End-to-end audio encoding and ASR and phonetics

(Interspeech 2023)

Can we understand what Wav2vec2.0 distinguishes?

- In Wav2vec2.0
  - CNN is used as feature encoder
  - transformer layers are used to map the high-dimensional CNN representations to the elements of some lexicon.
- We fine-tuned a Wav2vec2.0 system for producing broad phonetic transcriptions of Dutch.
- We investigate to what extent internal representations of a pre-trained and fine-tuned Wav2vec2.0 model reflect widely-shared phonetic knowledge.

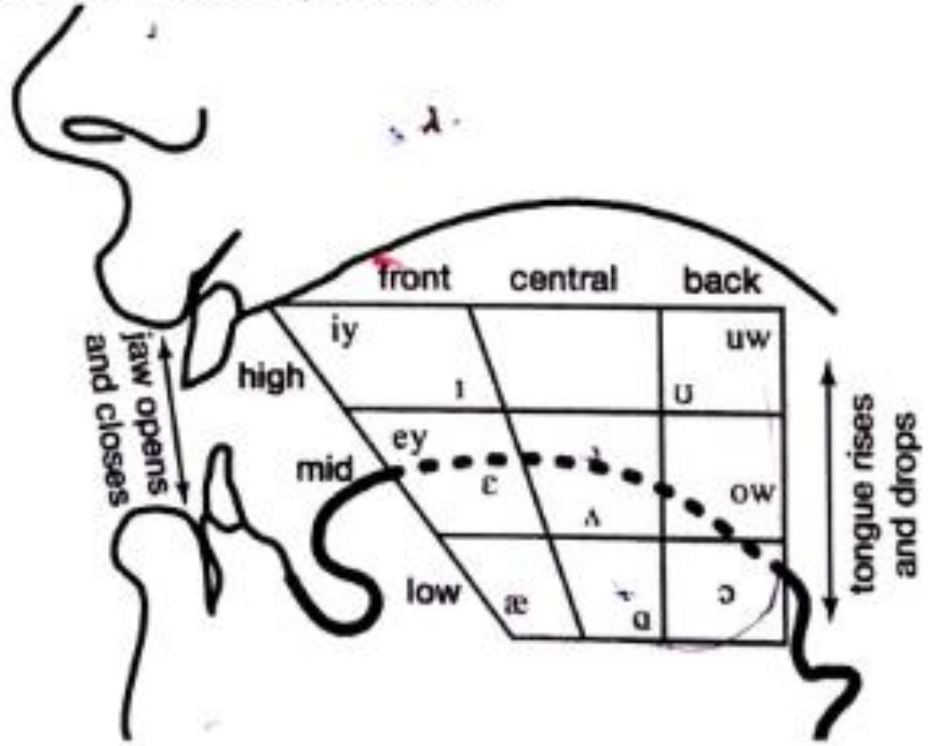
Software (more details later)

- Enormous amount of scripts available on the web
- https://docs.pytorch.org/audio/stable/tutorials/speech_recognition_pipeline_tutorial.html
- Huggingface
- Many more
- You can choose yourself

Vowel structure

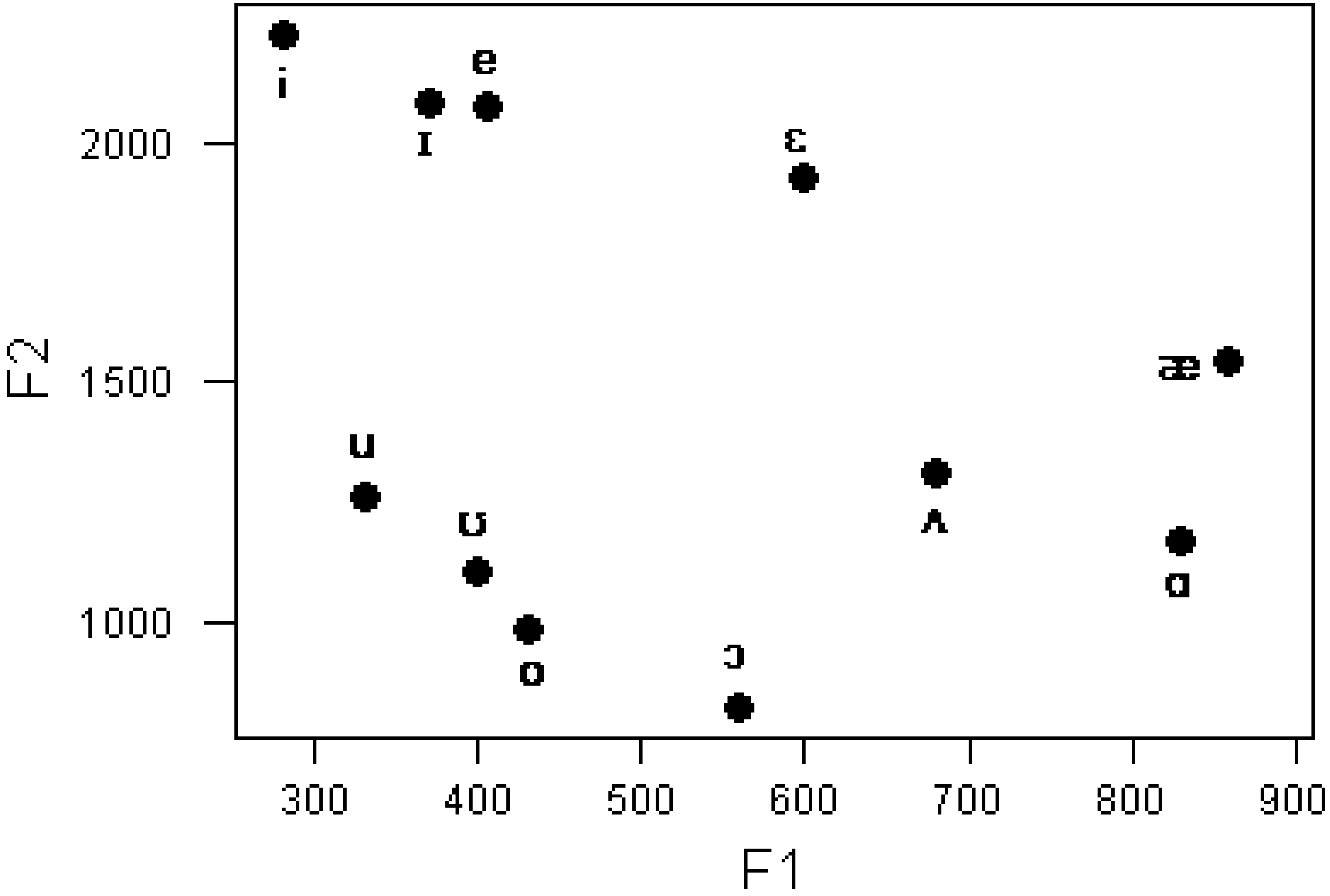
- vowels show phonetic structure

English Vowel System: vowel quadrant



Phonetic theo

- Phonetic theory "predicts" how vowels are organized

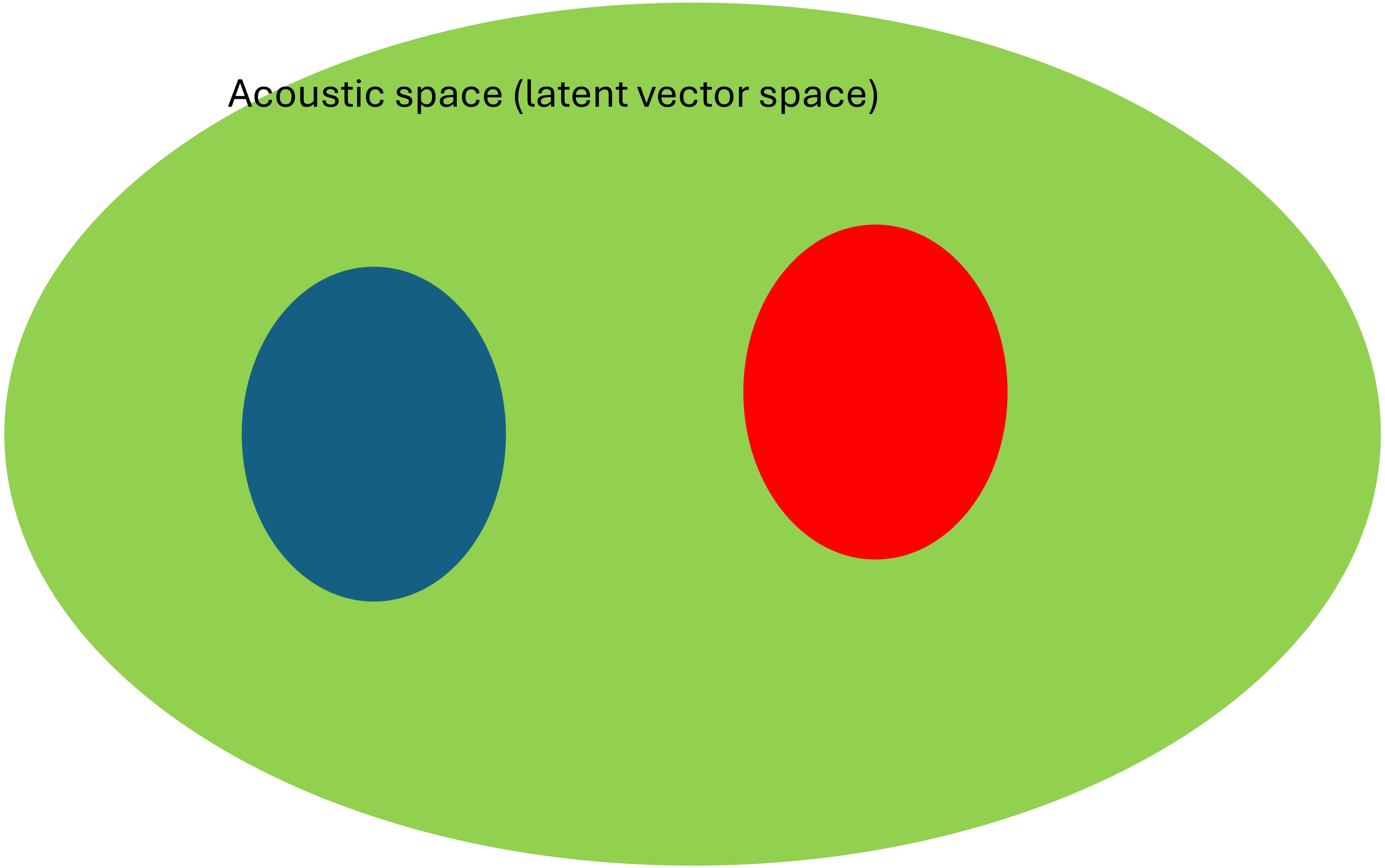


- Can we see the same in the latent representations in Wav2vec2?

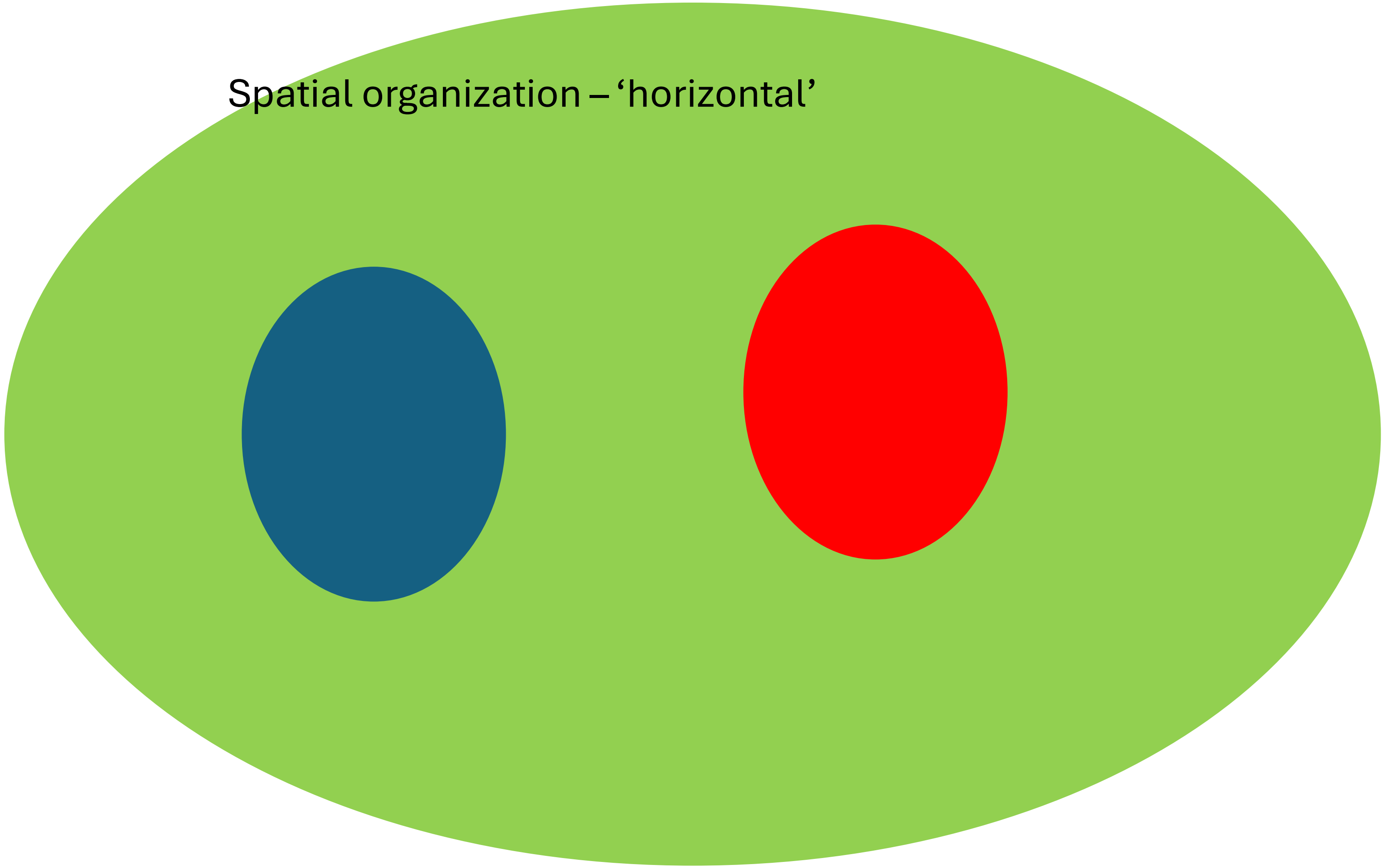
Horizontal versus vertical

- We could view the phone-phone competition in an end-to-end model by looking in three ways:
  - the Kullback-Leibler divergence between phone distributions in some vector space,
  - the within-frame phone ambiguity
  - the interpretation of the competition between winning and runner-up phone in terms of ’classical’ phonetic structure

  - (details see paper)

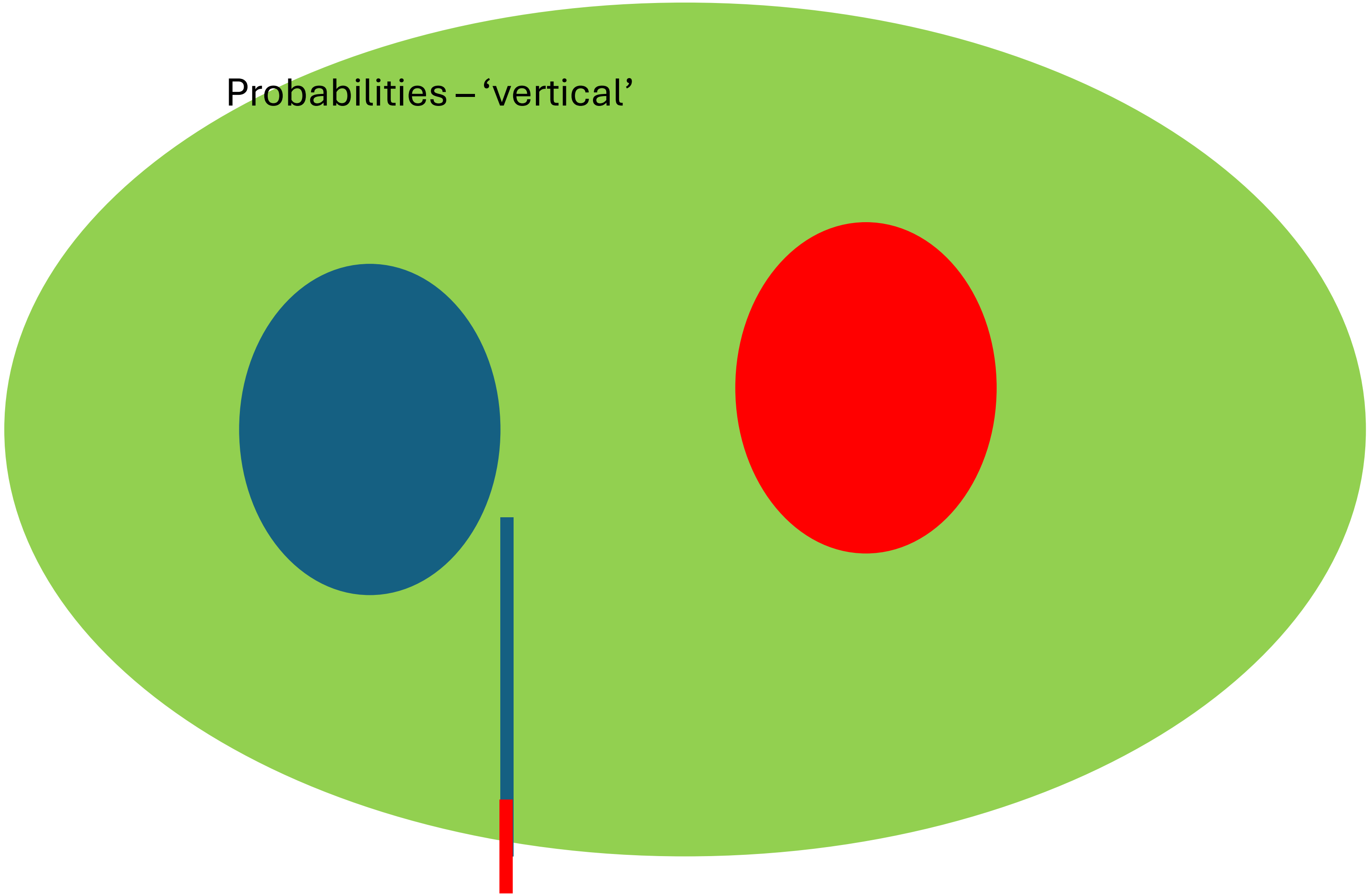


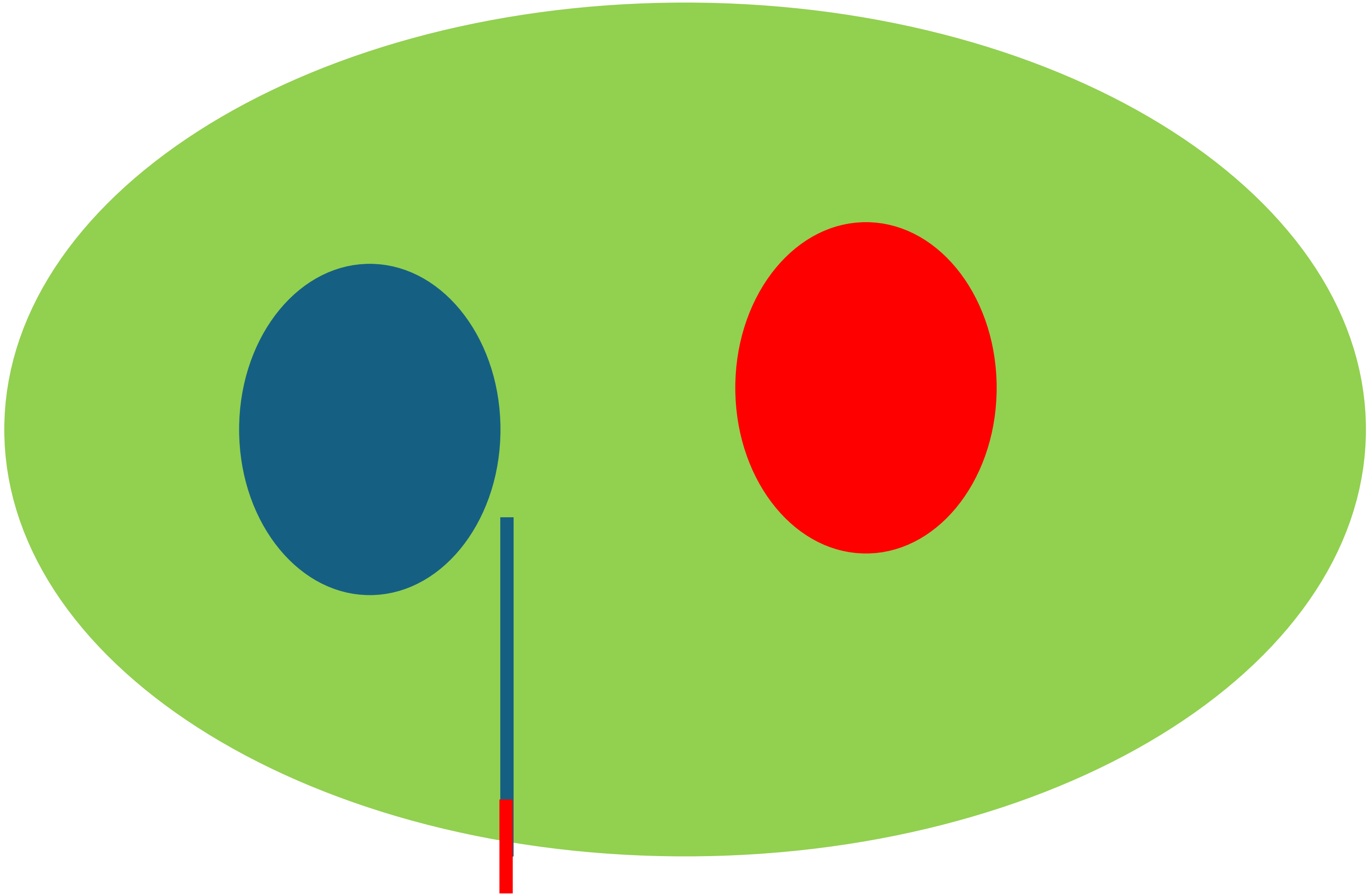
Acoustic space (latent vector space)



Spatial organization – ‘horizontal’

Probabilities – ‘vertical’





Probability vector → competition → entropy

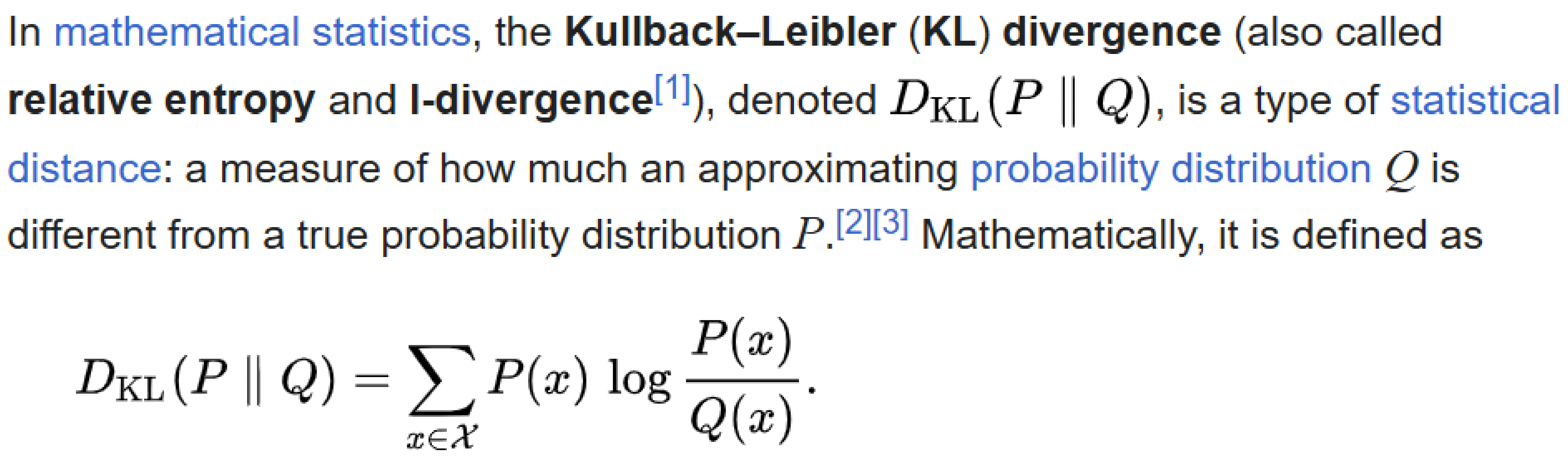
Method 1 (‘horizontal’, ‘spatial’)

- Store the 1024-D representations on all hidden layers for all utterances
  - And the 512-D CNN representations, which are as close as we can come to the speech signals

- KMeans clustering to create codebooks with 512 entries
  - Euclidean distance

- Each phone → a histogram that describes p(state | phone)

- the Kullback-Leibler (KL) divergence for all phone pairs



In mathematical statistics, the **Kullback–Leibler** (**KL**) **divergence** (also called **relative entropy** and **I-divergence**[1]), denoted $D_{\text{KL}}(P \parallel Q)$, is a type of statistical distance: a measure of how much an approximating probability distribution Q is different from a true probability distribution P.[2][3] Mathematically, it is defined as

$$D_{\text{KL}}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \, \log \frac{P(x)}{Q(x)}.$$

KL divergence

- Figure on next slide shows the (non-symmetric) KL divergence for all phone pairs for the representations on the CNN output layer and layers nr 12 and 24 in the transformer.
- It can be seen that on the CNN layer there is some evidence of acoustic-phonetic structure, but on the (higher) transformer layers that structure is much less apparent.
- Importantly, this does not mean that those representations contain weaker phonetic information.

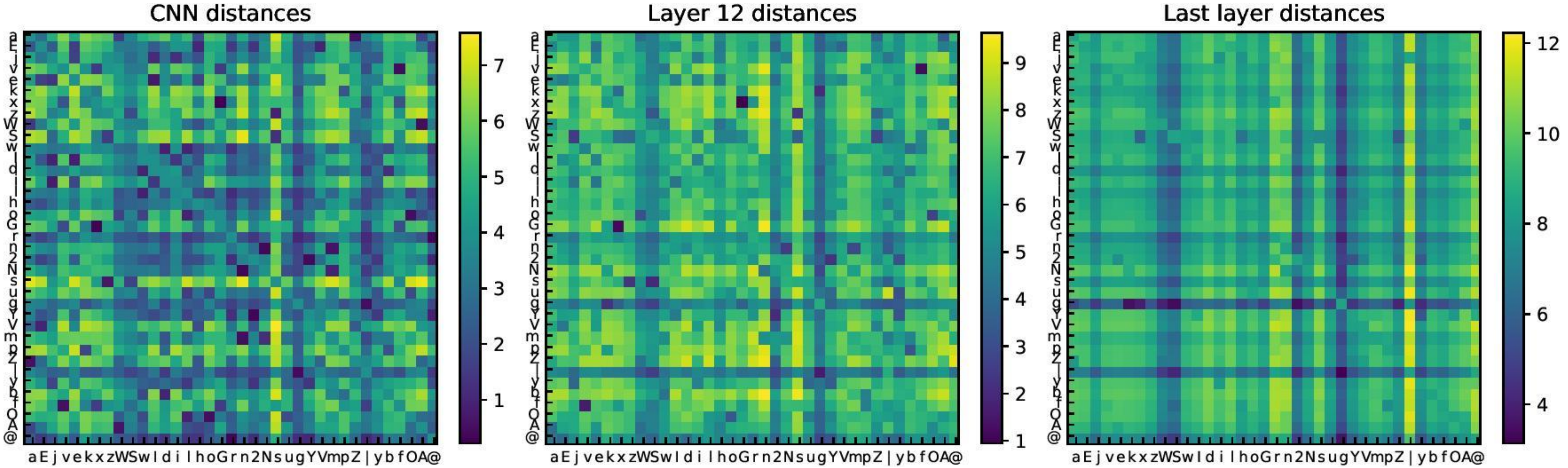


Figure shows (non-symmetric) KL divergence for all phone pairs
for the representations on the CNN output layer and layers nr 12 and 24 in the transformer

CNN layer: there is some evidence of acoustic-phonetic structure,
but on the (higher) transformer layers that structure is much less apparent

- Figure 2 shows that, across the board, there is a sub-stantial difference between BPCs and the random phone sets in terms of the KL divergence with the observed MLP-based phone PDFs.
- Clearly, the comparison differs for the different layers in the Wav2vec2 model, and is different for the pretrained and finetuned variant.
- Interestingly, the KL-divergence does not differ for layer 24 from the CTC fine-tuned model, while the MLP phone classifier performed best in the phone classification

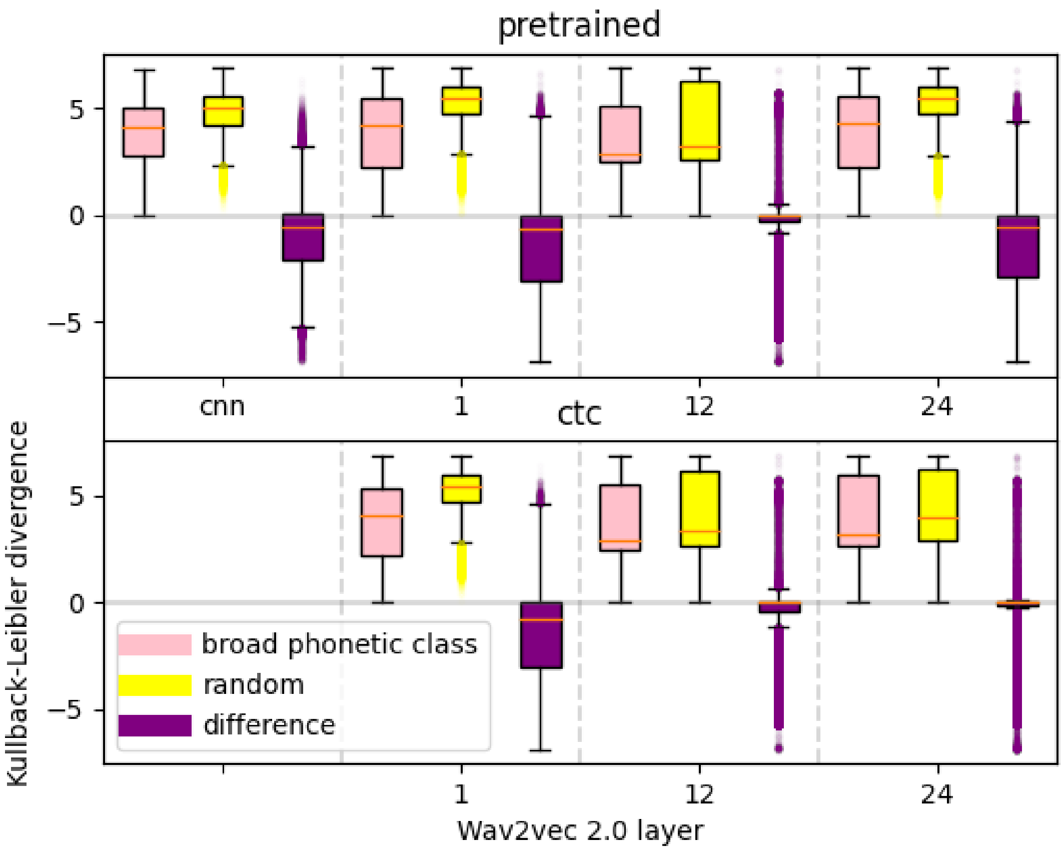


Figure 2: The distribution of KL-divergences scores for BPC PDF, random PDF and the difference between the KL-divergence scores for BPC and random.

The negative values are only present for the difference between the BPC and random based KL-divergence scores

Clearly, the comparison differs for the different layers in the Wav2vec2 model, and is different for the pretrained and finetuned variant. Interestingly, the KL-divergence does not differ for layer 24 from the CTC fine-tuned model, while the MLP phone classifier performed best in the phone classification

Probing (‘horizontal’)

- We show that MLP-based classifiers perform adequately on the latent representations for all layers while the internal phone-phone structure shows a varying pattern across layers.
- This suggests that while phone classifiers are able to produce a likely correct winner, given a latent representation, a phonetically motivated phone-phone structure in the latent space is not necessary.
- Details see paper

Probing classifier (MLP)

The table presents the **phone classification accuracy** (%) on an independent held out test set from component ’o’ in the Spoken Dutch Corpus

Layer	Pretrained	Finetuned (CTC)
CNN	69	
1	84	83
12	92	94
24	94	98

Method 2: ‘vertical’

- Given trained MLPs, we study the competition between phones according to the MLP output vector.
- by counting the number of times a given phone appears as the runner-up of a winning phone.
- Figure 3 gives an overview of the percentage of runner-up phones sharing the same BPC as the winning phone, across all MLP-output vectors on a particular layer.
- BPC = broad phonetic class
  - {p,t,k}, {b,g,d}, {s,z} {f, v} {i, e}, {..}

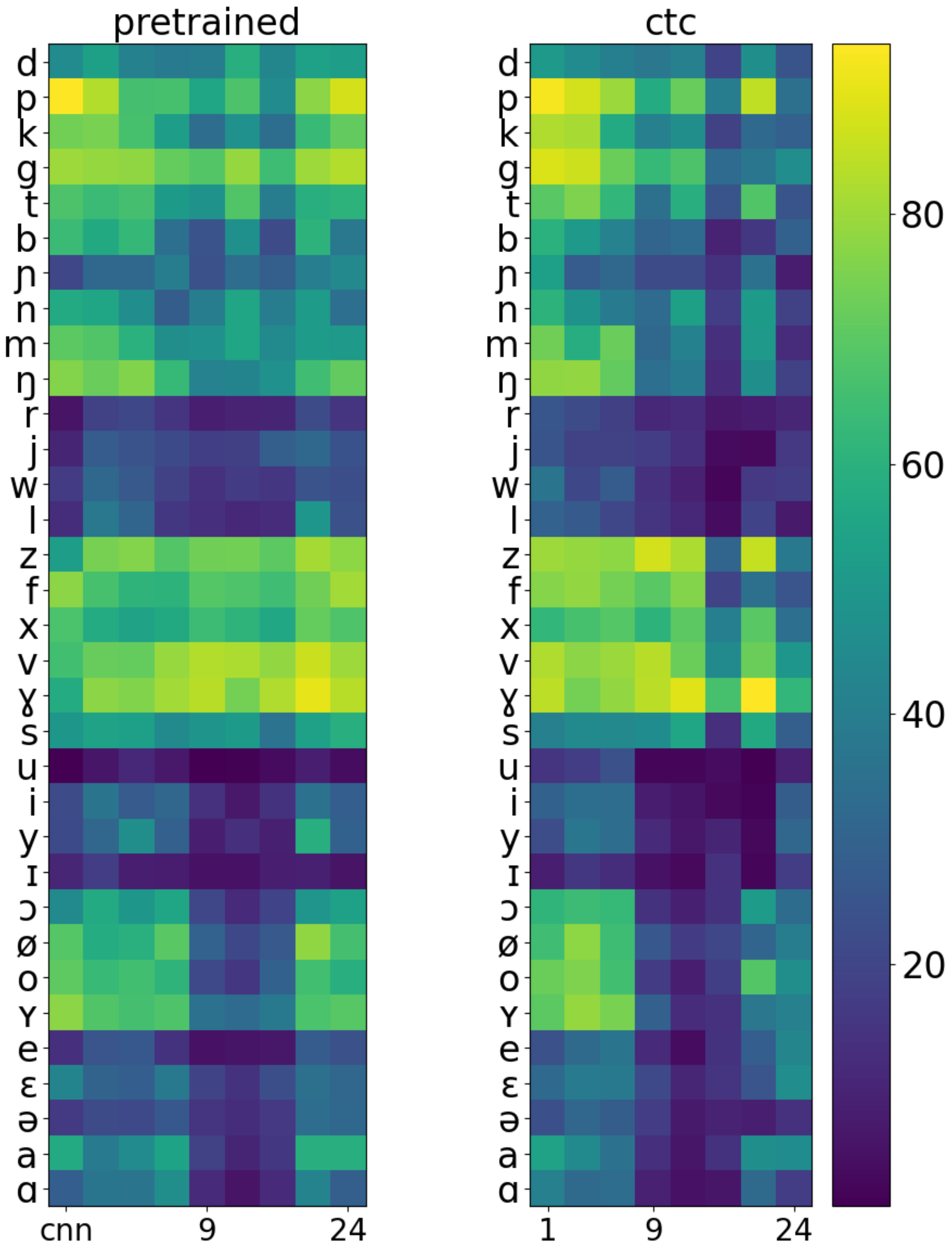
Phone–phone competition

- Fig. 3 shows the intrinsic phone-phone competition per MLP input vector, that is, at a given location in the latent space.
- It raises the question to what extent these runners-up reveal information about the intrinsic phonetic structure. Evidently it can be expected that the overall statistics of the runner-up is informed by the phonetic neighborhood of the winning phone.
- For the plosives,nasals and fricatives this occurs fairly often even in the middle layers of the model, while this is not the case for the high, some middle and low vowels.

Figure 3: The percentage of second-best phones sharing the same broad phonetic class as the winning phone (y-axis), as classified by the multi-layer perceptron trained on a given layer(x-axis) of the Wav2vec2 model.

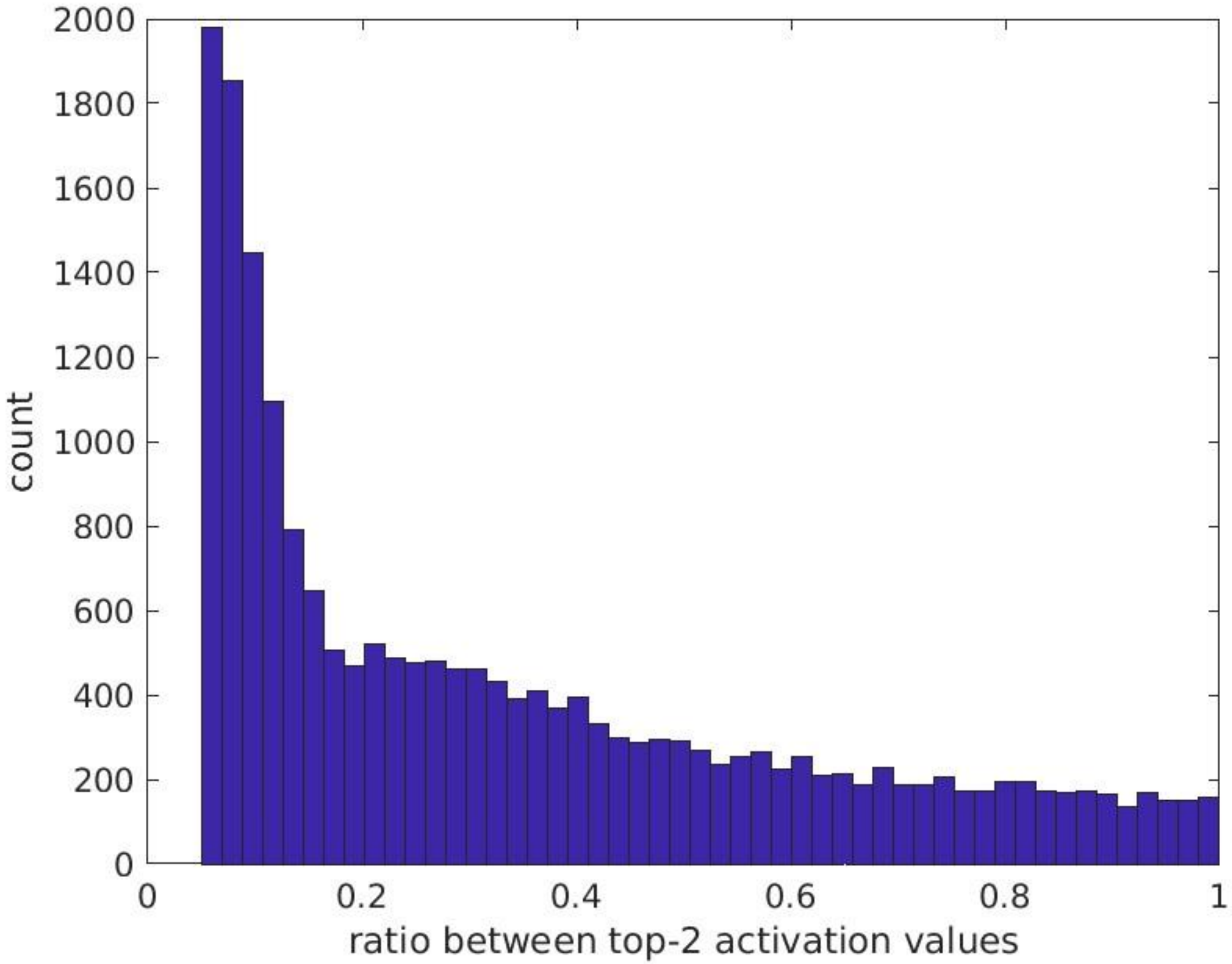
Three million (winner, runner-up) pairs

Left: pre-trained
Right: ctc fine-tuned.



Probability → distance

- $d \sim 1/(N + R)$
- in which d denotes the distance between winning phone and runner-up, N denotes the number of counts of the runner-up given the winner phone, and R is a regularisation term that corrects for very infrequent combinations.
  - These infrequent combinations will certainly appear since, for example, /p/ will hardly occur as second candidate after /a/.
  - In those cases, the formula predicts a large distance d between /a/ and /p/, and its numerical upperbound is determined by R.
- Next, we use multi-dimensional scaling (MDS) to build a global 2D map, based on the phone-phone distances d.
  - We used MDS rather than other visual-ization methods such as t-Distributed Stochastic Neighbor Embedding (t-SNE) since we focus on the local and global relations between phones instead of on clustering.



Study 3. Competition

This figure shows the competition between the winner phone and the runner up in terms of their scores in the logit matrix (the ratio)

The bulk has a low entropy

Results suggest that

- MLP-based classifiers perform adequately on the latent representations for all layers
- phone-phone structure shows a varying pattern across layers

→a phonetically motivated phone-phone structure in the latent space is not necessary

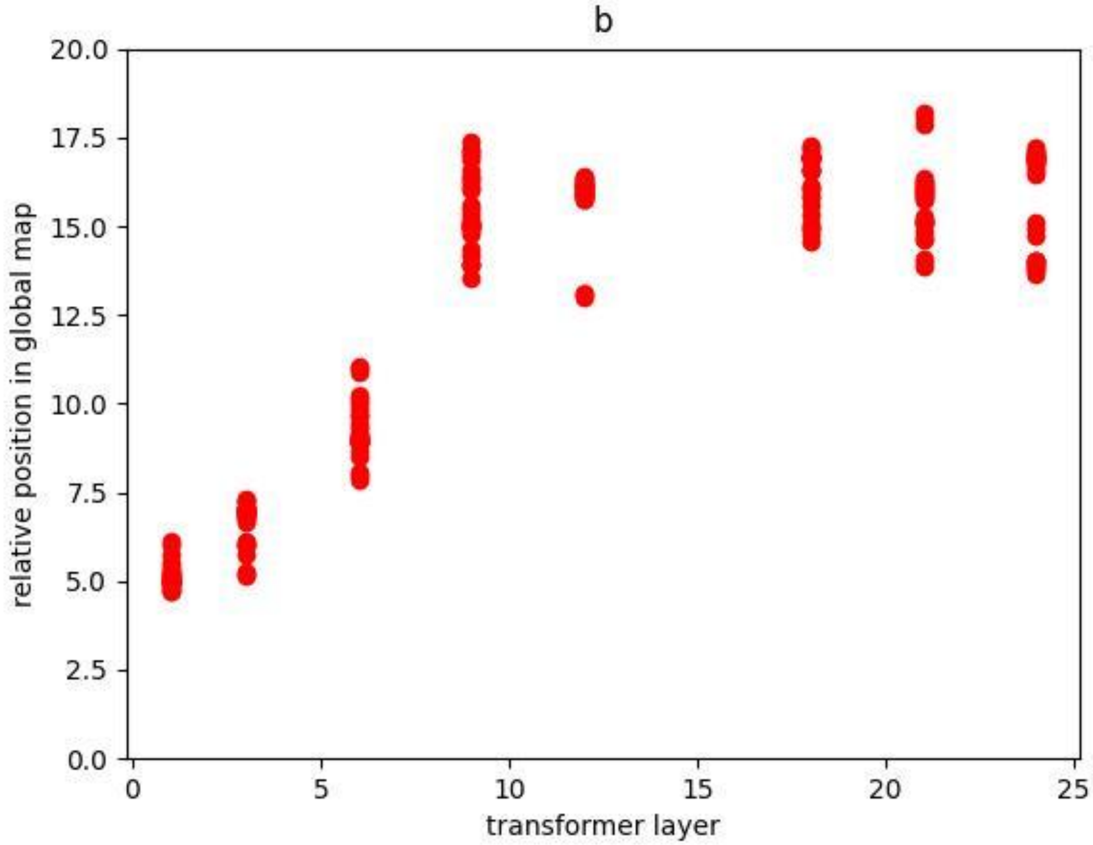
In order to investigate this further, we study the competition between phones according to the MLP classification

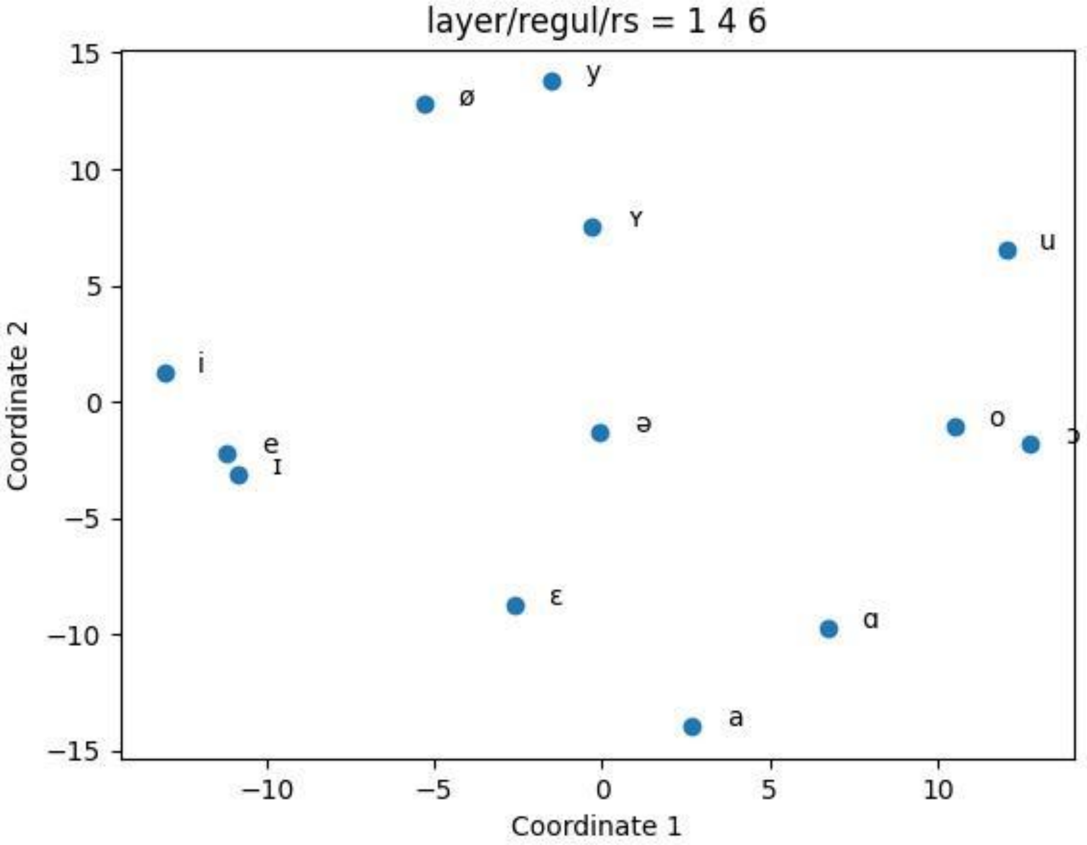
(winner, runner-up) statistics

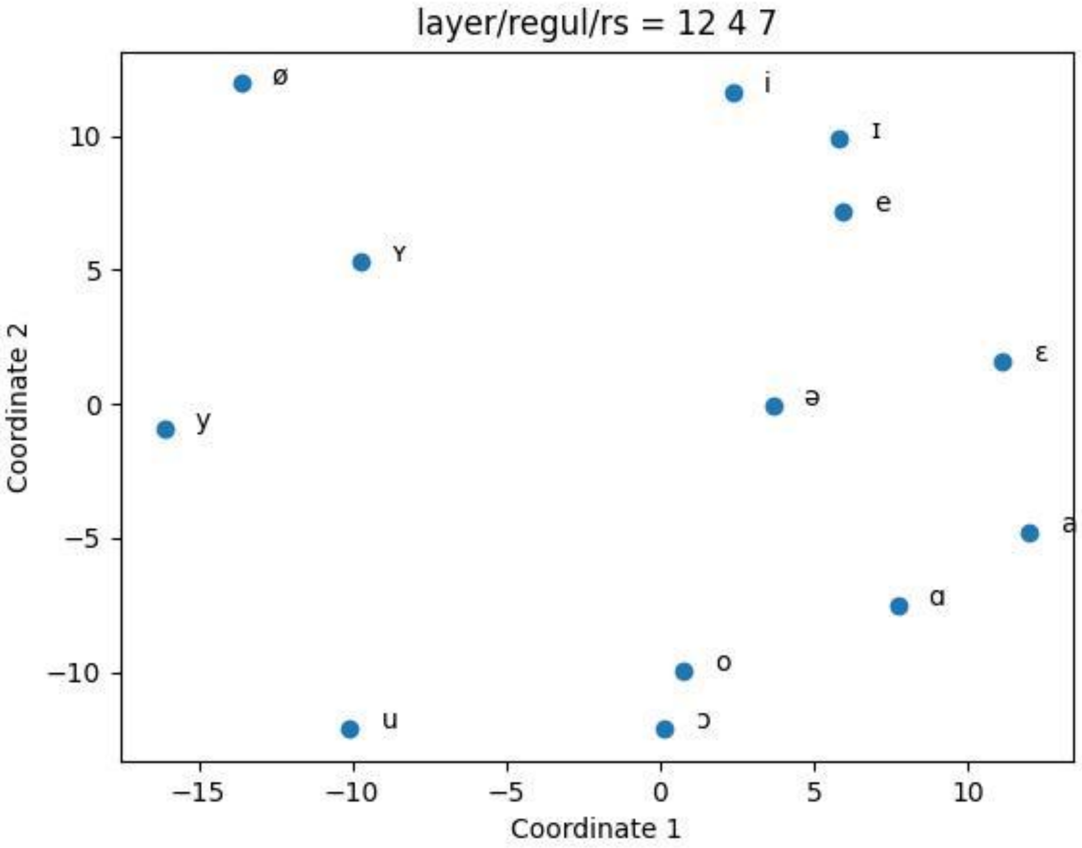
- We zoom in into the statistics of the recognized phone and its runner-up, by counting the number of times a given phone appears as the runner-up of a winning phone

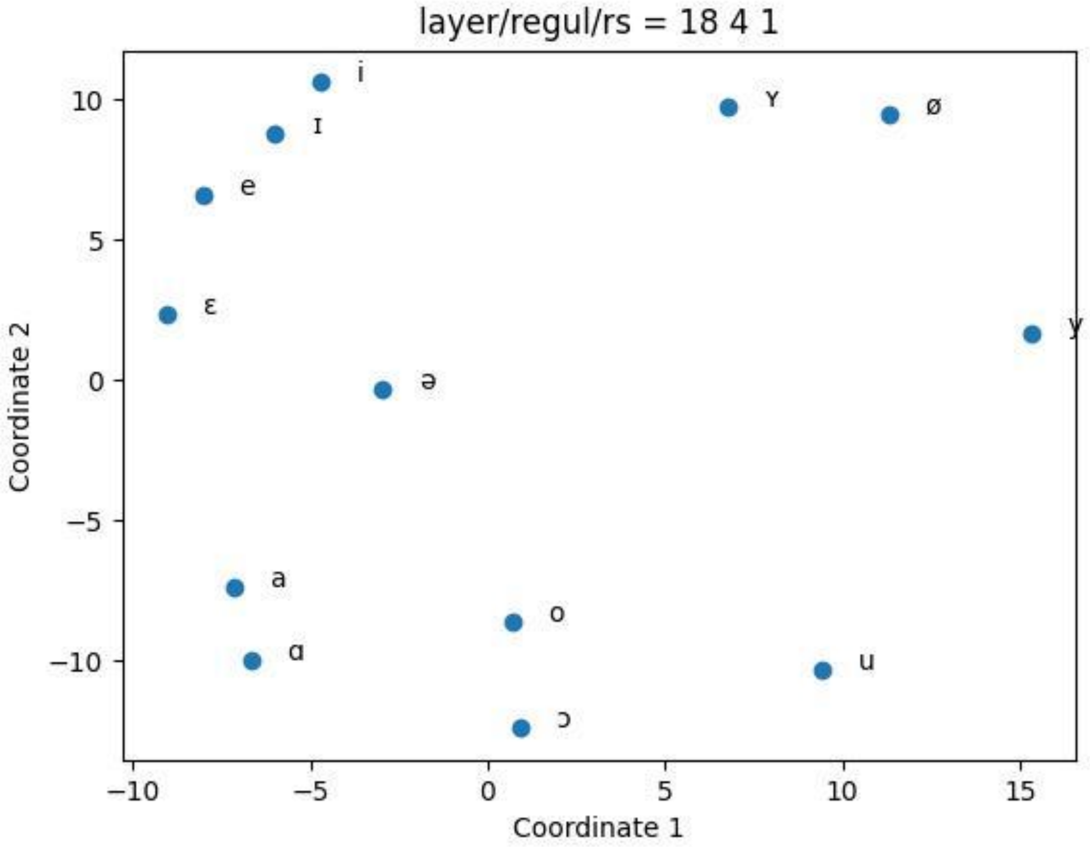
N(runner_up | winning phone)

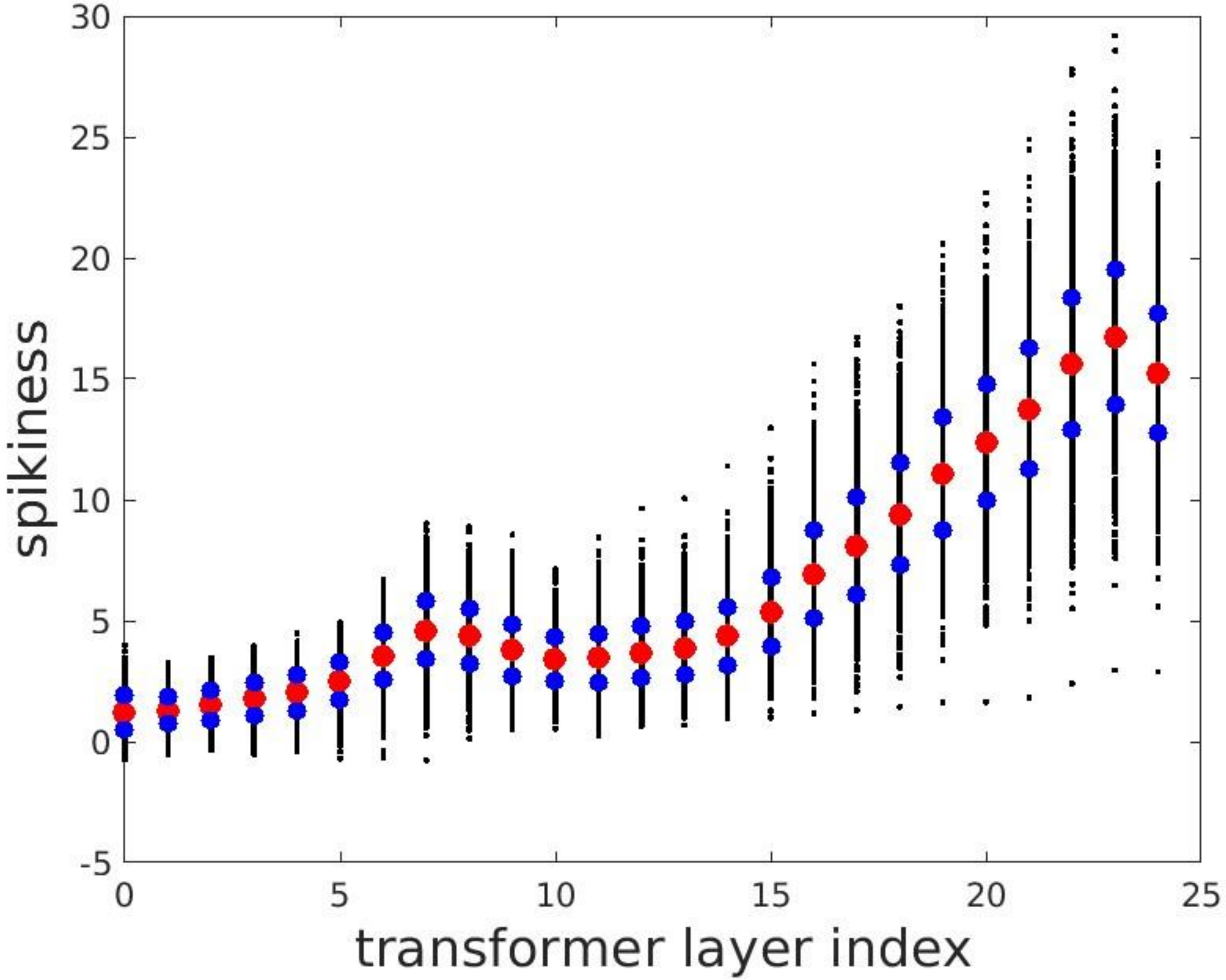
- Next figure gives a pictorial overview of the percentage of runner-up phones sharing the same broad phonetic class (BPC) as the winning phone, a cross all MLP-output vectors on a particular layer.


b
relative position in global map
transformer layer
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
0
5
10
15
20
25


layer/regul/rs = 1 4 6
ø
y
ʏ
u
i
ə
o
ɔ
e
ɪ
ɛ
ɑ
a
15
10
5
0
−5
−10
−15
−10
−5
0
5
10
Coordinate 1
Coordinate 2

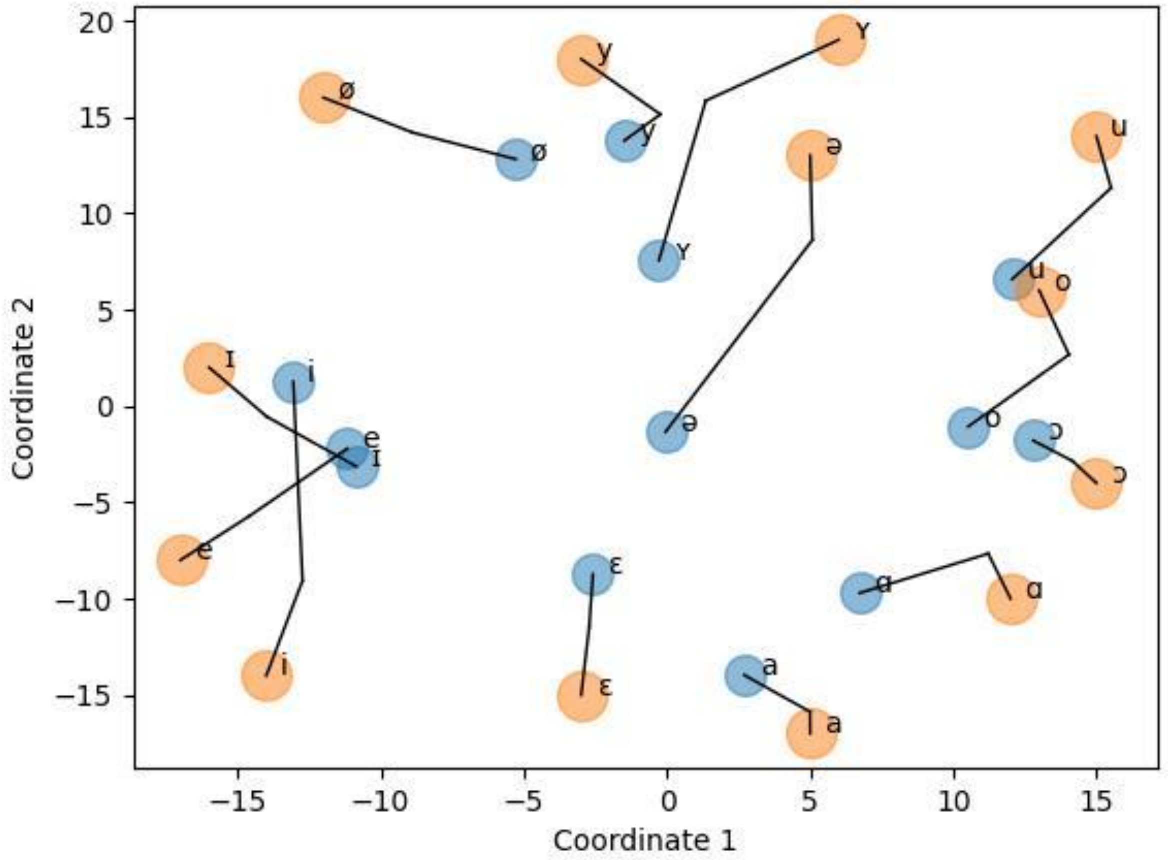

layer/regul/rs = 12 4 7
ø
i
ɪ
e
ʏ
ɛ
ə
y
a
ɑ
o
u
ɔ
10
5
0
−5
−10
−15
−10
−5
0
5
10
Coordinate 1
Coordinate 2


layer/regul/rs = 18 4 1
i
ɪ
e
ɛ
ə
a
ɑ
o
ɔ
u
ʏ
ø
y
Coordinate 2
10
5
0
−5
−10
Coordinate 1
−10
−5
0
5
10
15



spikiness
transformer layer index
30
25
20
15
10
5
0
-5
0
5
10
15
20
25

Figure 4:
MDS locations of Dutch vowels (monophthongs), after averaging over 50 different MDS runs;
PCA-based corrective rotations are used before creating the overlay.
The smaller light-blue markers denote the positions based on transformer layer 1; the larger light-orange markers denote the positions at layer 18; the connecting lines indicate the (simplified) route of the MDS solutions going from transformer layer 1 to transformer layer 18.
(Applicability of the Euclidean distance assumed.)



Preparing for linear separability??

- Figure 4 shows the MDS positions of vowels based on the distances obtained via Eq. 2 from the runner-up distribution in layers 1 and 18 of the transformerblock, respectively.
- For layer 1 (represented by smaller light-blue markers) the vowel triangle (cf. [5]) is clearly respected,and the topology of the resulting configuration is very close tothe Dutch vowel system.
- For higher layers, the MDS solution moves away from the configuration found for layer 1 – vowels are pushed apart towards a much more convex vowel configu-ration.
- An increasing convexity of the configuration can be understood as preparing the latent representation for enabling the ultimate layer to make decisions using linear hyperplanes.

Results

- Finetuned Wav2vec2 models can produce high-quality transcriptions on the phone level (8.3% on TIMIT test).

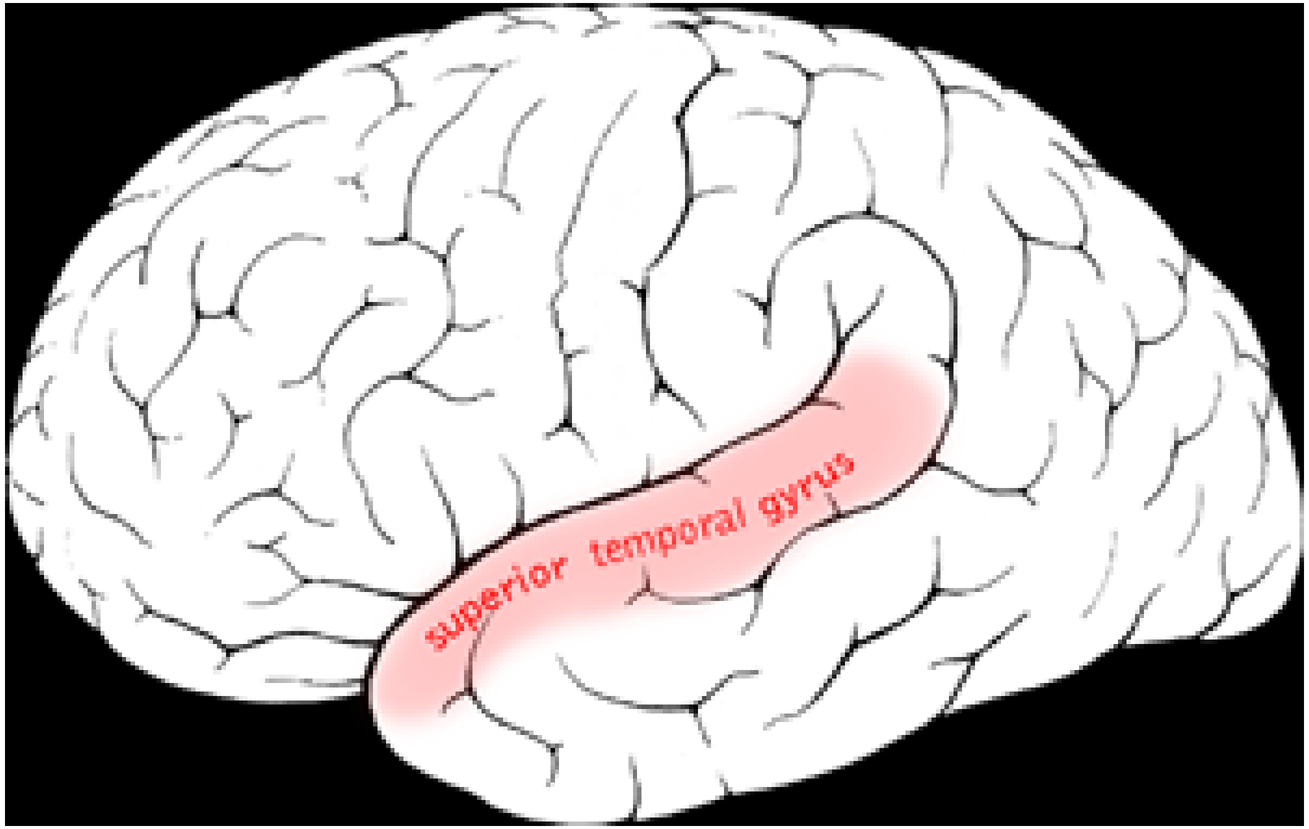
- Wav2vec2 does not show how confusable phones compete.

- Due to the CTC algorithm, the score distance between the winner and all competitors is (usually) very large, and the rank order of the competing phones may be seemingly random,

- Wav2vec2 model fine-tuned with CTC is a good phone decision machine, but a poor machine when it comes to describing phonetic structure

What could be a relation with the brain?

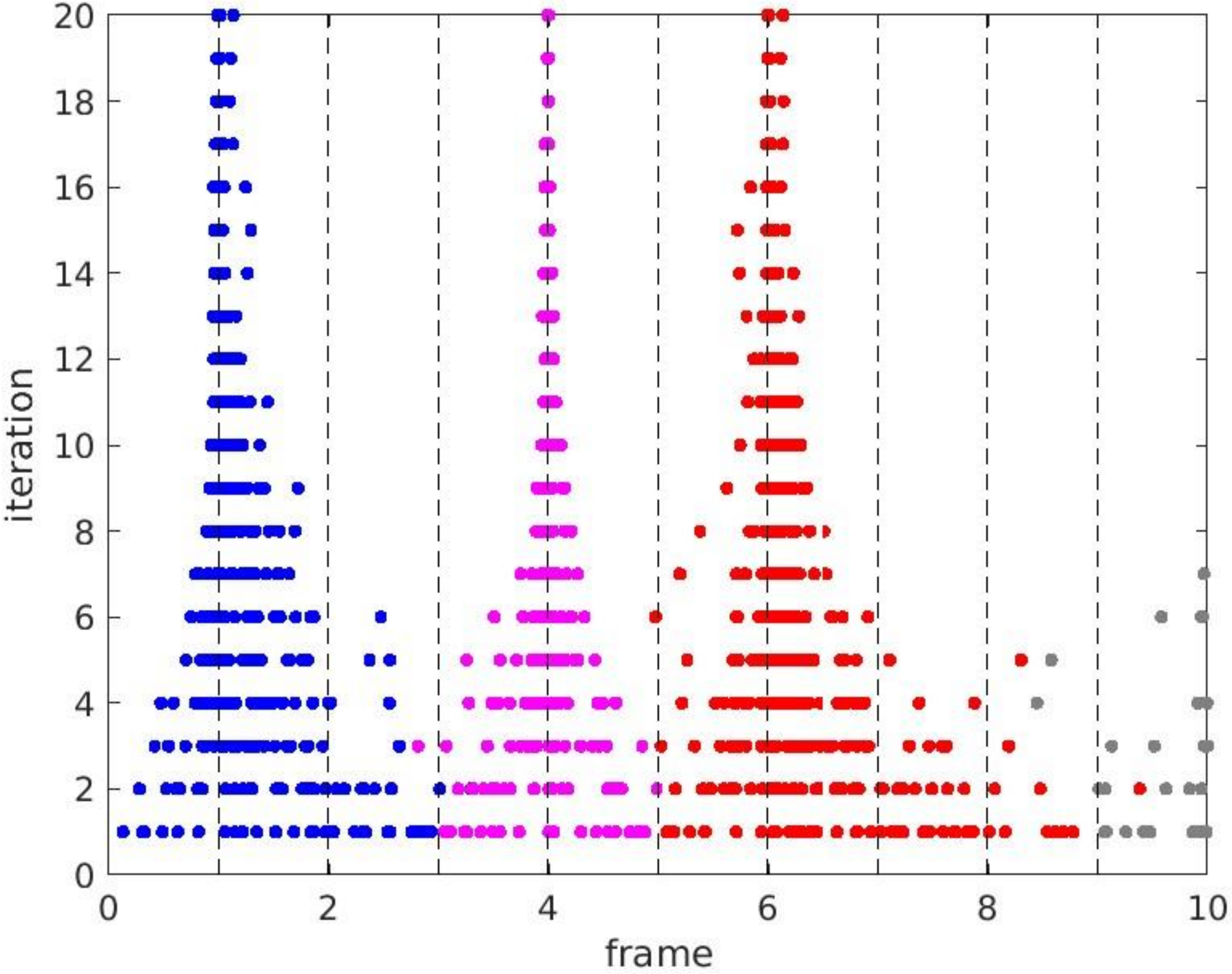
- The Superior Temporal Gyrus (STG) represents the input speech via nonlinear processes



superior temporal gyrus (STG)

- Speech processing in the human brain is based on a transformation from acoustic speech signals into internal representations via non-linear operations in the STG.
- The STG contains the *non primary* auditory cortex where phonological processing is known to take place.
    - It represents the input speech via nonlinear processes that include spectro-temporal processing, normalization and restoration based on context, and involves complex auditory encoding for acoustic-phonetic and prosodic features.

Side issue: Effect of CTC



iteration
20
18
16
14
12
10
8
6
4
2
0
0
2
4
6
8
10
frame