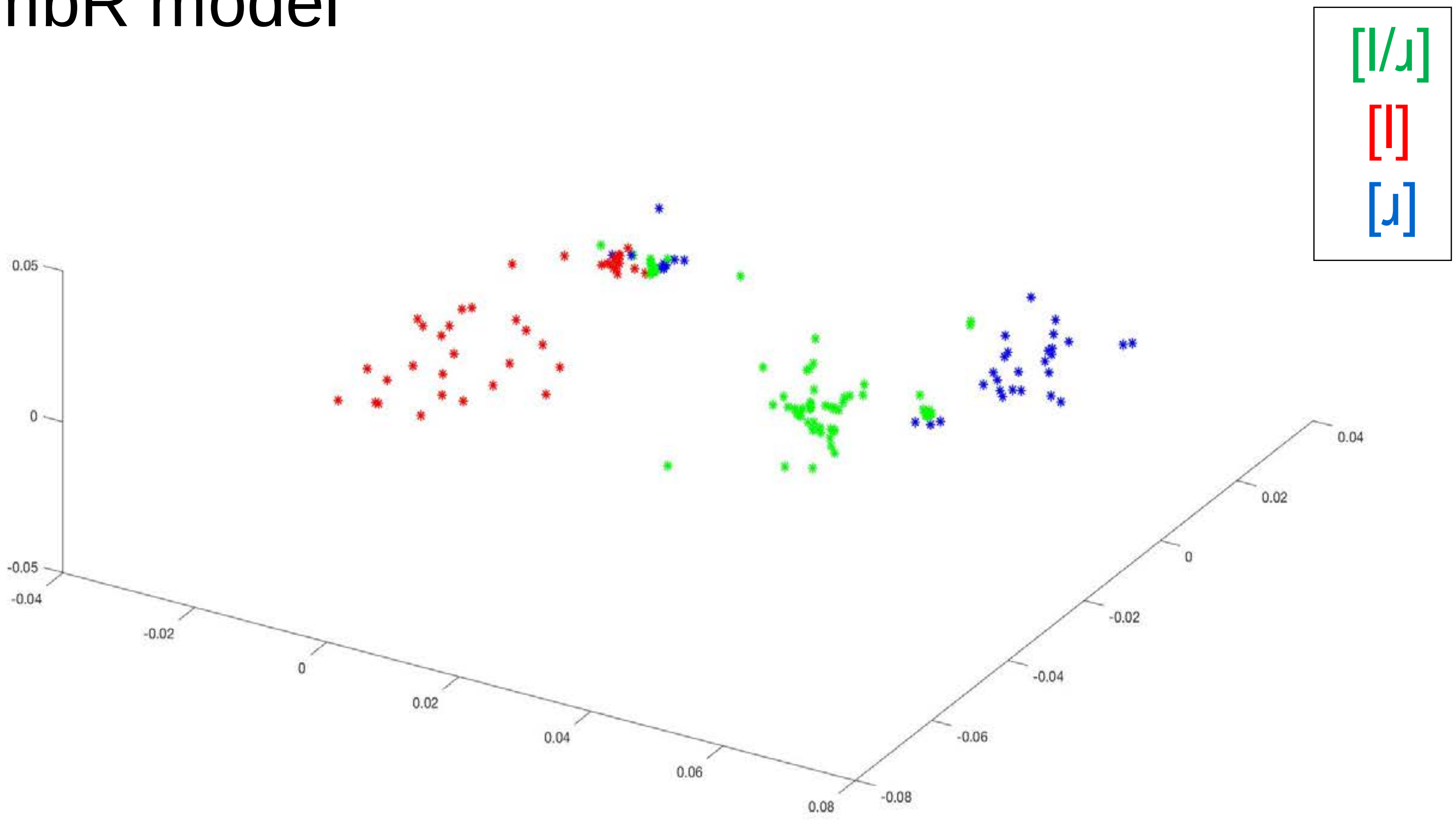
Baseline model



TU Delft

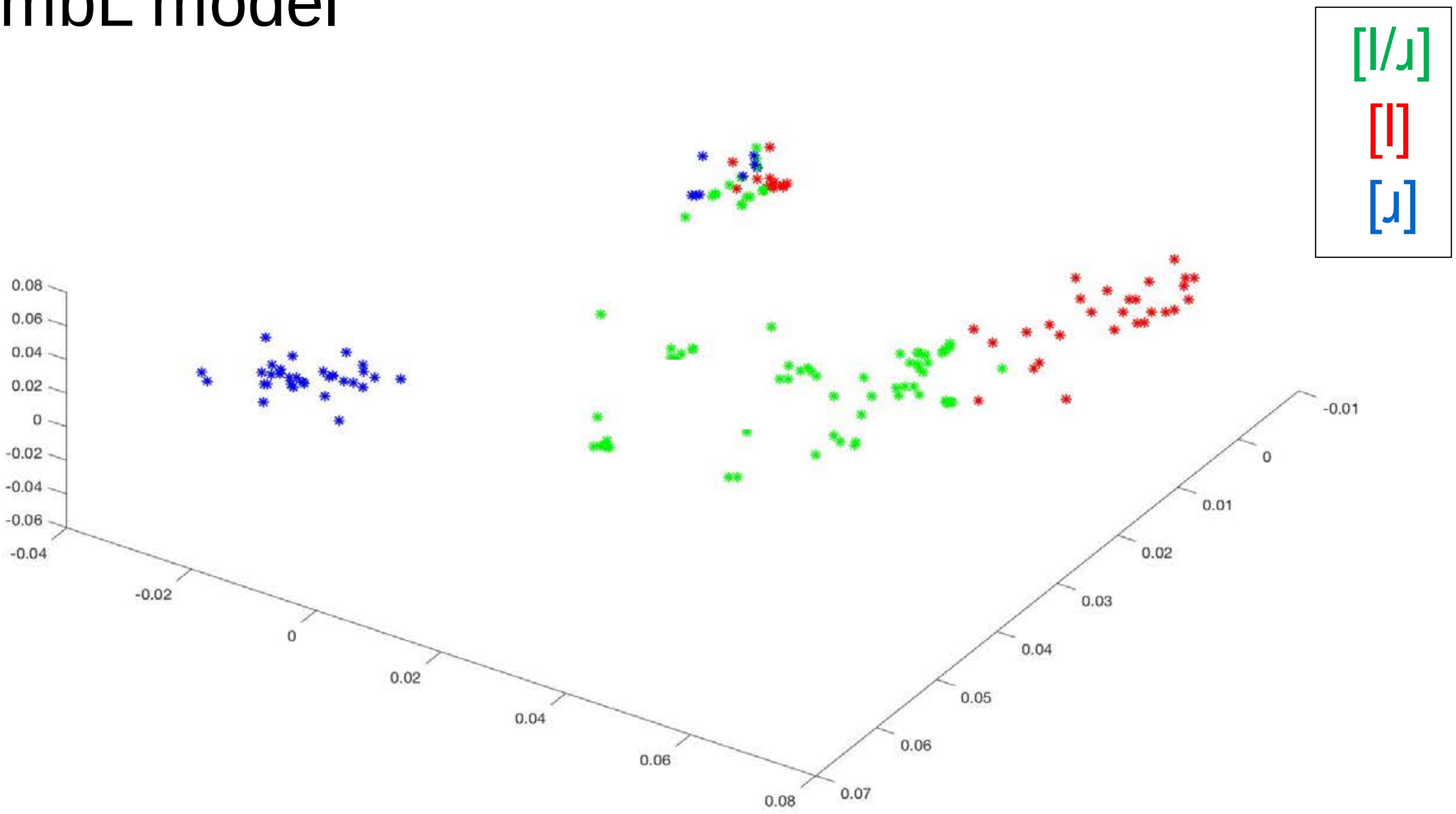
PCA visualisations of the 4th hidden layers

AmbR model



PCA visualisations of the 4th hidden layers

AmbL model



So …

- DNNs trained with ambiguous sounds show perceptual learning

- Not only at the output layer but also at intermediate layers

Perceptual learning in DNNs over time

- Human listeners only need 10-15 ambiguous items

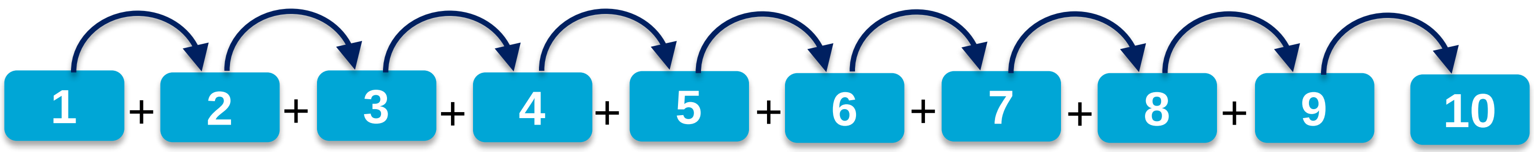
- How about DNNs?

[Ni, Hasegawa-Johnson, Scharenborg, 2019]

Same experimental set-up

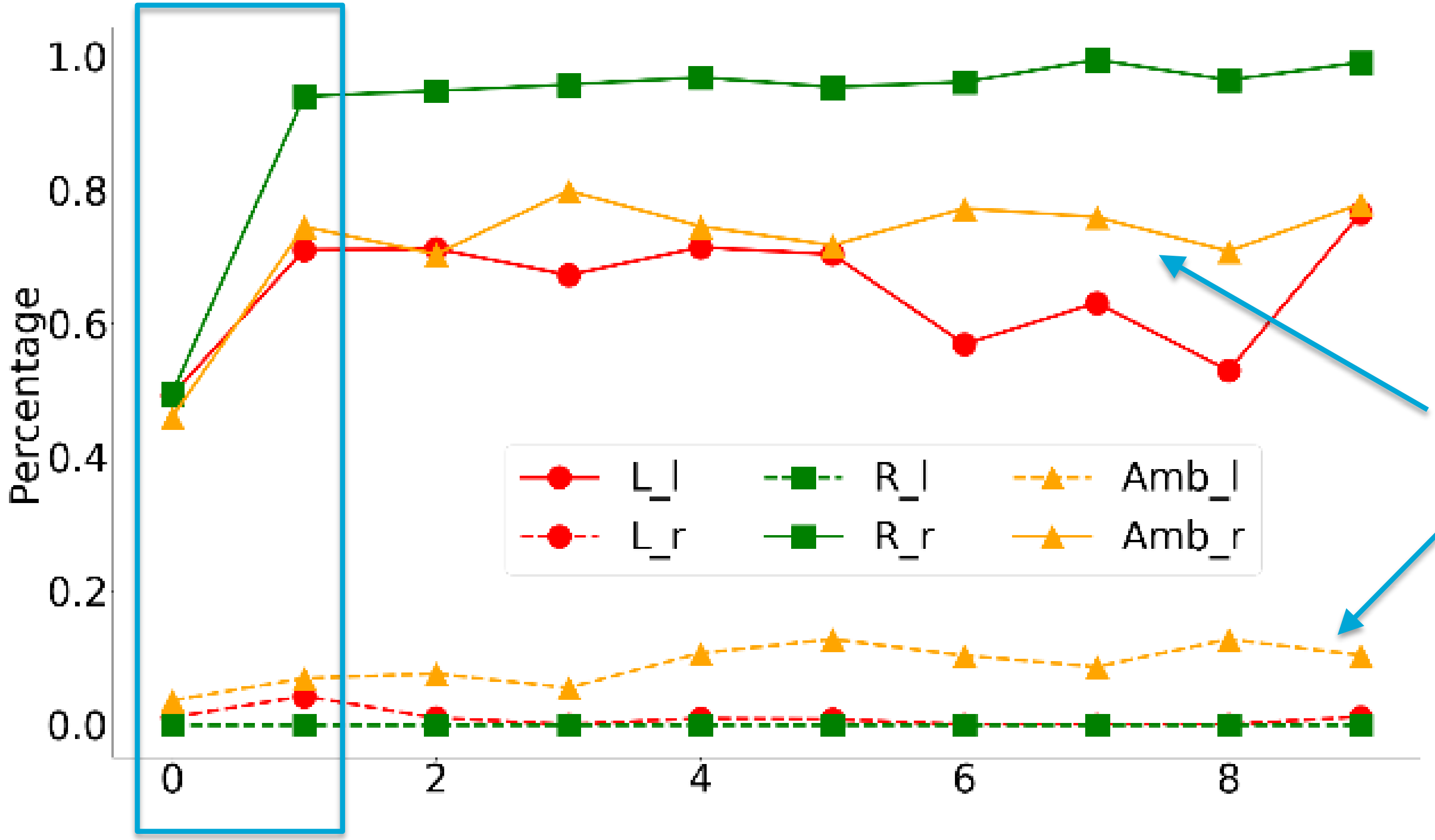
- Retraining in 10 bins of 4 ambiguous items

- Test on unseen data from next bin



Proportion of [ɹ] responses over time

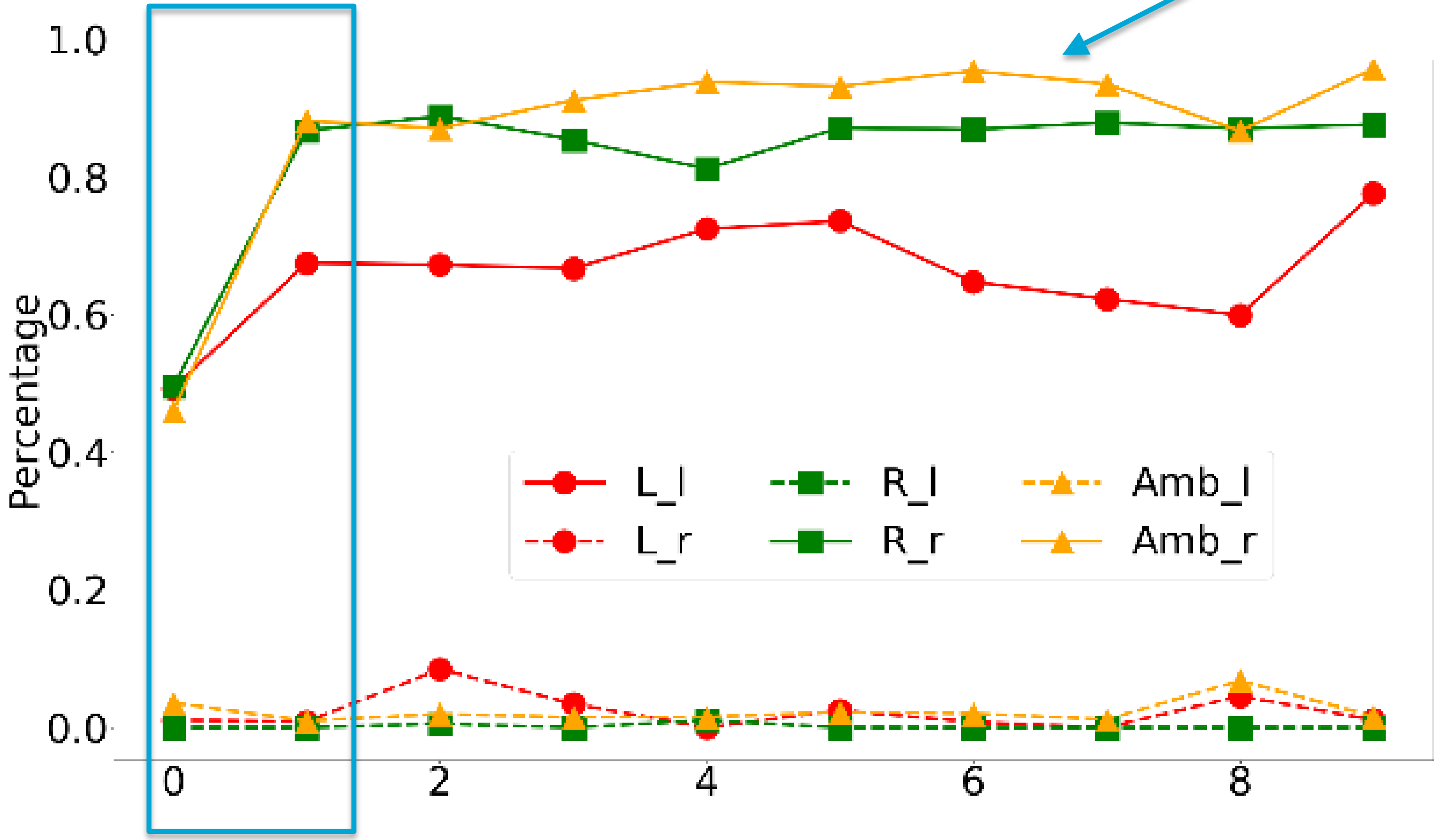
Baseline model



TU Delft

Proportion of [ɹ] responses over time

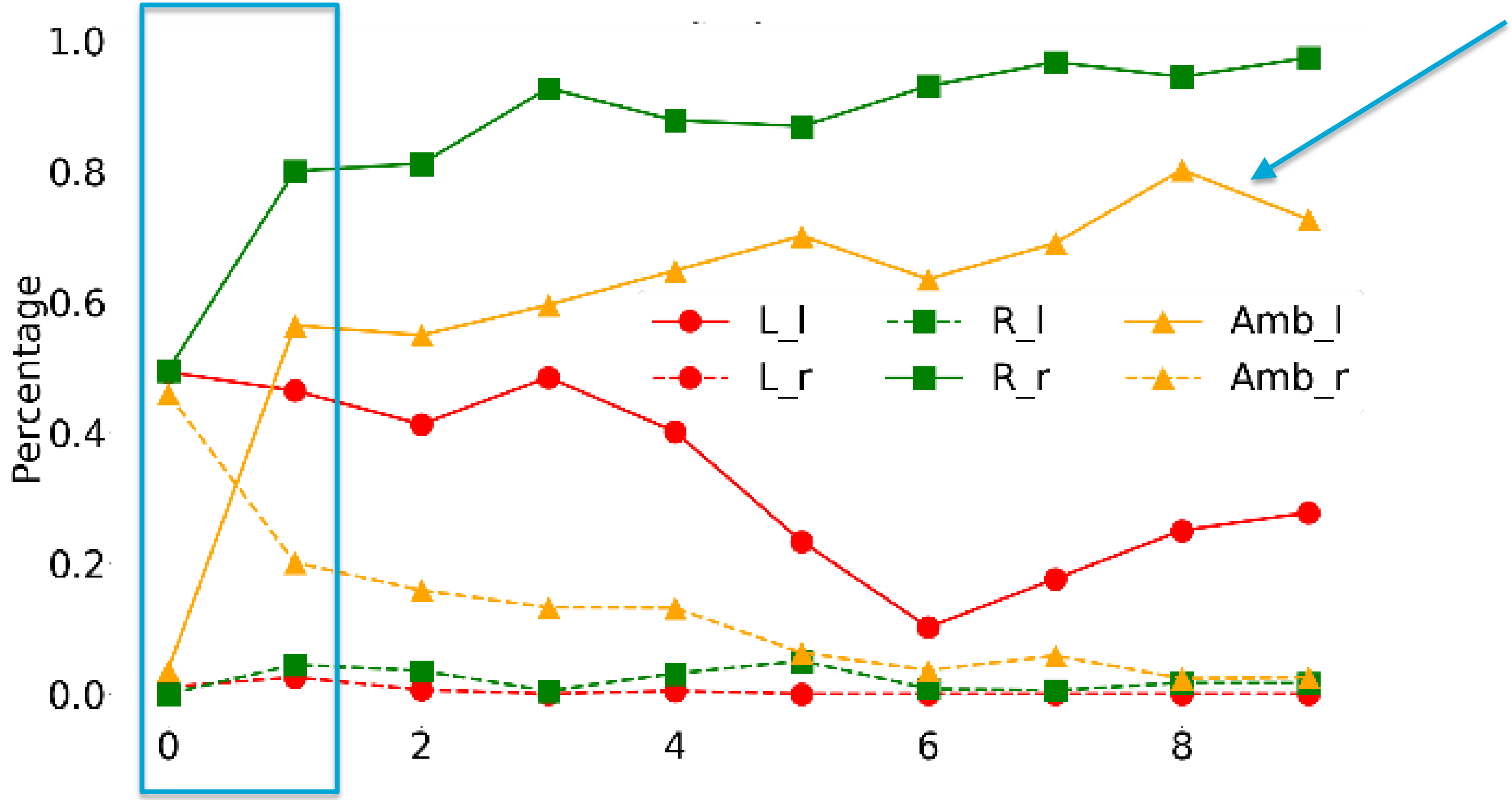
AmbR



TU Delft

Proportion of [ɹ] responses over time

AmbL model



TU Delft

So …

DNNs

- Need only a few ambiguous items

- Show a similar step-like function as humans

Representation of speech in DNNs

- Free reign to the DNN

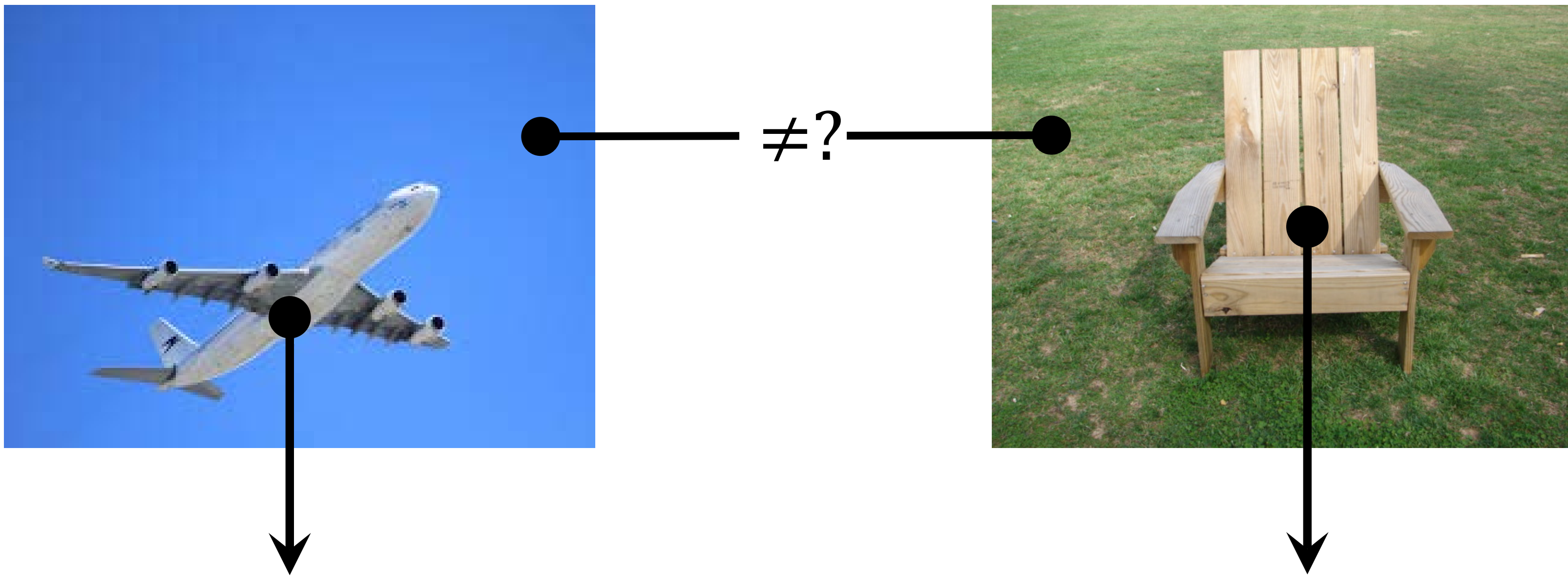
➔What speech representations are learned when faced with the large variabiliy in speech?

[Scharenborg, van der Gouw, Larson, Marchiori, 2019]

Research question

Can a **generic** DNN-based ASR trained to **distinguish high-level speech sounds** learn the **underlying structures** used by human listeners to understand speech?

An example from vision



Extract and learn features associated with planes and chairs?

Methodology

- Naïve, generic feed-forward deep neural network
  - 3 hidden layers with 1024 nodes each
- Corpus Spoken Dutch, 64h read speech
- Consonsant/vowel classification task
- Visualize the activations of the speech representations at the hidden layers
  - Using **different linguistic labels**, to check for clustering

TU Delft

Linguistic labels

- Vowels/consonants differ in the way they are produced: absence/presence of a constriction in the vocal tract
- Phoneme: the smallest unit that changes the meaning of a word e.g. ***ball*** and ***wall***
- Articulatory feature: acoustic correlate of the articulatory properties of speech sounds
  - *Manner of articulation:* type of constriction (e.g., full closure, narrowing)
  - *Place of articulation:* location of the constriction (consonants only)
  - *Tongue position:* location of the bunch of the tongue (vowels only)

C/V classification results

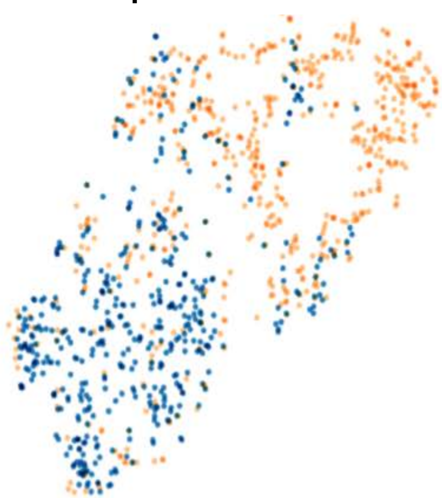
- Frame level accuracy: 85.5%
  - Averaged over 5 training runs
  - Consonants: 85.19% / Vowels: 86.69% correct

Visualisations: V/C classification task

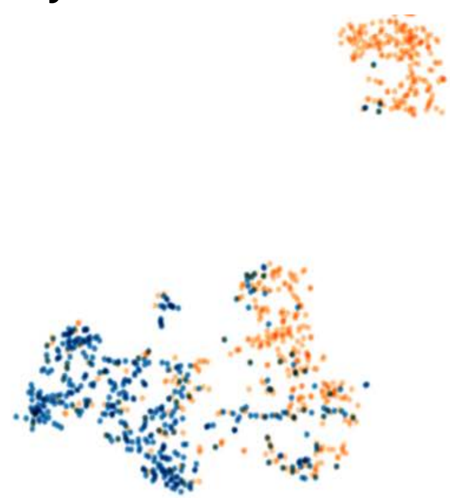
- vowel
- consonant

Original input of the network

Layer 1 of the network



Layer 2 of the network

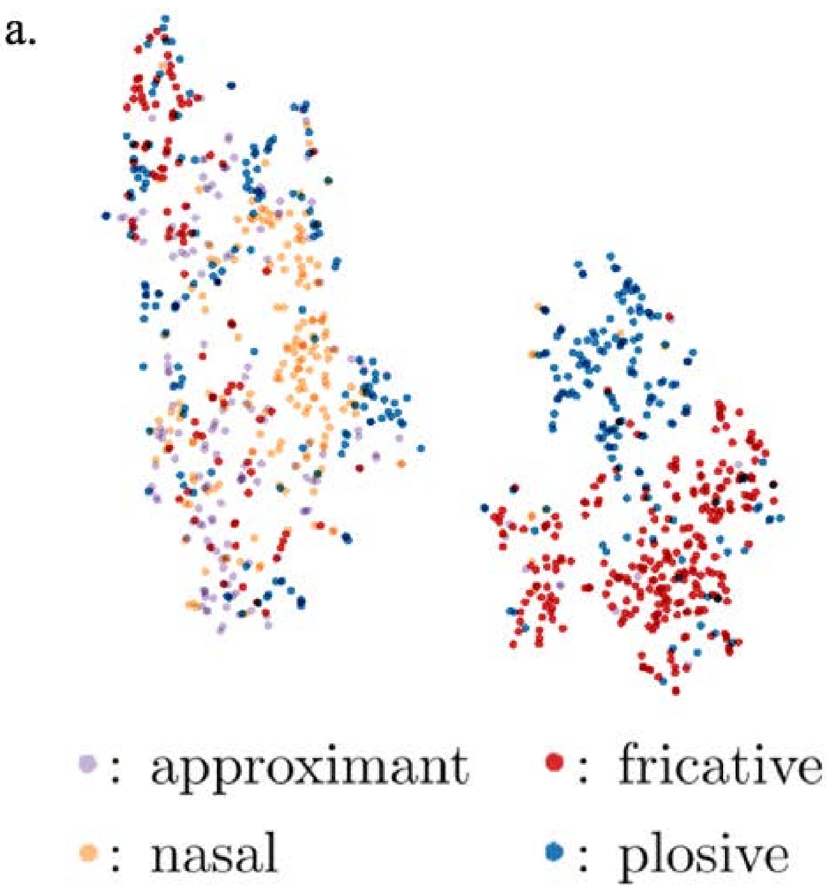


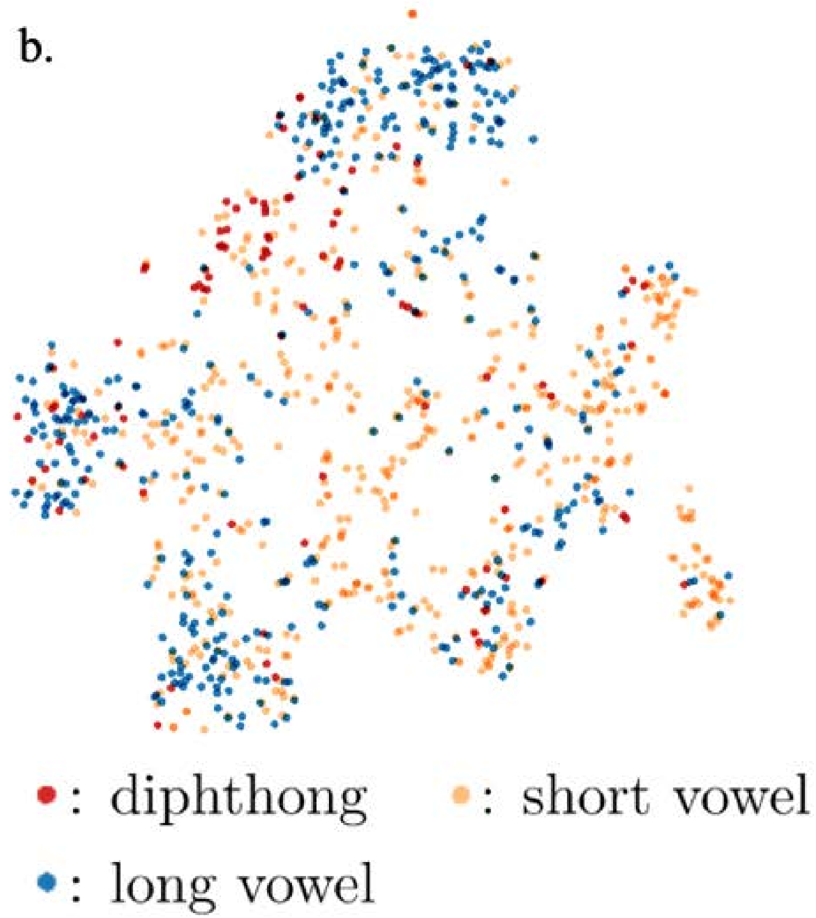
Layer 3 of the network

Manner of articulation, 3rd layer

Consonants

Vowels



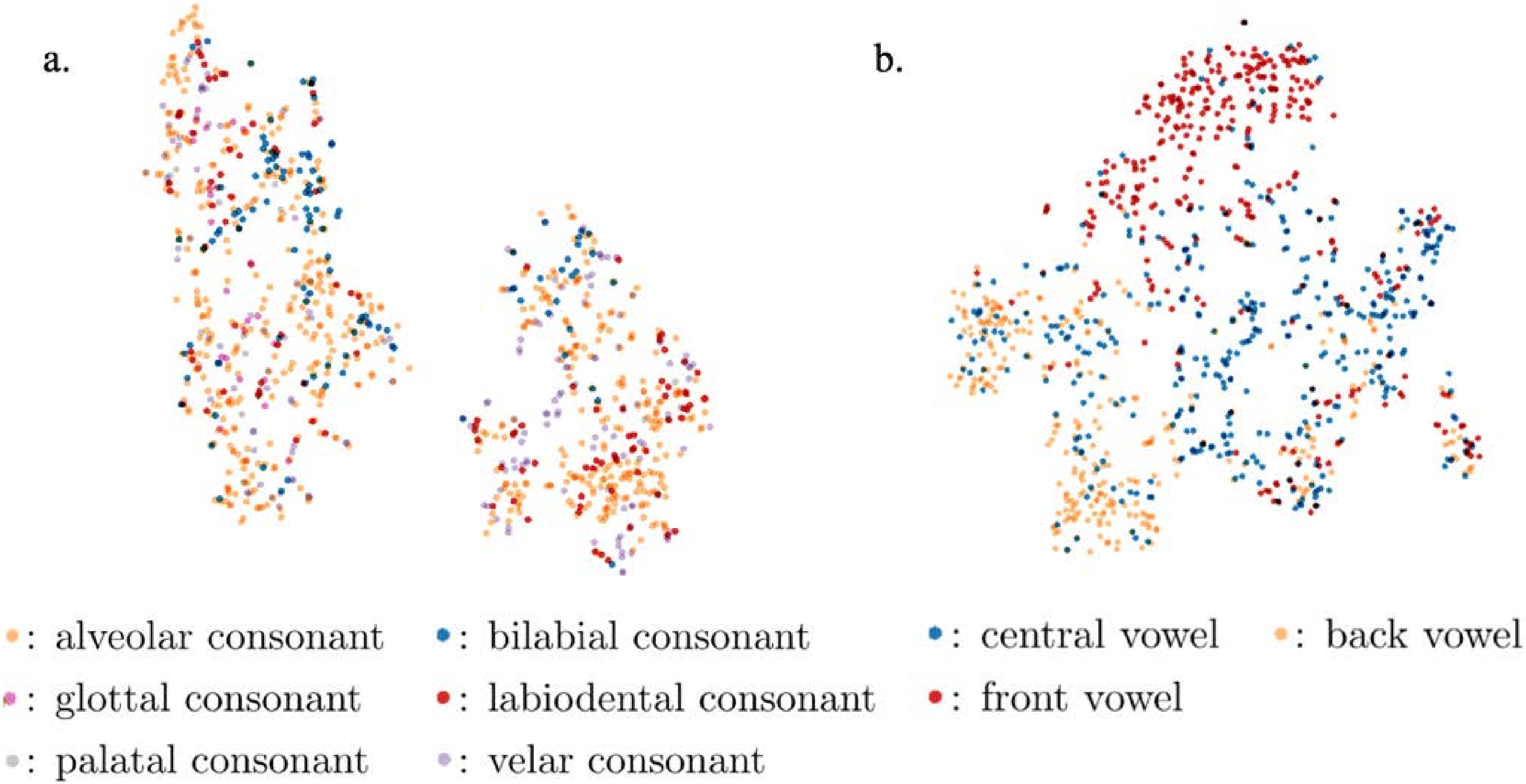


TU Delft

Place of articulation, 3rd layer

Consonants

Vowels



TU Delft

Summary

DNNs

- Progressive abstraction in subsequent hidden layers

- Capture the structure in speech by clustering the speech signal, *without explicit training*

- Adapt to nonstandard speech on the basis of only a few labelled examples, showing a step-like function

TU Delft

Our needs

- Invariant units of speech which transfer easily and accurately to other languages

➔ *DNN phone categories are flexible*

➔ *DNNs automatically create linguistic categories*

- ASR system that can flexibly adapt to new types of speech

➔ *DNNs are able to do that*

Next: ASR system that can *decide* when to create a new sound category, and do so