Analyzing self-supervised speech representations

Encoding structures of speaker information and phonetic context

Oli Danyi Liu

University of Edinburgh

Advisors: Sharon Goldwater, Hao Tang, Naomi Feldman

UKRI CENTRE FOR DOCTORAL TRAINING

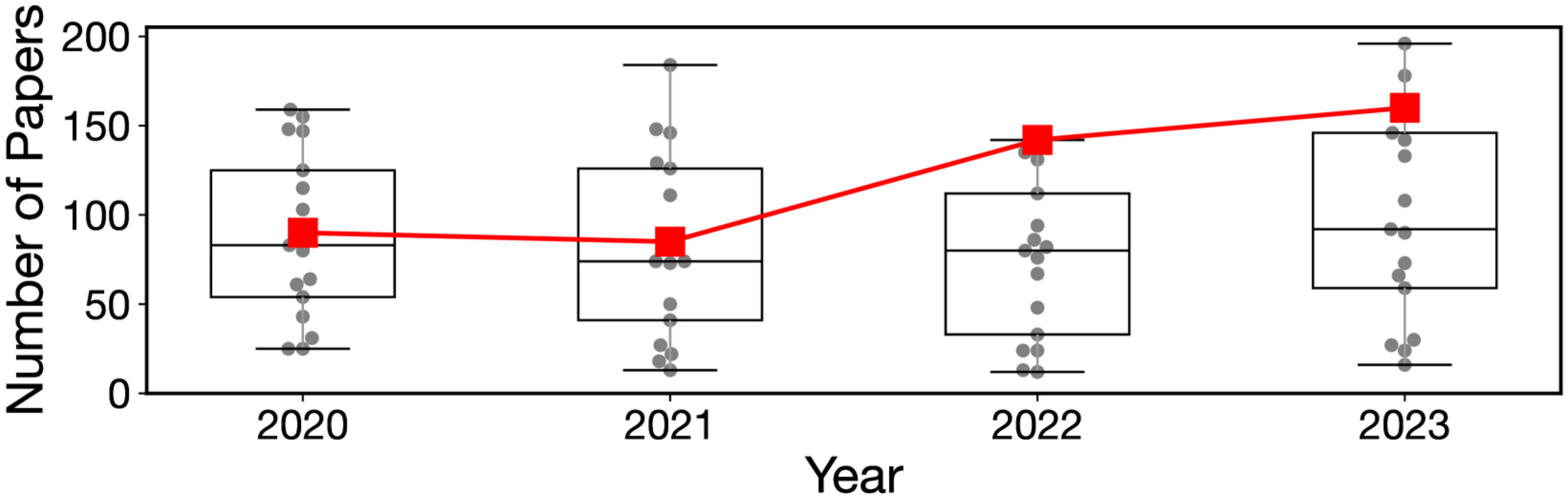
Current language technology systems are impressive

How well do state-of-the-art speech processing systems perform?

State-of-the-art speech processing systems, including automatic speech recognition (ASR) and text-to-speech (TTS) systems, perform very well in controlled environments with clear speech and standard accents. They can achieve high accuracy rates and produce natural-sounding speech.

- Self-supervised learning models play an important role
- Yet they are still largely black boxes.

Interpretability and Analysis of models



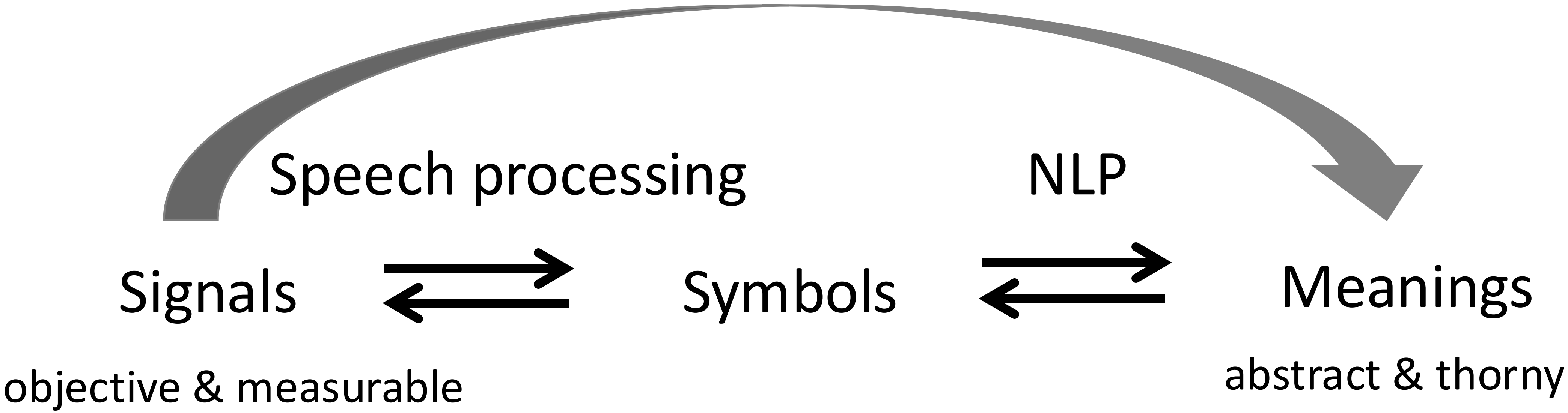
Model interpretability has been growing within NLP.

Researchers in other subfields build on findings from interpretability.

There are much fewer interpretability work on speech models.

Mosbach et al. (2024) From Insights to Actions: The Impact of Interpretability and Analysis Research on NLP

Why study speech models for interpretability



- Could potentially shed light on how discrete symbols are represented in a distributed, continuous space
- Good performance can be achieved with simpler models
- Many findings and theories from speech perception and phonology
- Language is not just about text

Using self-supervised models to explore scientific questions

Self-supervised models have been shown to

- simulate human-like perceptual biases (Millet and Dunbar, 2022)
- predict brain activities of human listeners to some extent
  (Millet et al., 2022; Caucheteux et al., 2023; Tuckute et al., 2023)

These models exhibit non-trivial properties found in humans

What computational constraints are required for these properties to arise?

Speech contains a lot of information

"eat your raisins outdoors"

quiet environment

male speaker

annoyed

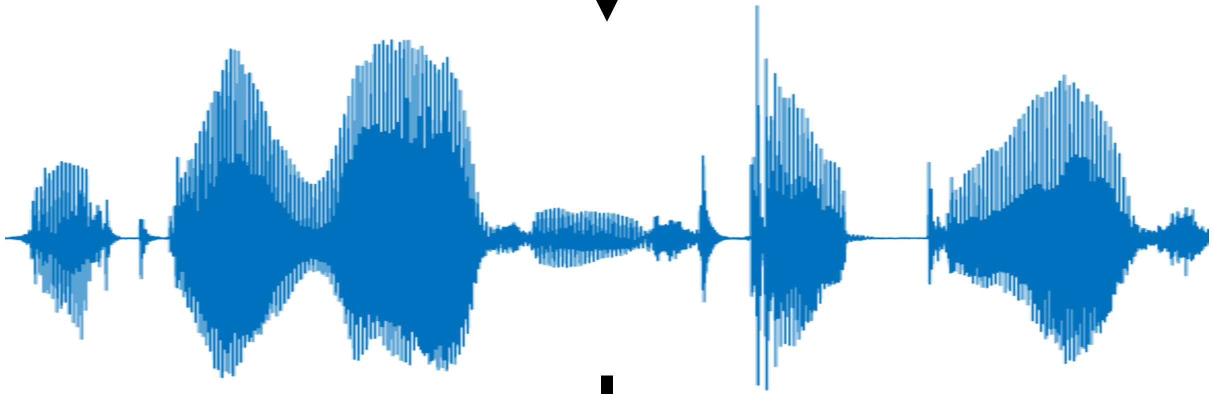
Speech contains a lot of information ⇔ variability

“eat your raisins outdoors”

speaker

environment

Non-trivial



Fourier transform

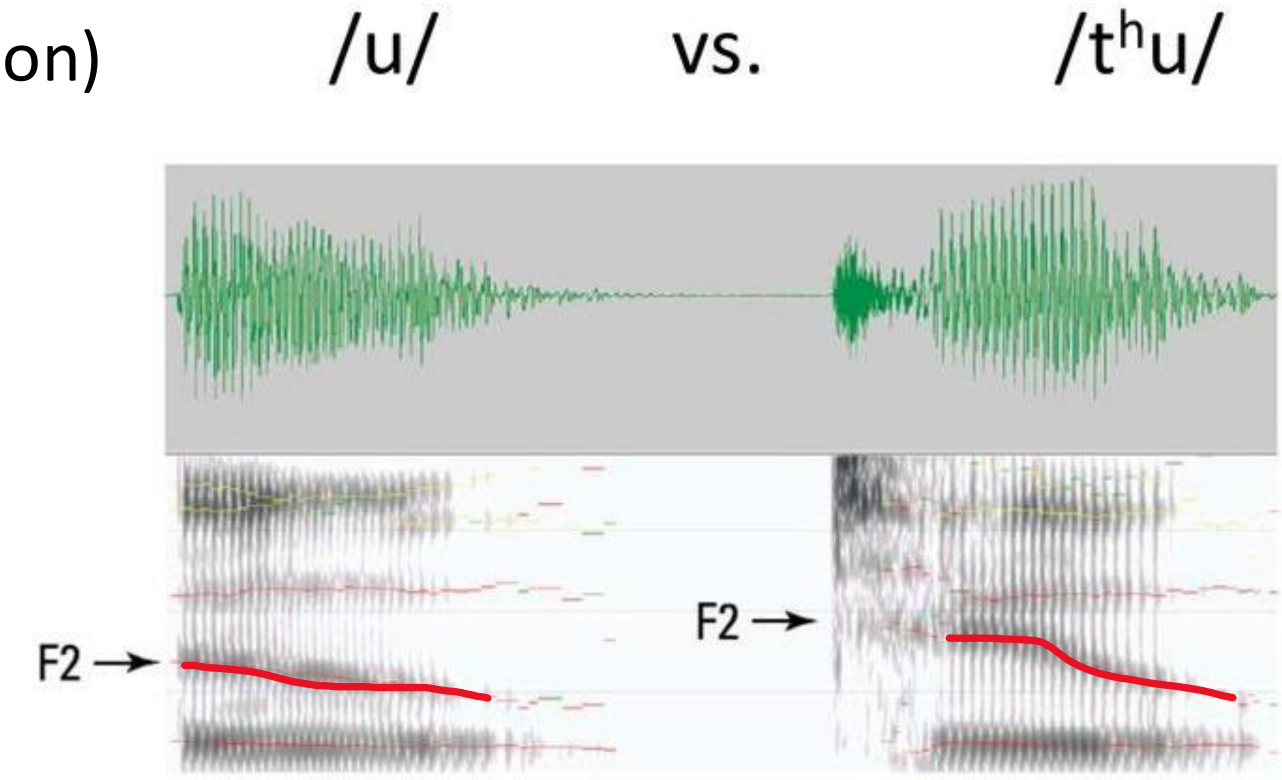
manipulation in the frequency domain

Acoustic features e.g. log Mel Spectrogram

Challenges in mapping acoustics to text

- Speaker variability
- Context sensitivity (coarticulation)
- Processing continuous speech
  - Tracking previous phones
  - Tracking their order

For example,
cats, task, tax, asked, acts
all consist of /k/, /æ/ , /t/, /s/



Example from William Katz's lecture notes

Outline

In the representation space of self-supervised learning models:

1. Speaker information is encoded orthogonally to phonetic information

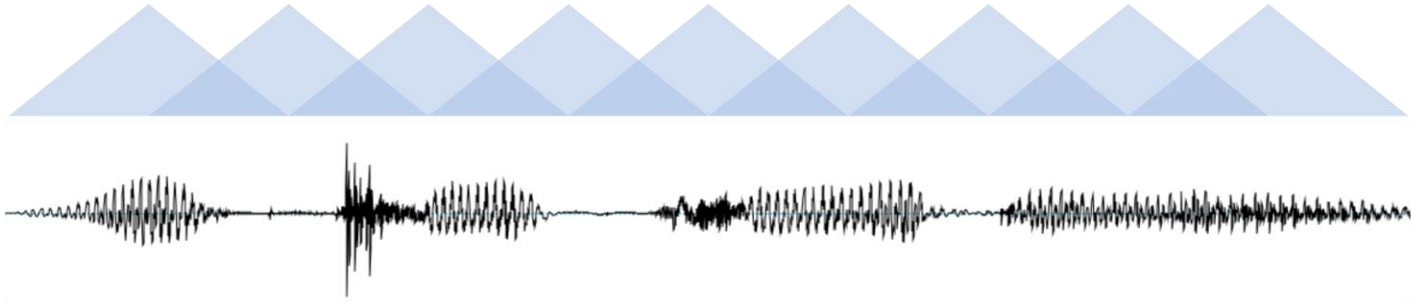
2. Multiple successive phones are encoded at the same time

3. There is some extent of cross-context generalizability

*2, 3 were also found in the neural encoding of human listeners

Self-supervised learning (SSL) model of speech

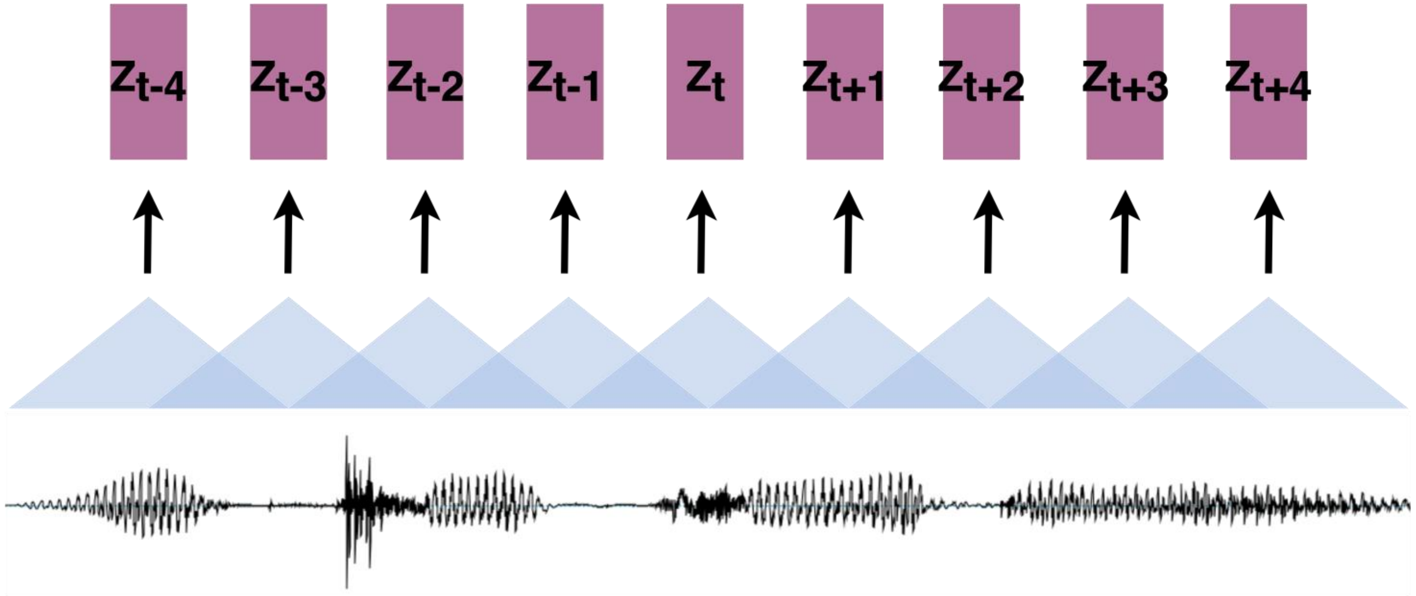
1-D convolution



Self-supervised learning (SSL) model of speech

Frame-level
Embedding
(1 frame = 10ms)

1-D convolution

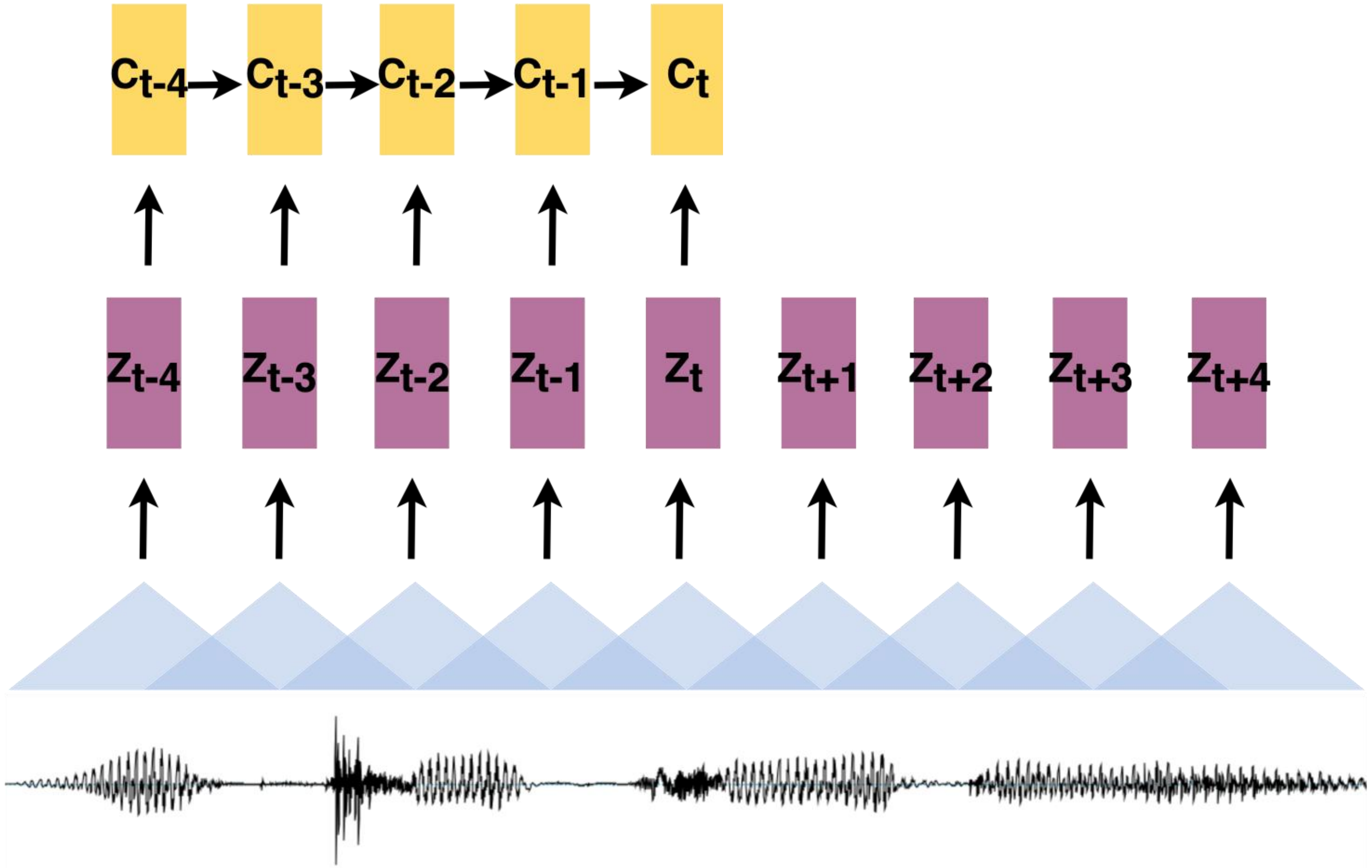


Self-supervised learning (SSL) model of speech

Contextualized embeddings (4-layer LSTM)

Frame-level embedding

1-D convolution

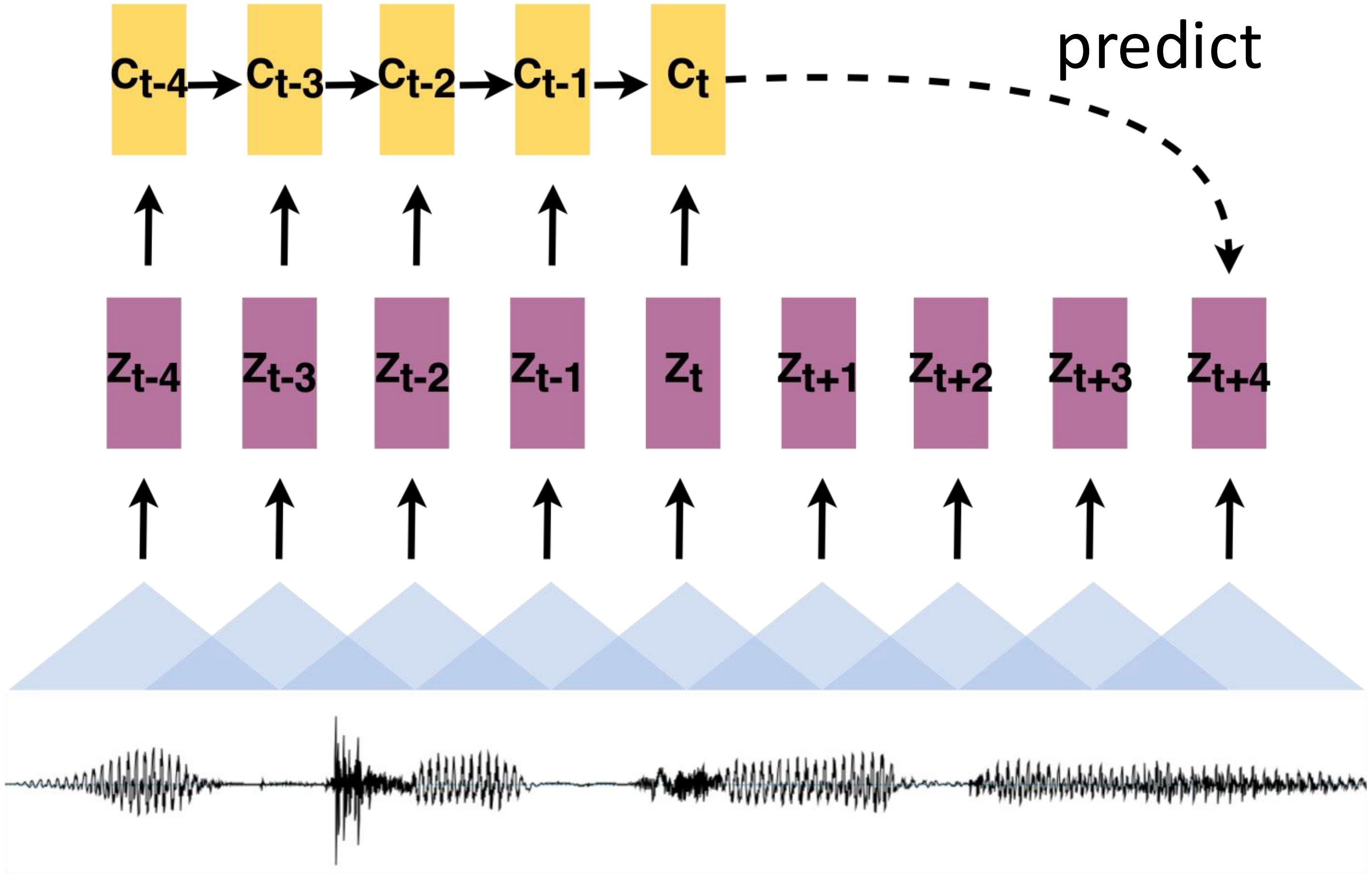


Self-supervised learning (SSL) model of speech

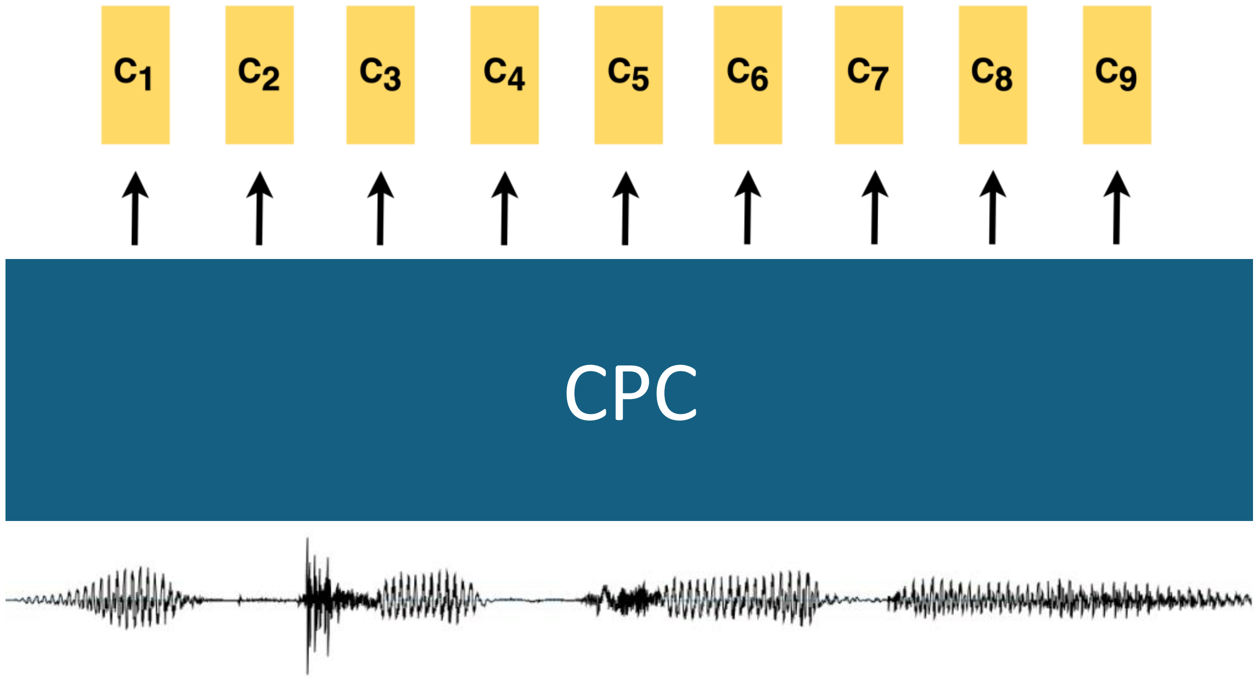
Contextualized embeddings (4-layer LSTM)

Frame-level embedding

1-D convolution



Contrastive predictive coding



- Forward prediction
- more cognitively plausible than masked prediction
- LSTM-based
- results from transformer-based models are consistent

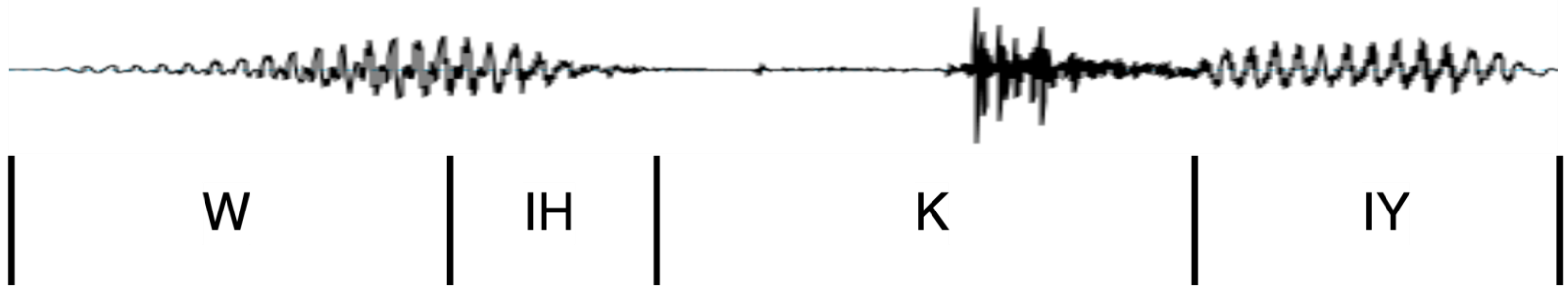
Oord et al. (2018) Representation learning with contrastive predictive coding

Outline

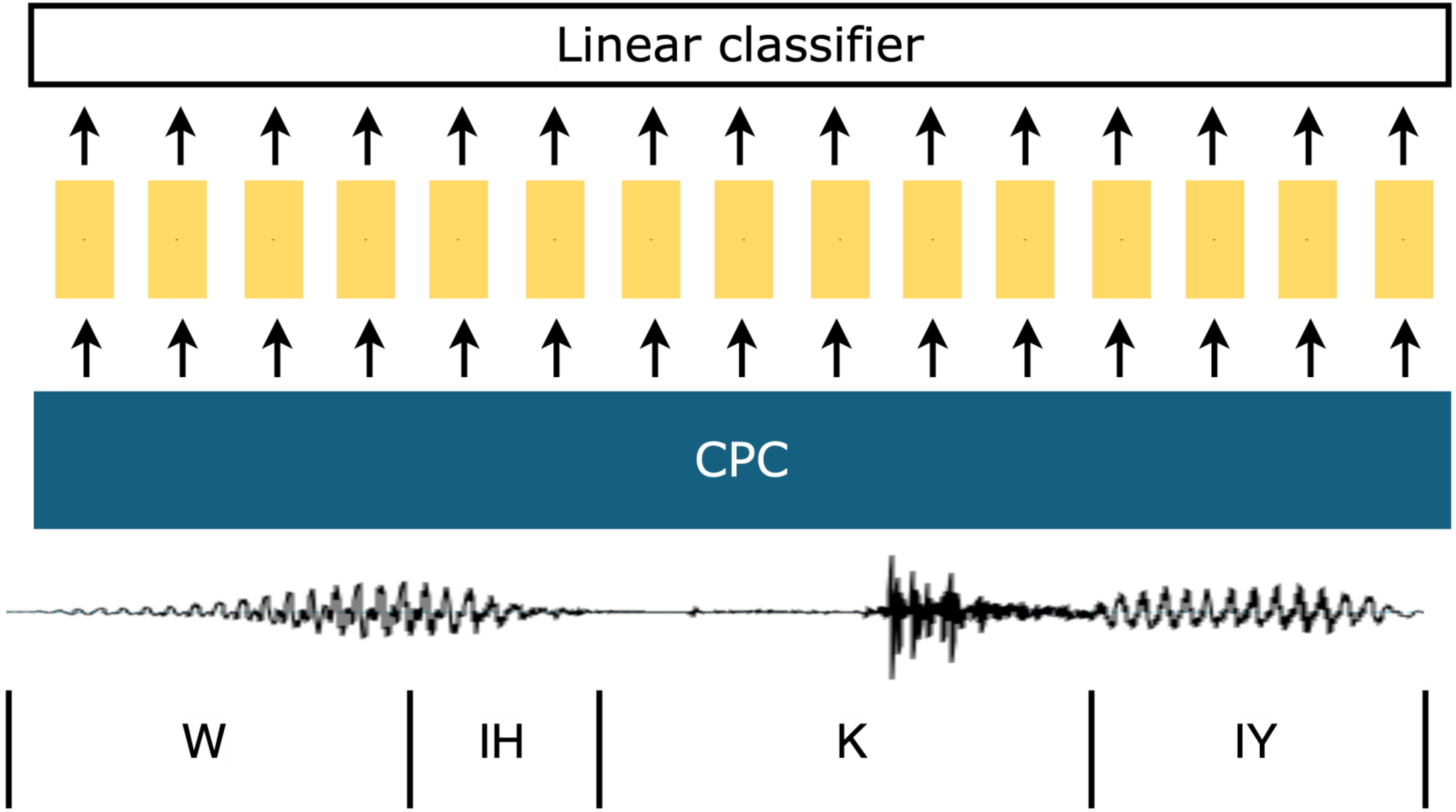
In the representation space of self-supervised learning models:

1. Speaker information is encoded orthogonally to phonetic information

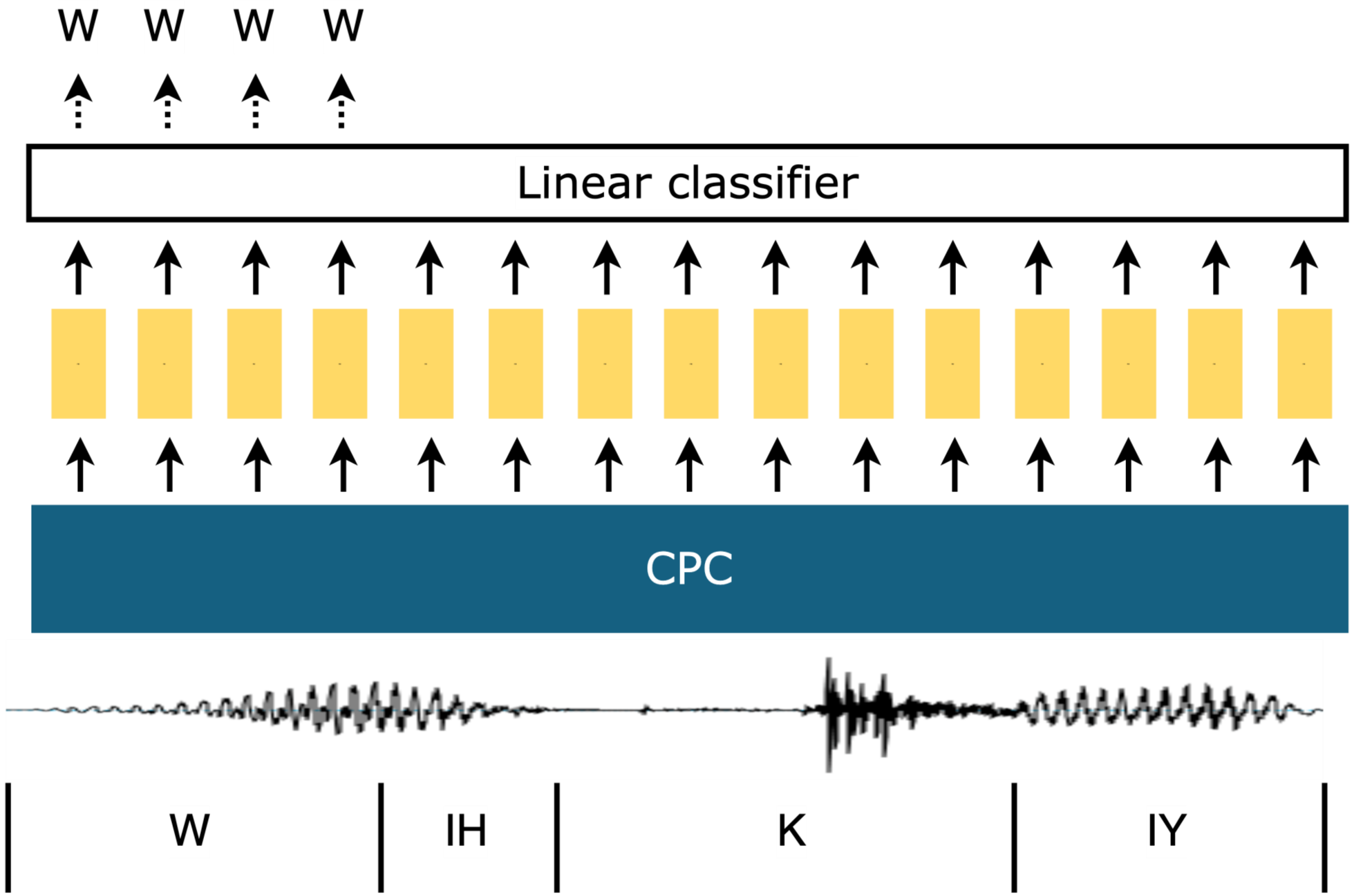
Probing for phonetic information



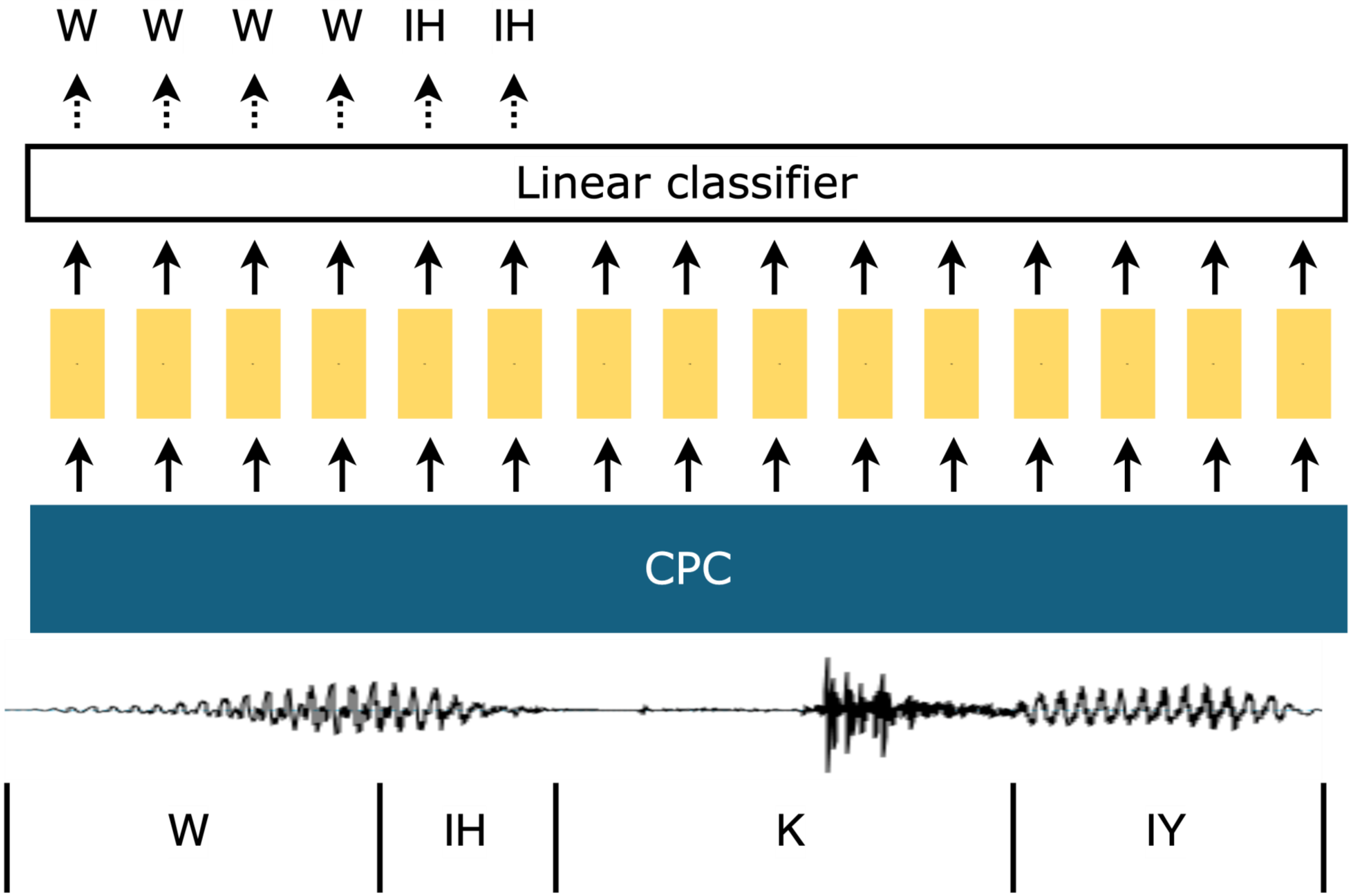
Probing for phonetic information



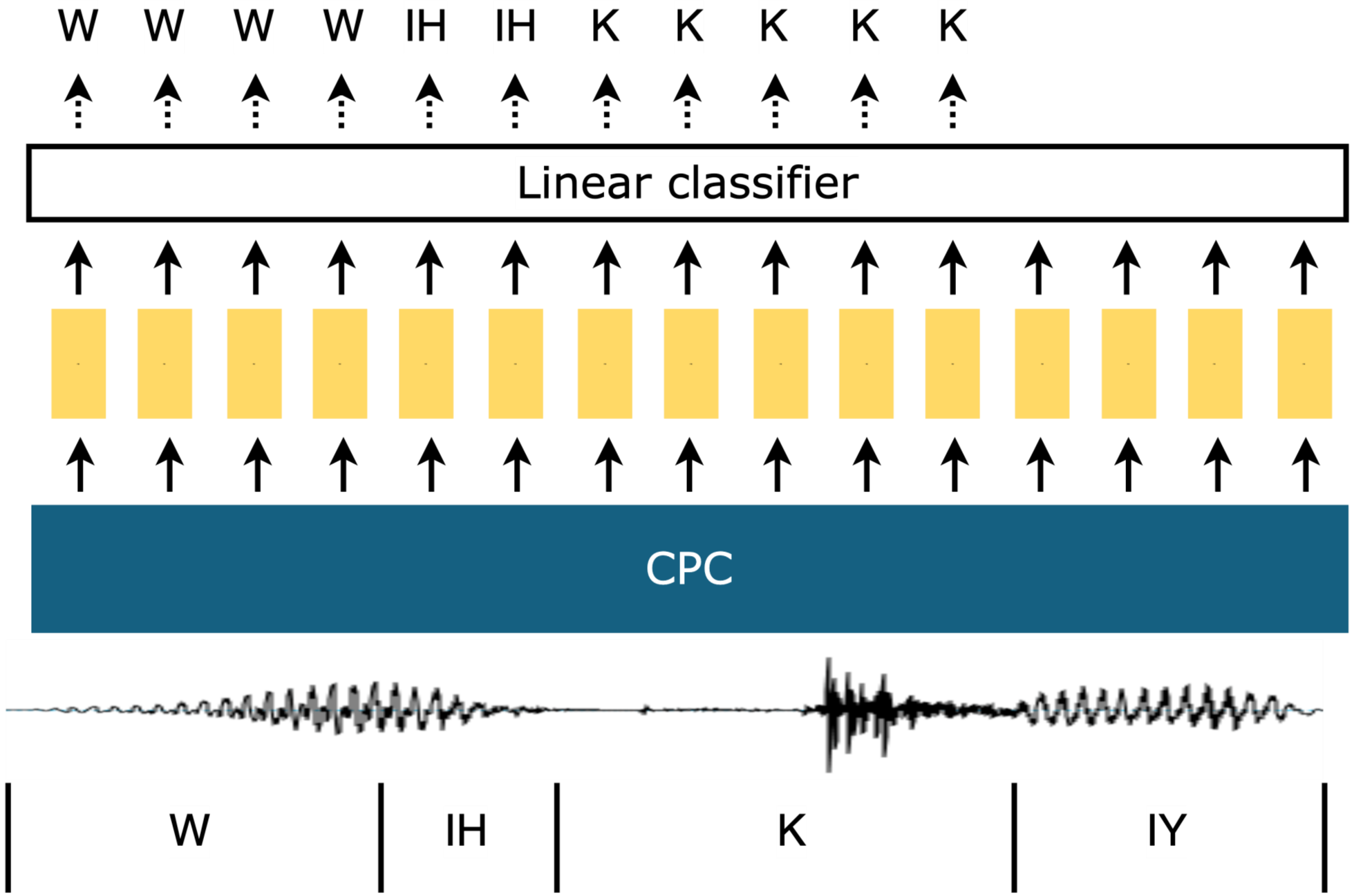
Probing for phonetic information



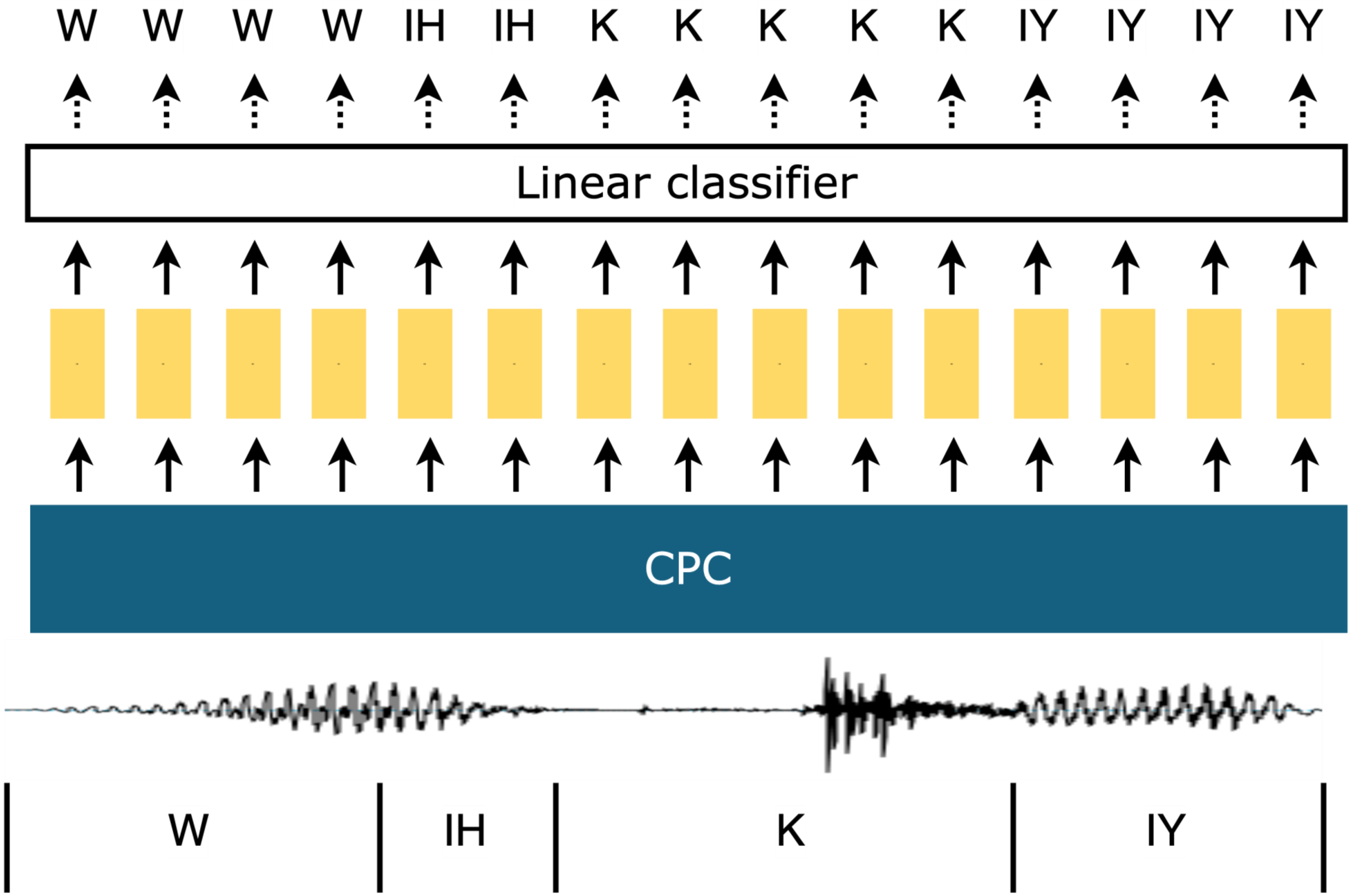
Probing for phonetic information



Probing for phonetic information

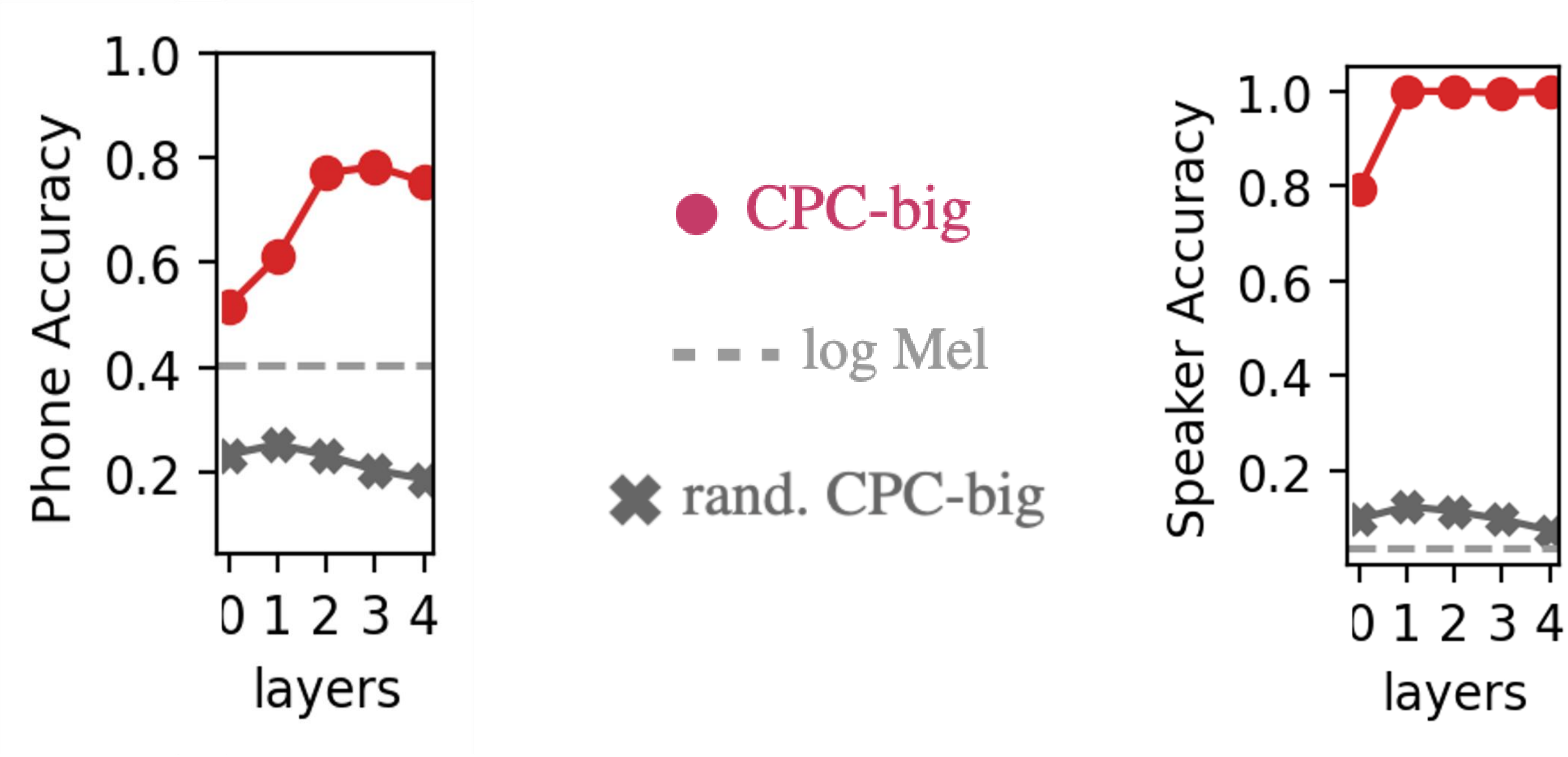


Probing for phonetic information



CPC encodes significant

phonetic information and speaker information



Previous work on analyzing SSL speech models

Representations in these models encode

- acoustic events (Wells et al., 2022)
- word-level context (Sanabria et al., 2022)
- speaker identity (van Niekerk et al, 2021)
- gender (de Seyssel et al., 2022)

What information is encoded

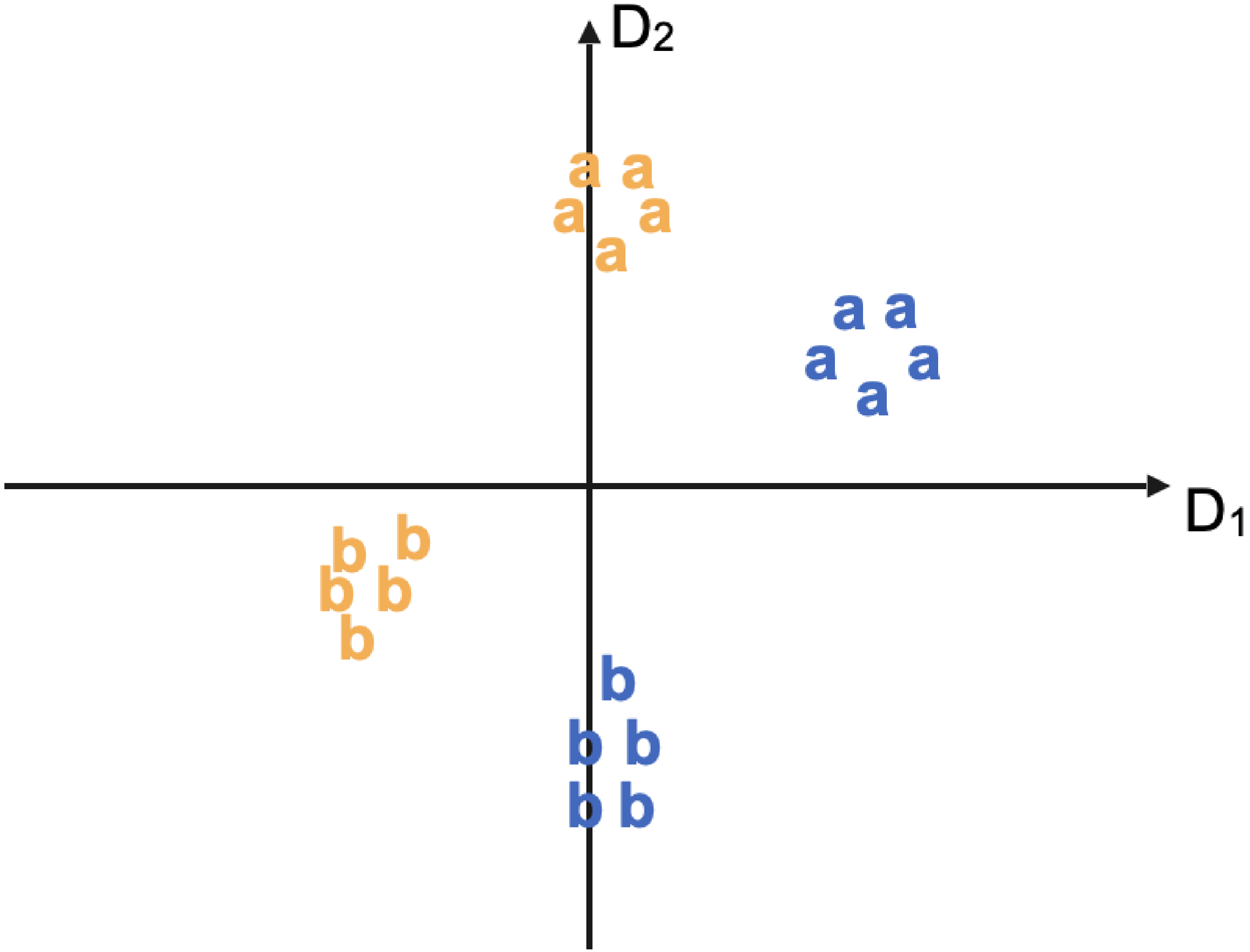
Which layers are different information more salient (Pasad et al., 2021; Pasad et al. 2023)

How are they organized in the representation space?

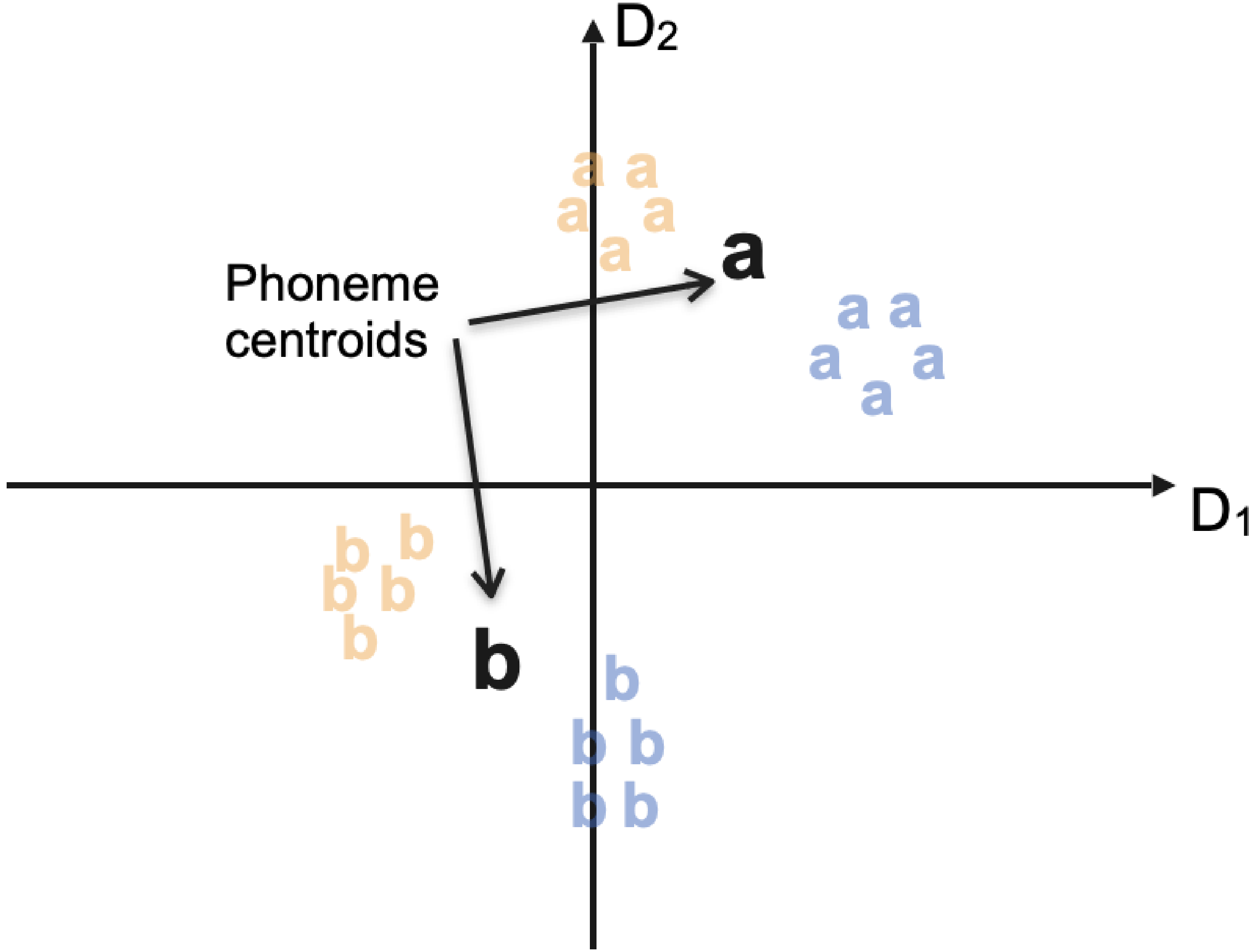
Our hypothesis

- Humans maintain acoustic details and can perceive speaker differences but can also easily abstract away speaker variability to recognize words.
- Speaker and linguistic information vary independently in producing speech.
- They could be encoded ***orthogonally***

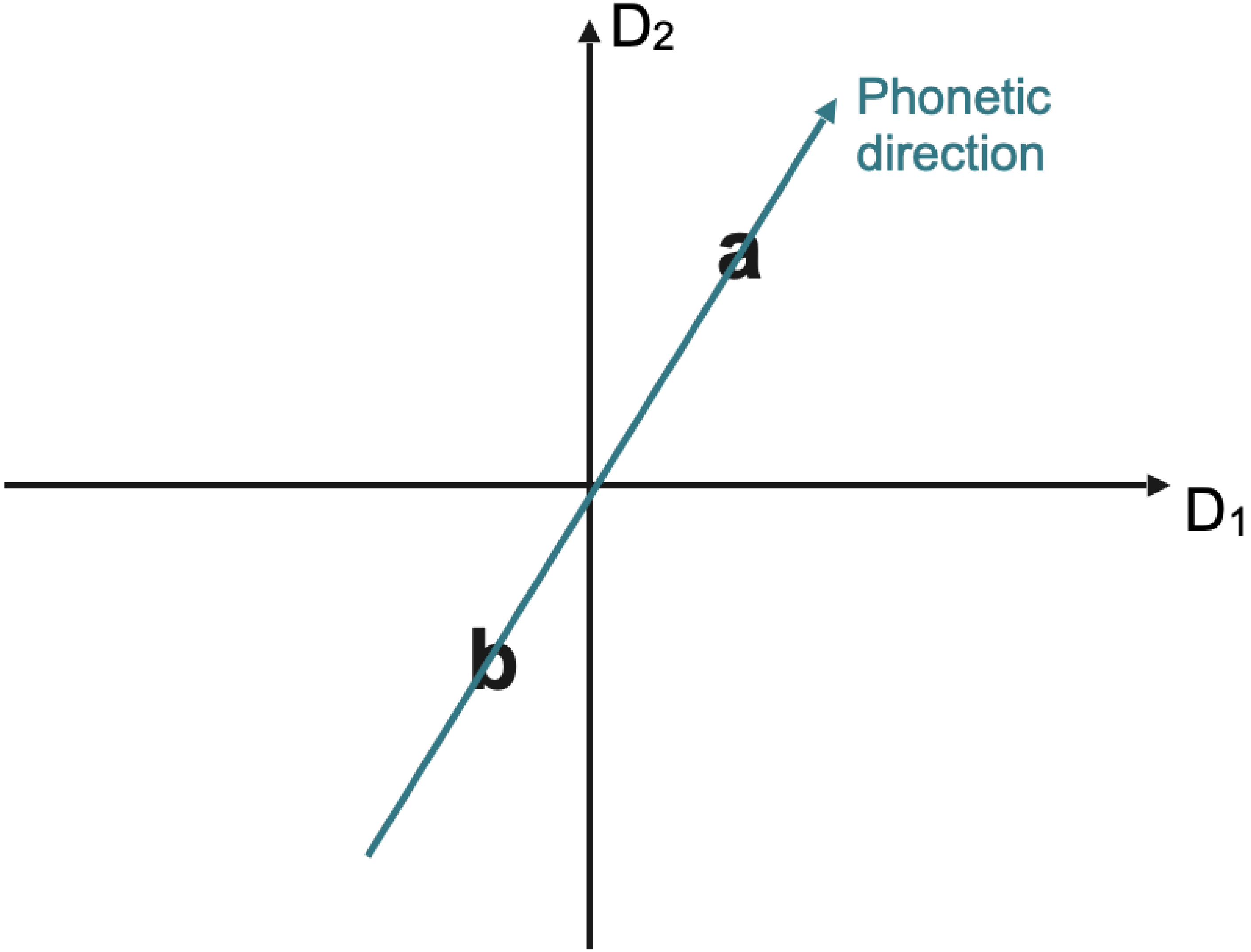
Evaluate orthogonality



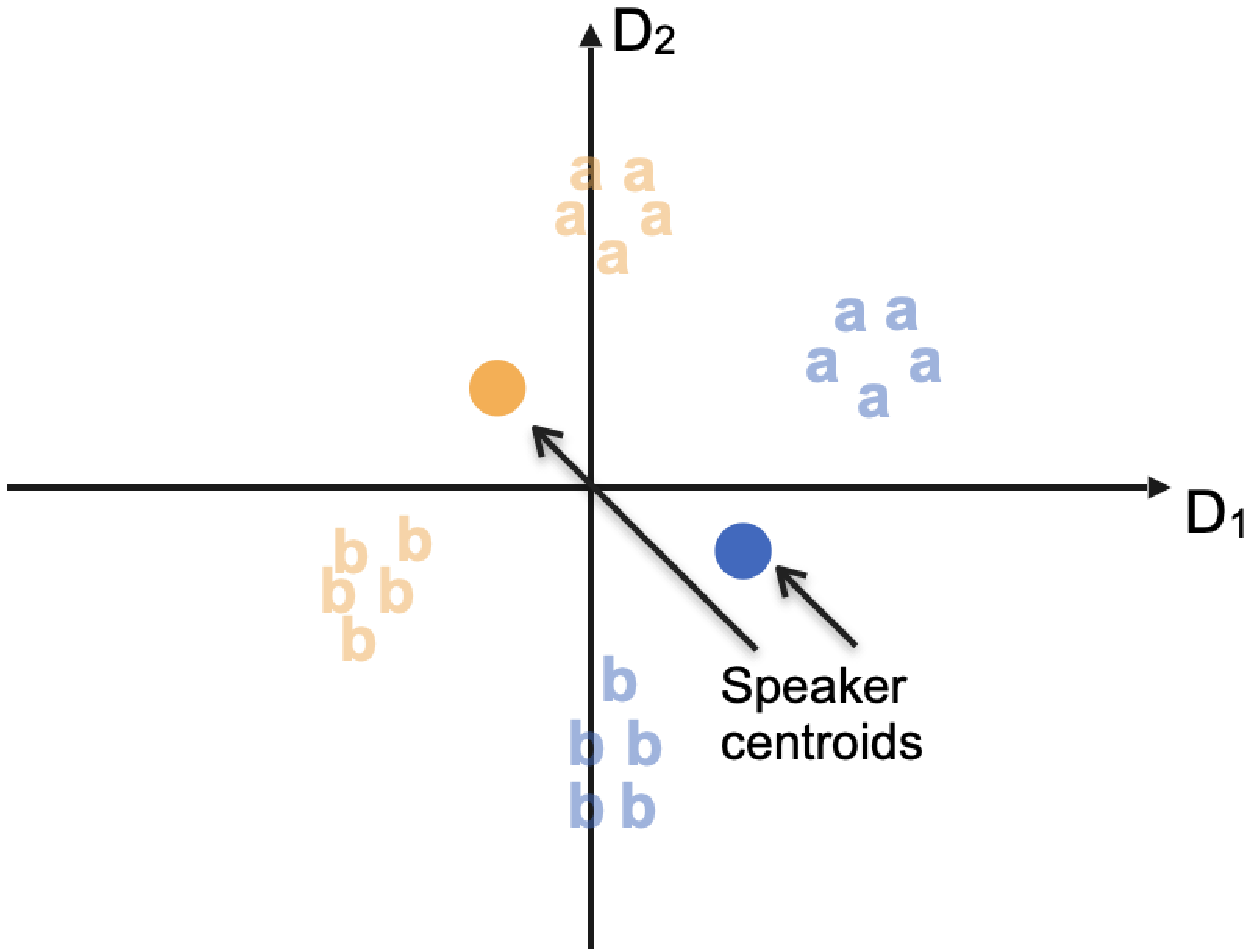
Evaluate orthogonality



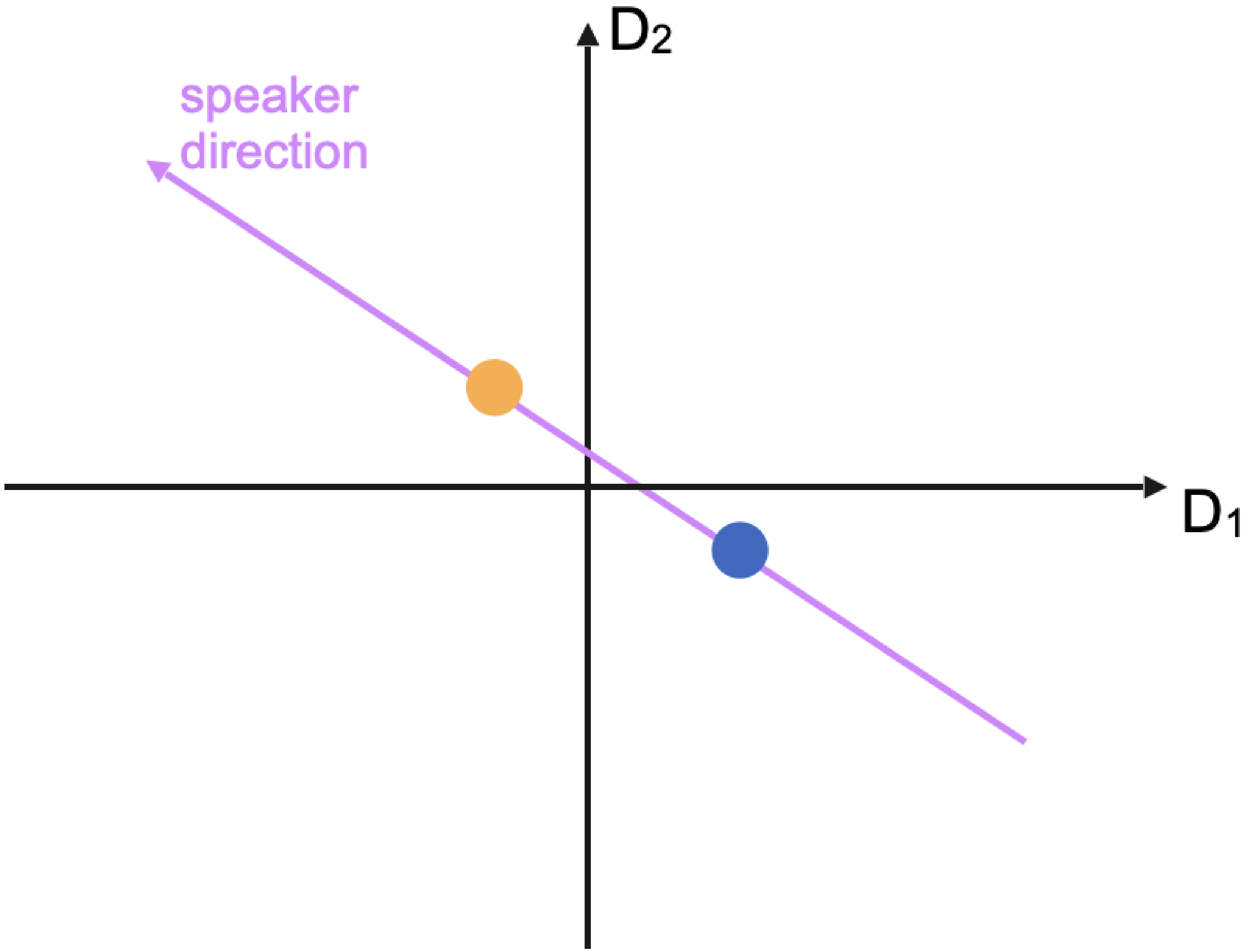
Evaluate orthogonality



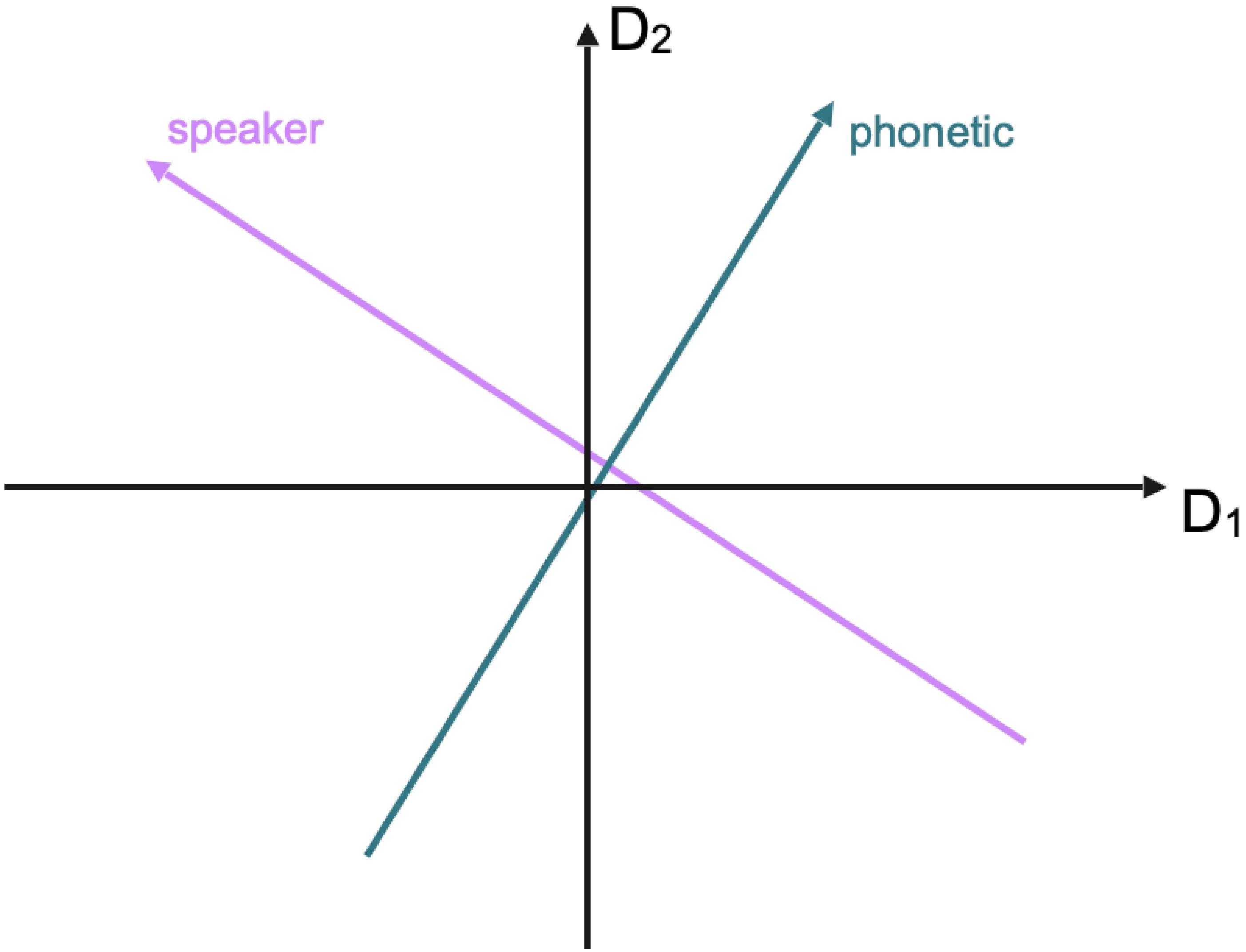
Evaluate orthogonality



Evaluate orthogonality



Evaluate orthogonality



Evaluating orthogonality

1. Identify the speaker subspace and the phonetic subspace
   Dataset: Librispeech (English audiobooks read by US native speakers)
   We used the dev-clean subset with 40 speakers (8 min per speaker)

phoneme label

↑

Each representation

↓

speaker label

aggregate ⇒

39 phoneme centroids

40 speaker centroids

PCA ⇒

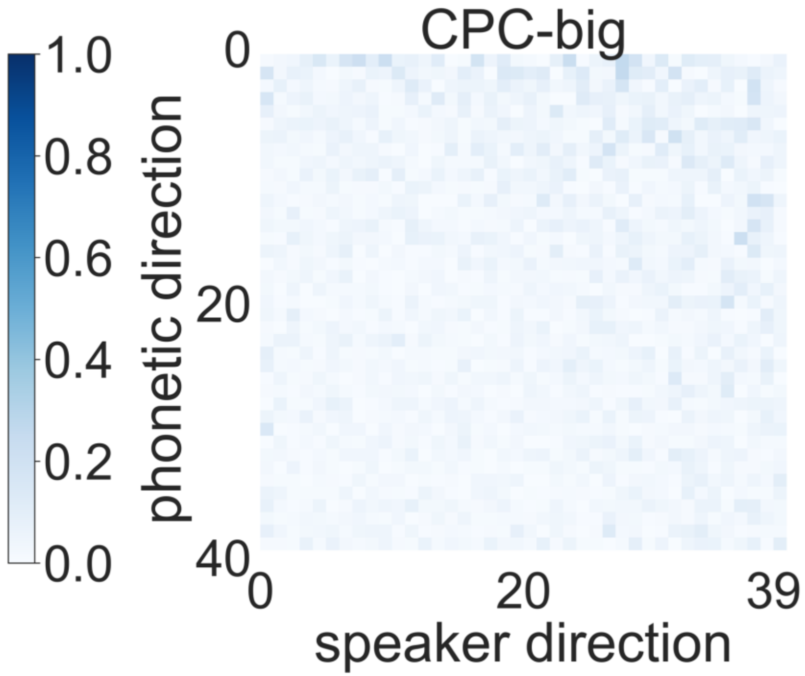
39 phonetic directions

40 speaker directions

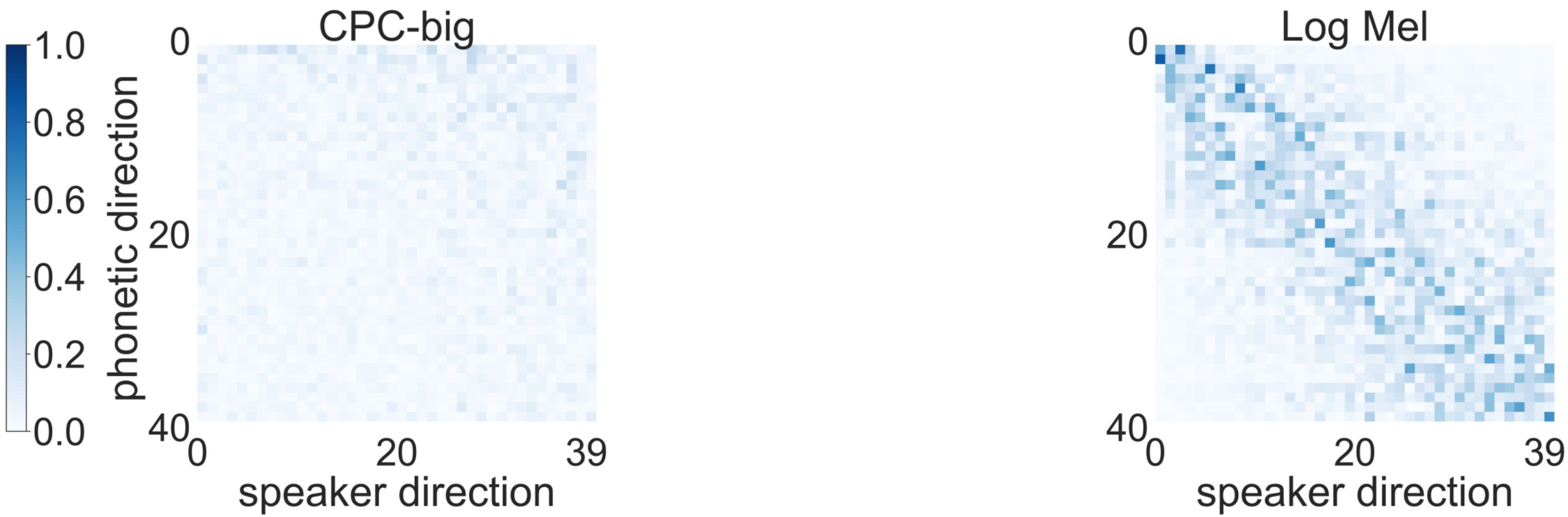
Evaluating orthogonality

1. Identify the speaker subspace and the phonetic subspace
2. Evaluate whether the two subspaces are orthogonal
   - Measure cosine similarity between speaker and phonetic directions.
     If orthogonal, they should be low.
   - "Collapse" the speaker subspace, i.e. project to its null space;
     measure phonetic information in the projected vector.
     If orthogonal, phonetic information should be intact.

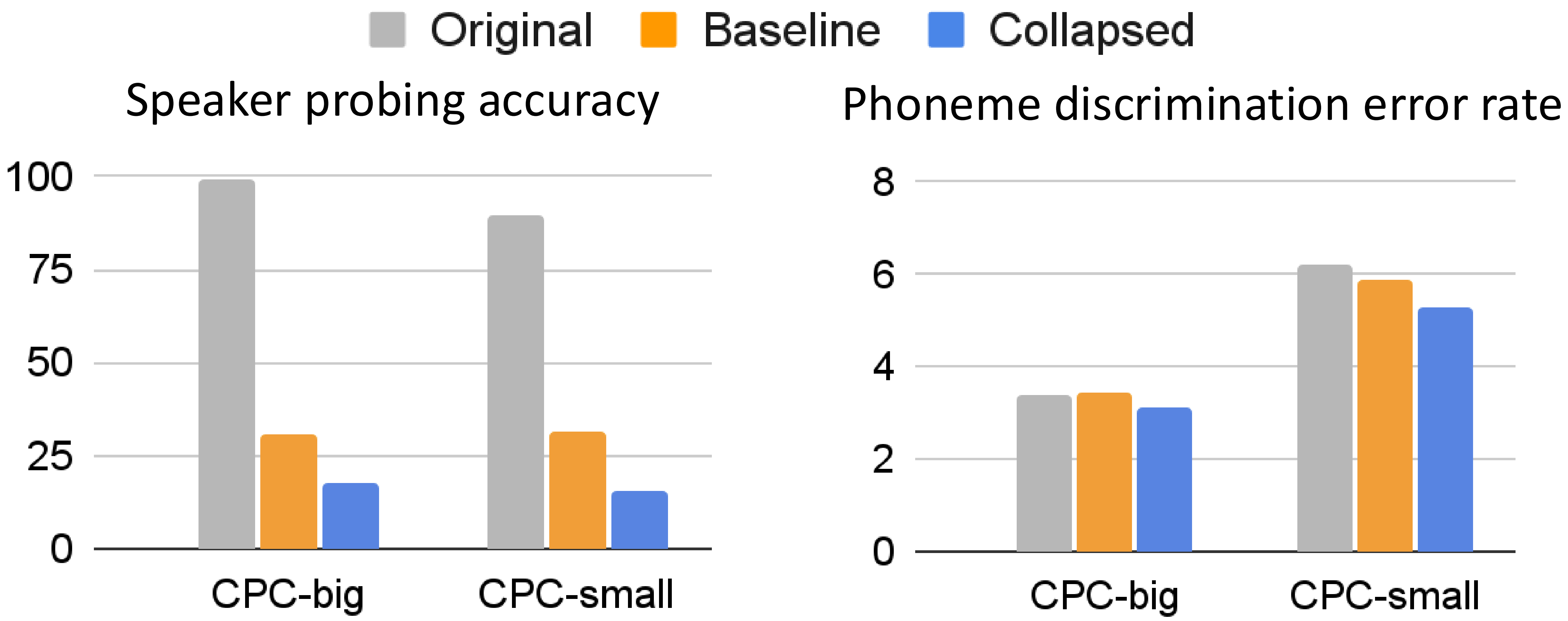
Cosine similarity between speaker and phonetic directions



Cosine similarity between speaker and phonetic directions



“Collapsing” the speaker subspace



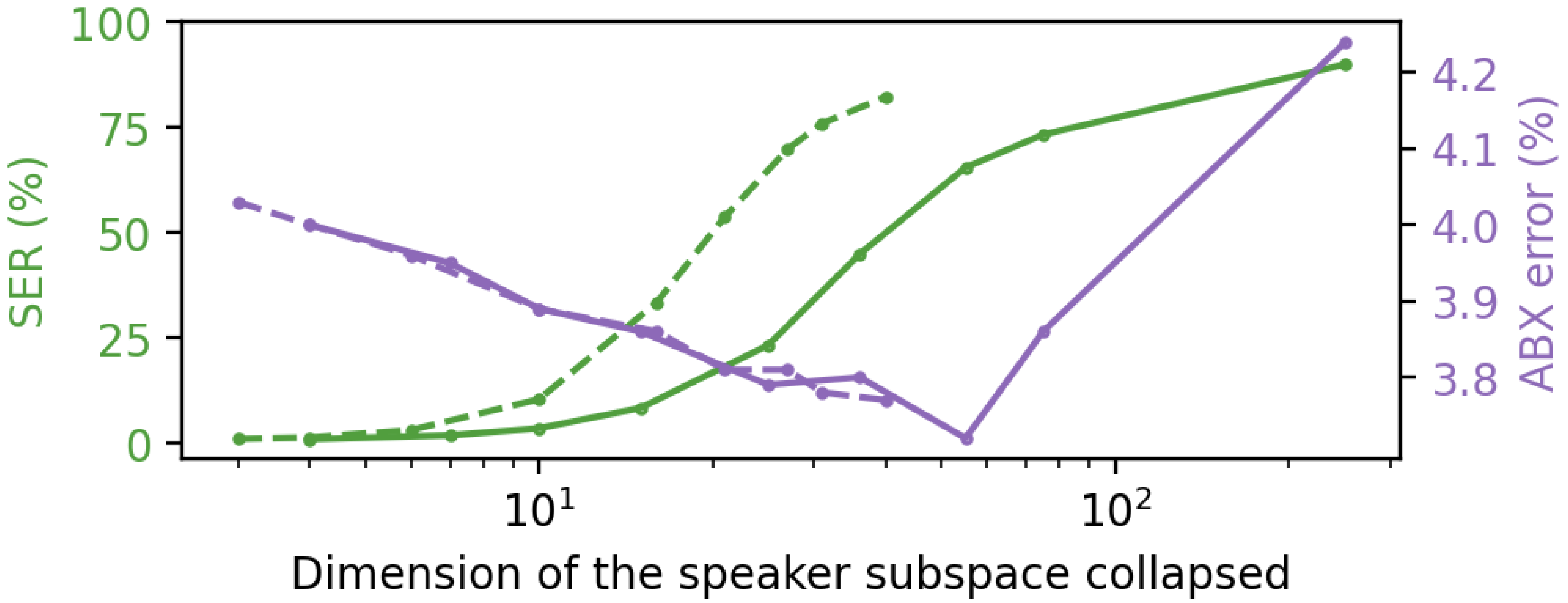
Remove speaker information

Improve phoneme discrimination

The learnt speaker subspace generalizes to unseen speakers

Collapsing a learnt speaker subspace on unseen speakers can

- Eliminate speaker information
- Improve phoneme discriminability



Conclusions (part 1)

Speaker and phonetic information are encoded in orthogonal subspaces

- This property lends itself to simple disentanglement
  - Could be used for speaker normalization
  - Are they orthogonal in neural encoding of brains?
- In a follow-up work, we proposed a quantitative measure for orthogonality and found that it correlates with phoneme probing accuracy

Mohamed et al. (2024) Orthogonality and isotropy of speaker and phonetic information in self-supervised speech representations (To appear in Interspeech)

Outline

In the representation space of self-supervised learning models:

1. Speaker information is encoded orthogonally to phonetic information

2. Multiple successive phones are encoded at the same time

Temporal dynamics of phone encoding

Phones need to be tracked and integrated to extract words.

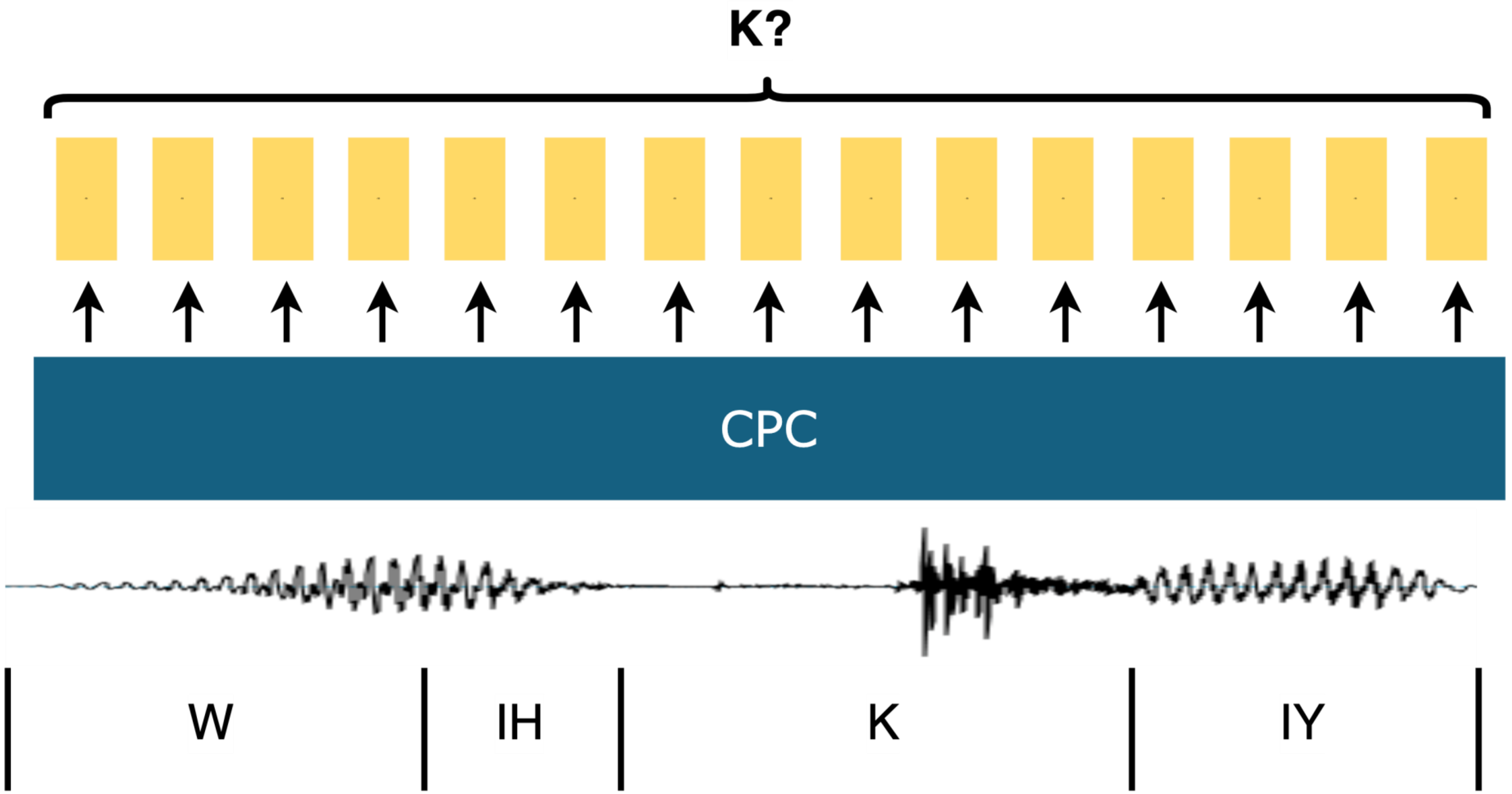
The average duration of a phone is about 80ms.

Gwilliams et al. (2022) analyzed MEG recordings from human listeners, and found that each phone is decodable for 400ms.

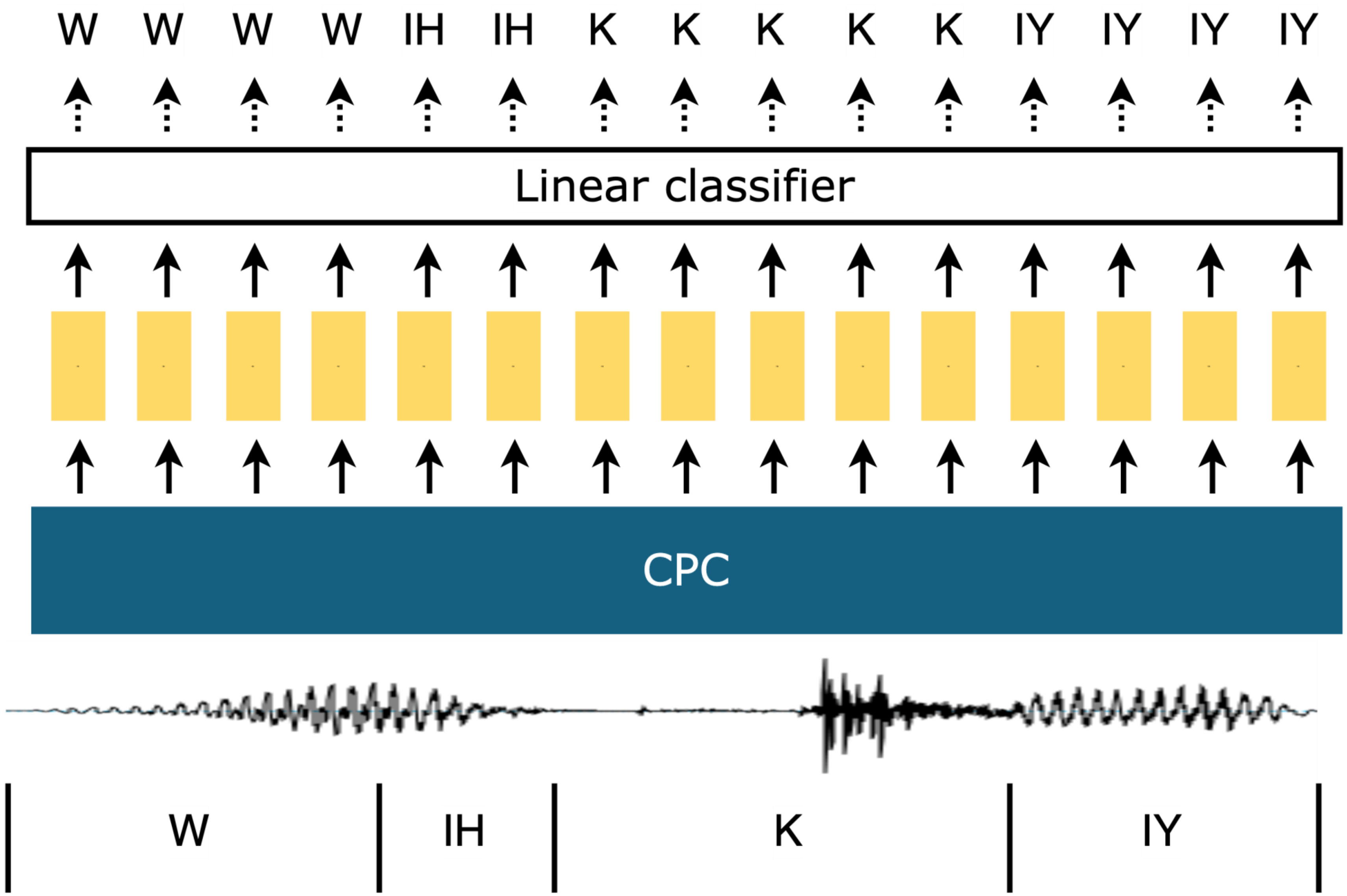
- Coarticulation could cause a phone to be encoded for $>$ 80ms
- A decodable window $\gg$ 80ms implies multiple phones are maintained simultaneously

Gwilliams et al. (2022) "Neural dynamics of phoneme sequences reveal position-invariant code for content and order." *Nature communications*

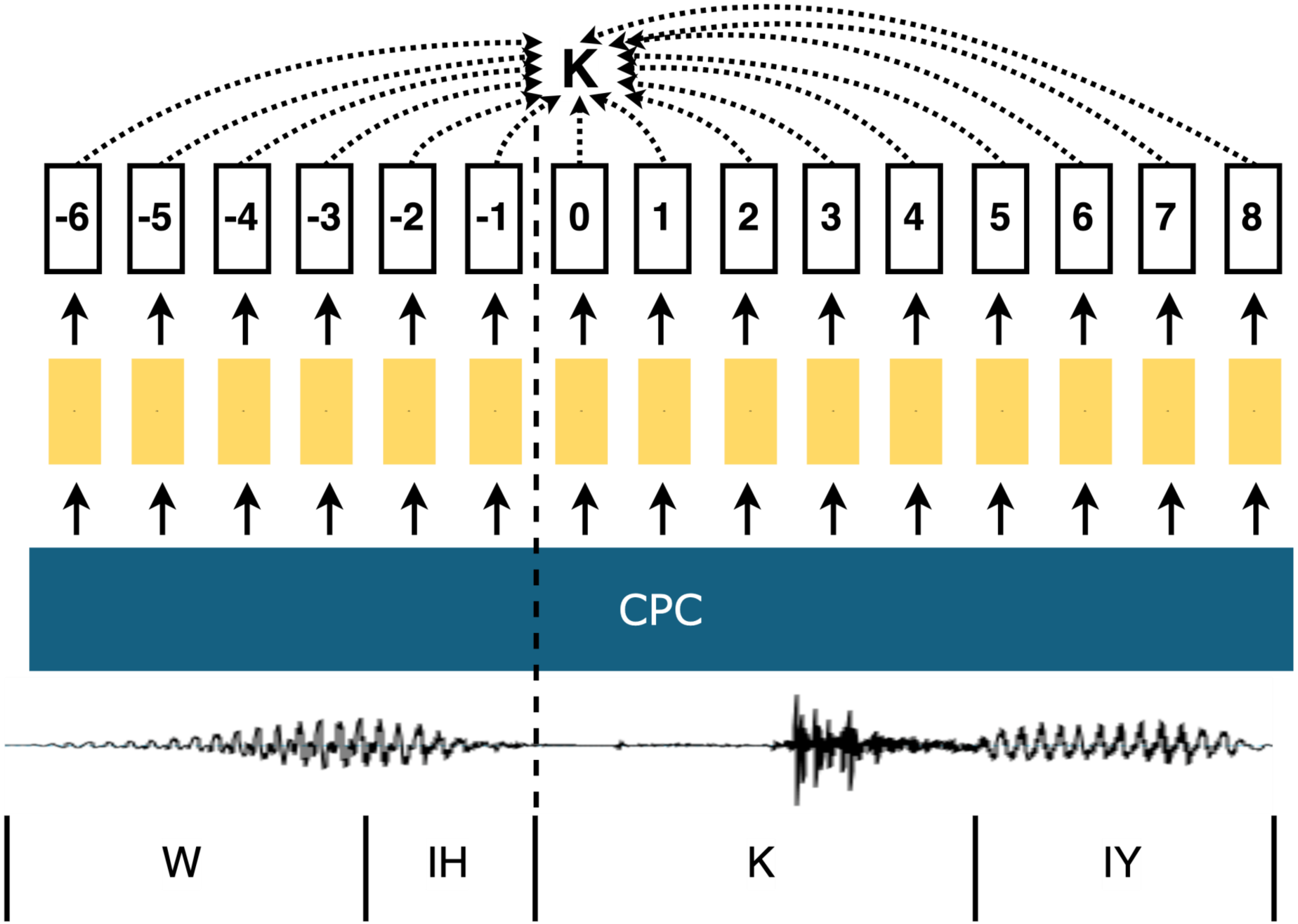
How long can we decode the phoneme with representations before and after it occurs in the acoustics?



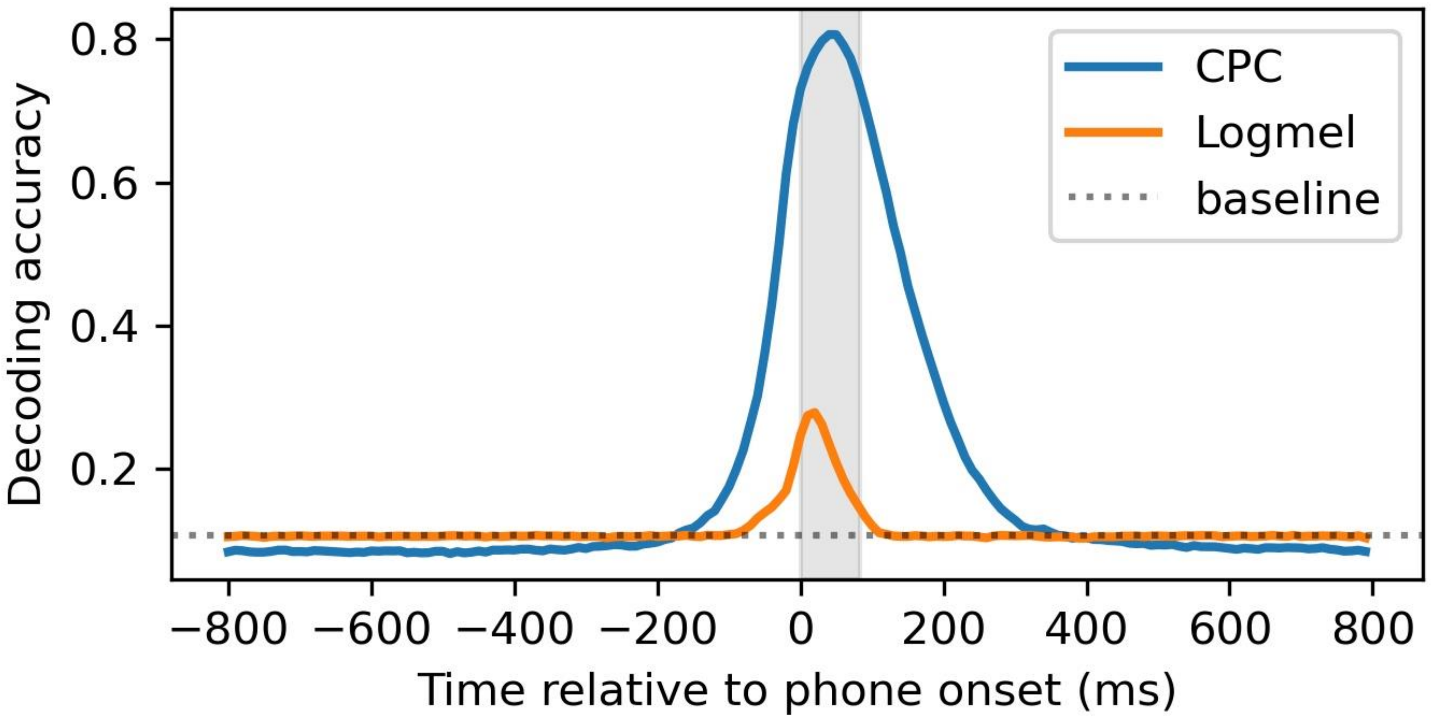
Recall standard probing



Decoding a phone from neighboring frames

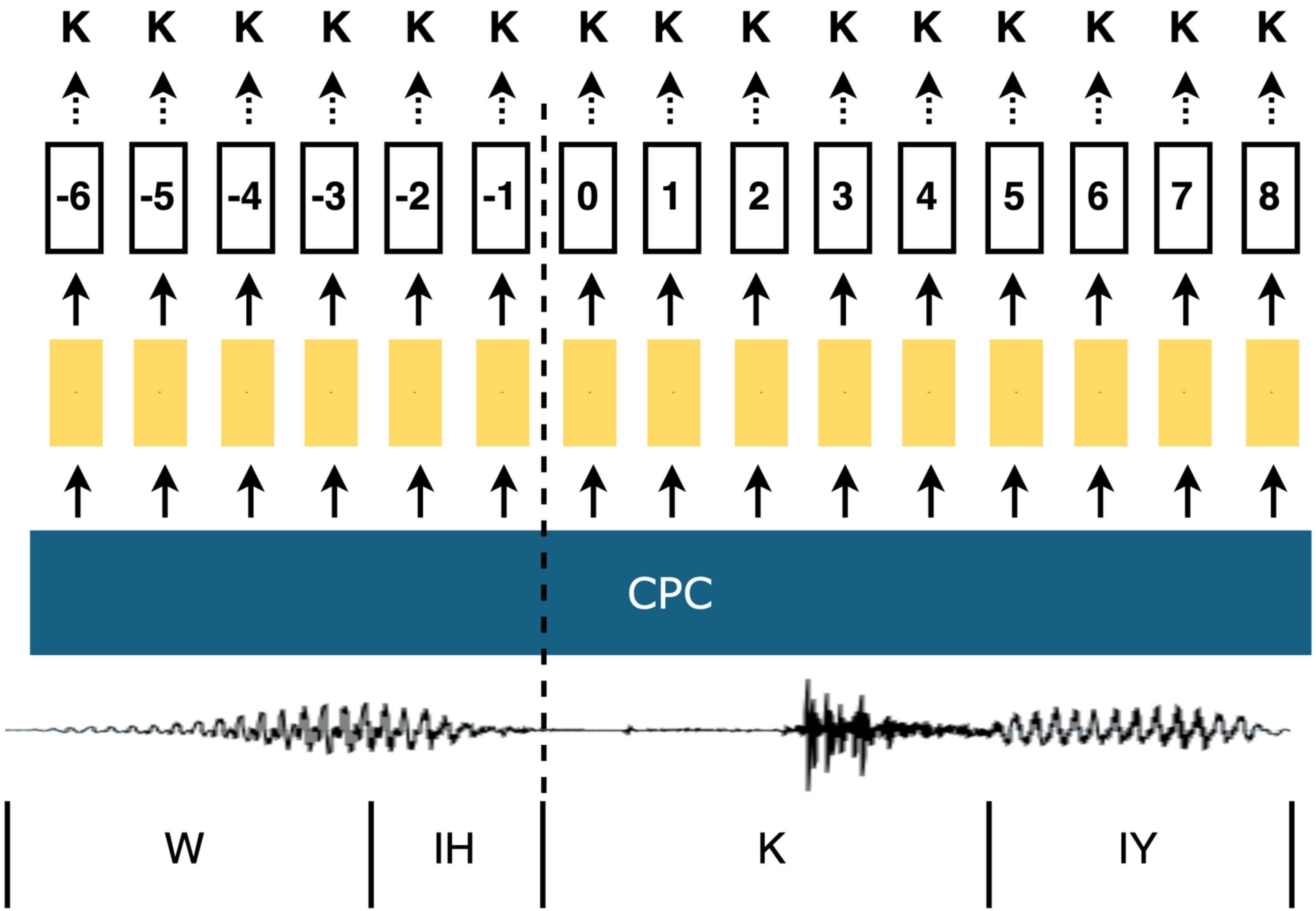


The window of phonetic decodability

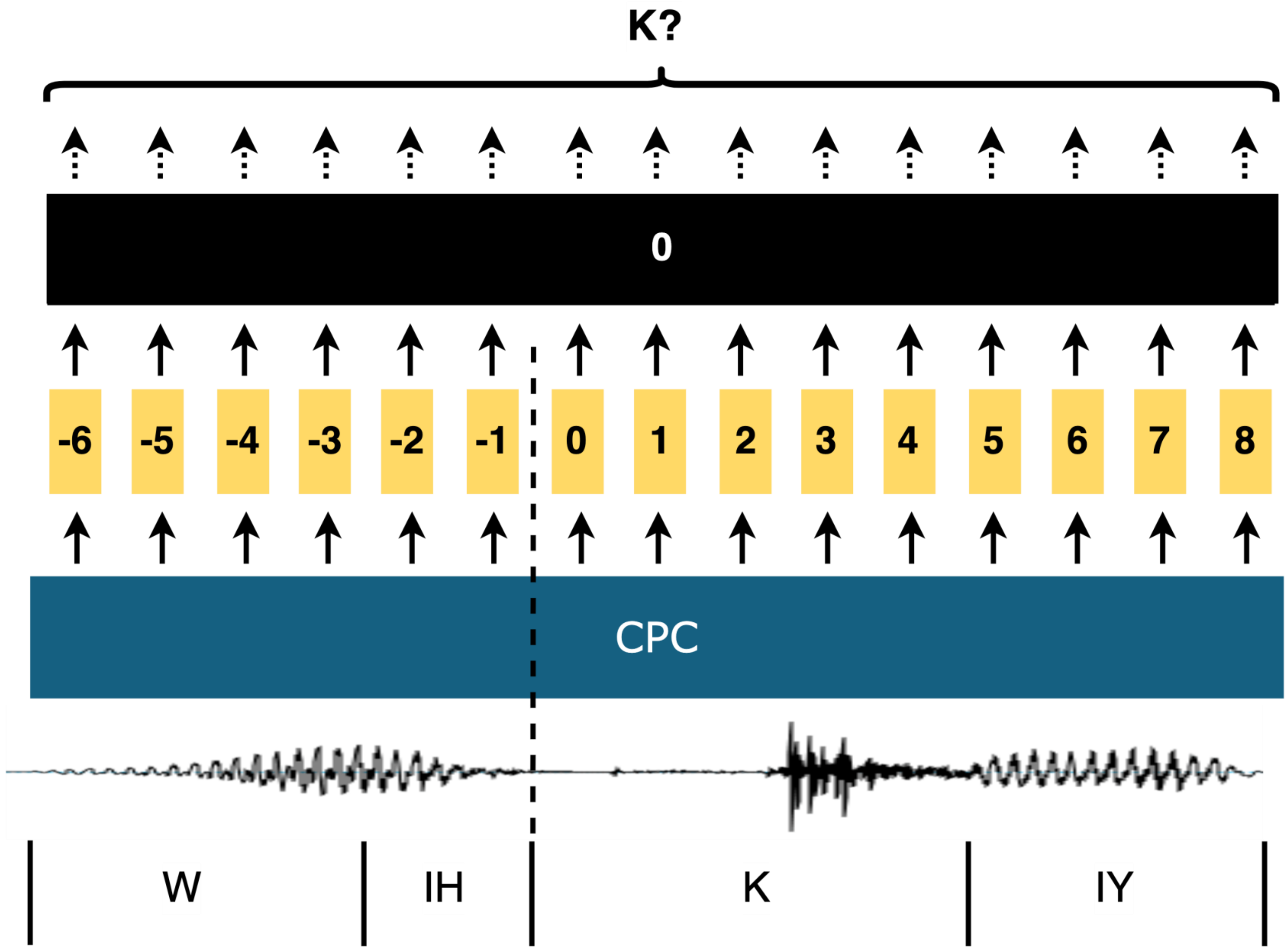


Brain recordings – about 400ms

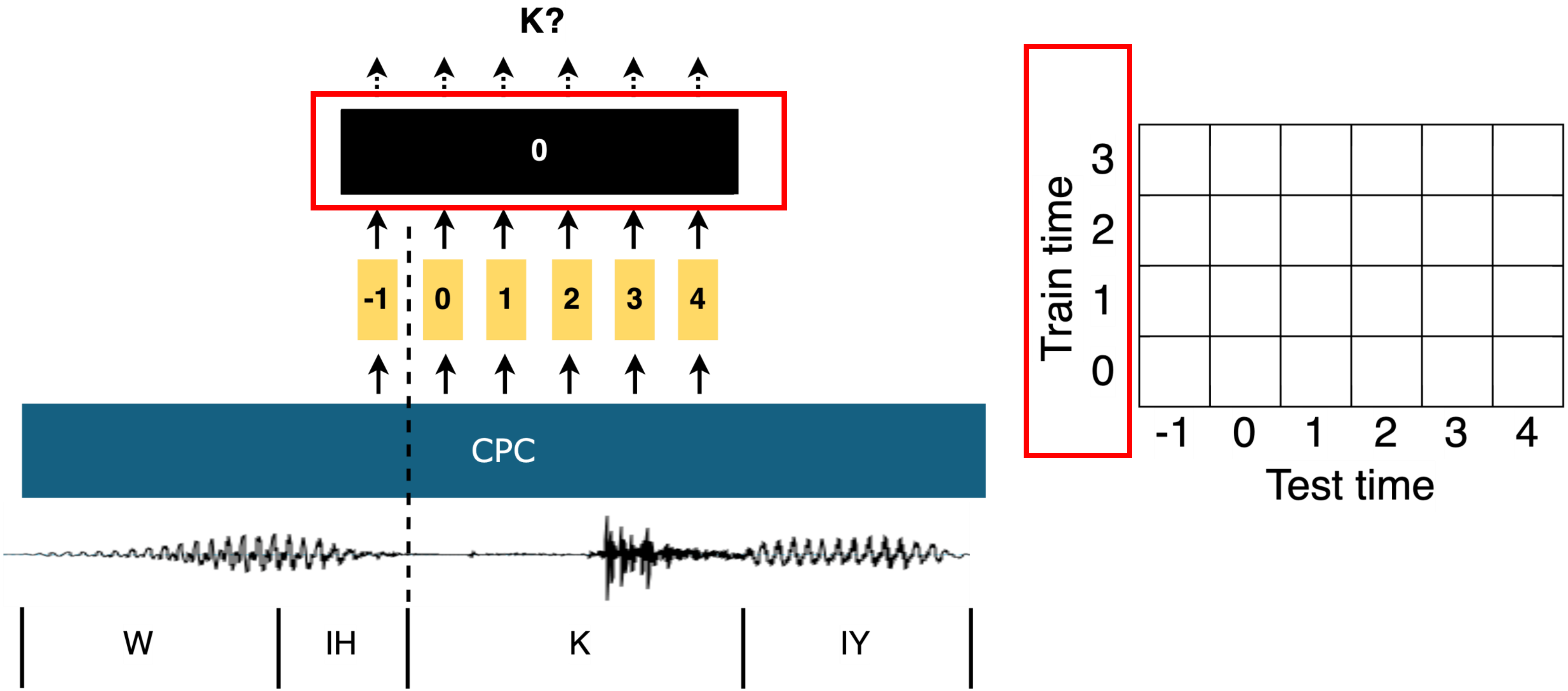
Does the encoding pattern change in this window?



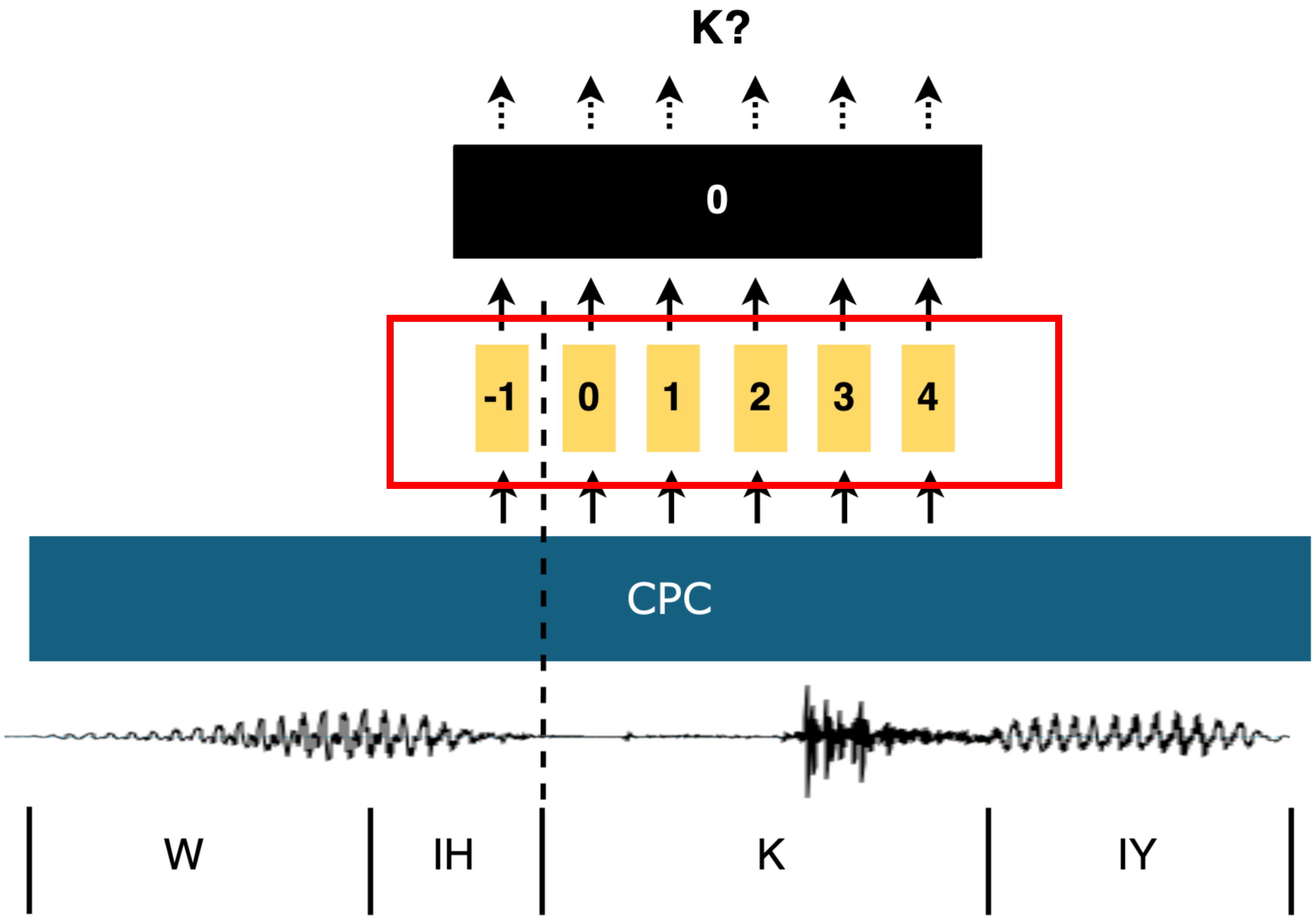
Does the encoding pattern change in this window?

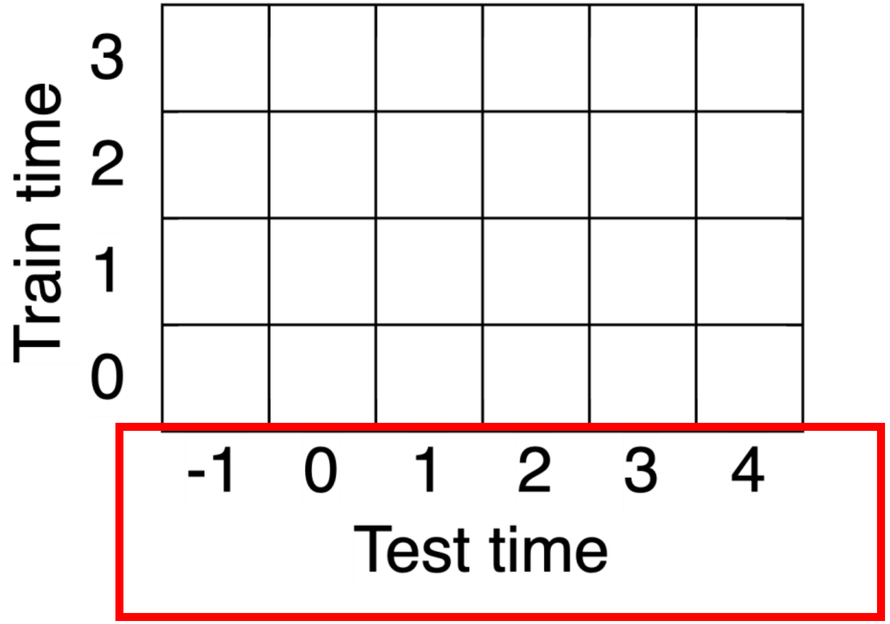


Does the encoding pattern change in this window?

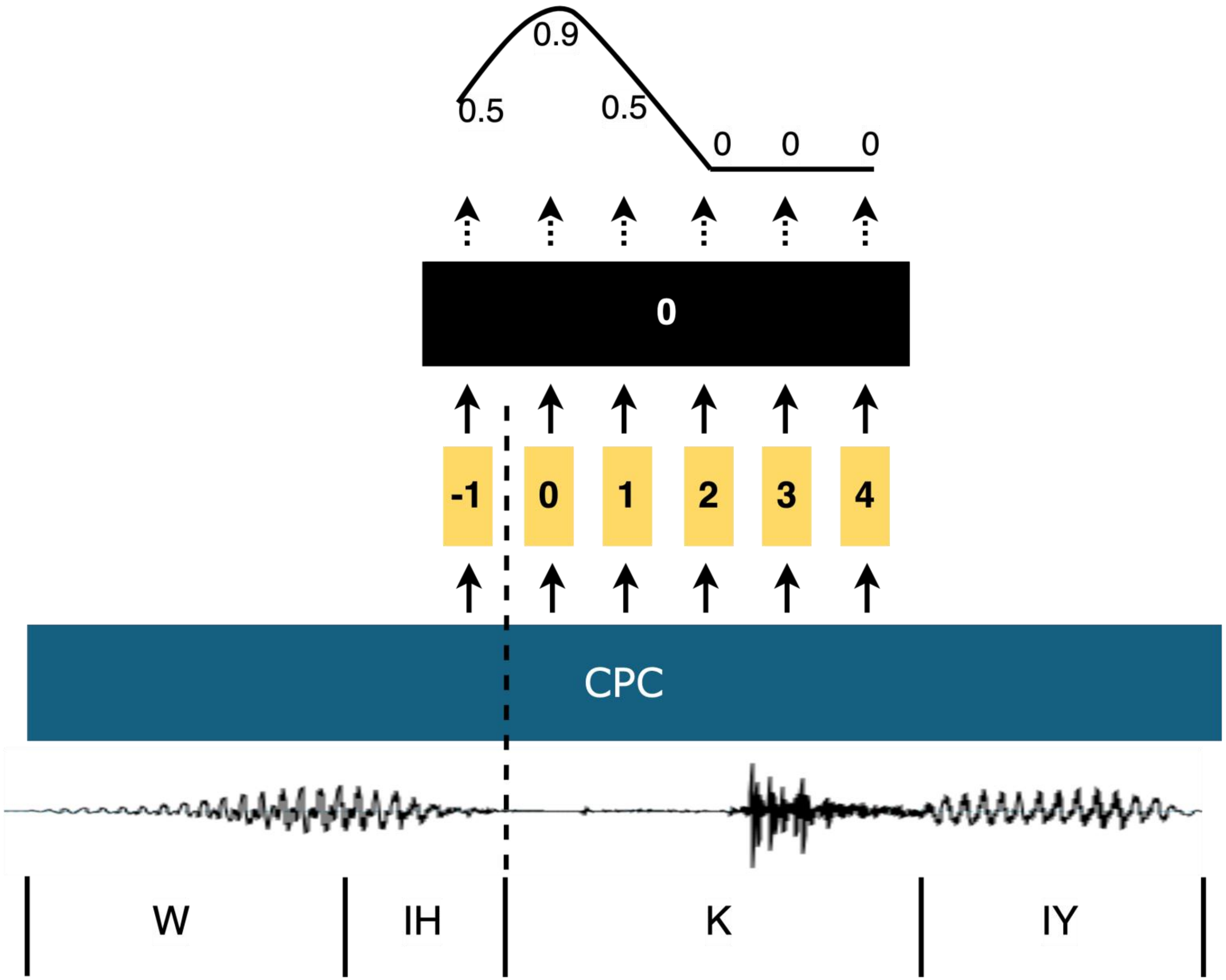


Does the encoding pattern change in this window?





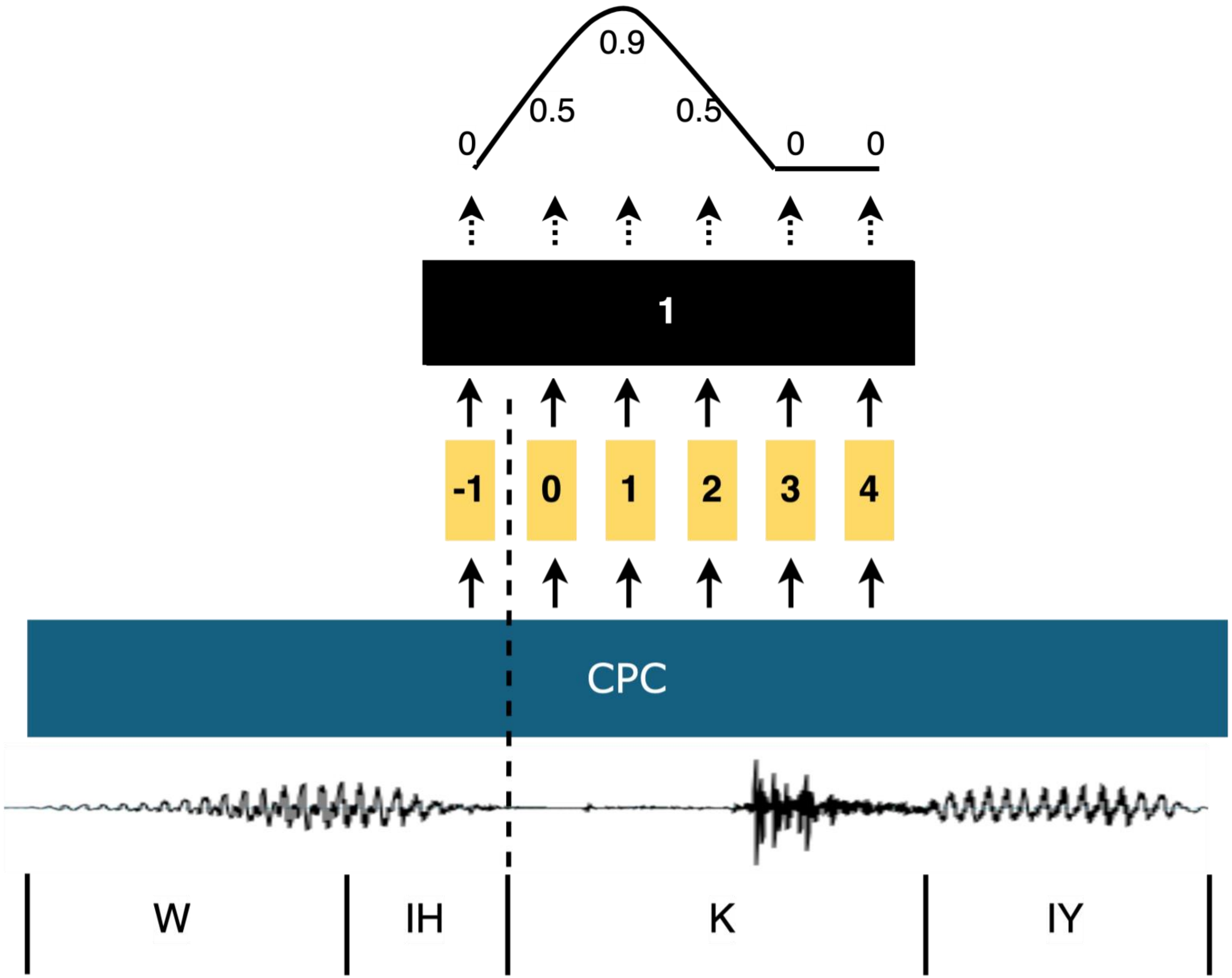
Does the encoding pattern change in this window?



Train time						
3						
2						
1						
0	0.5	0.9	0.5	0	0	0
	-1	0	1	2	3	4

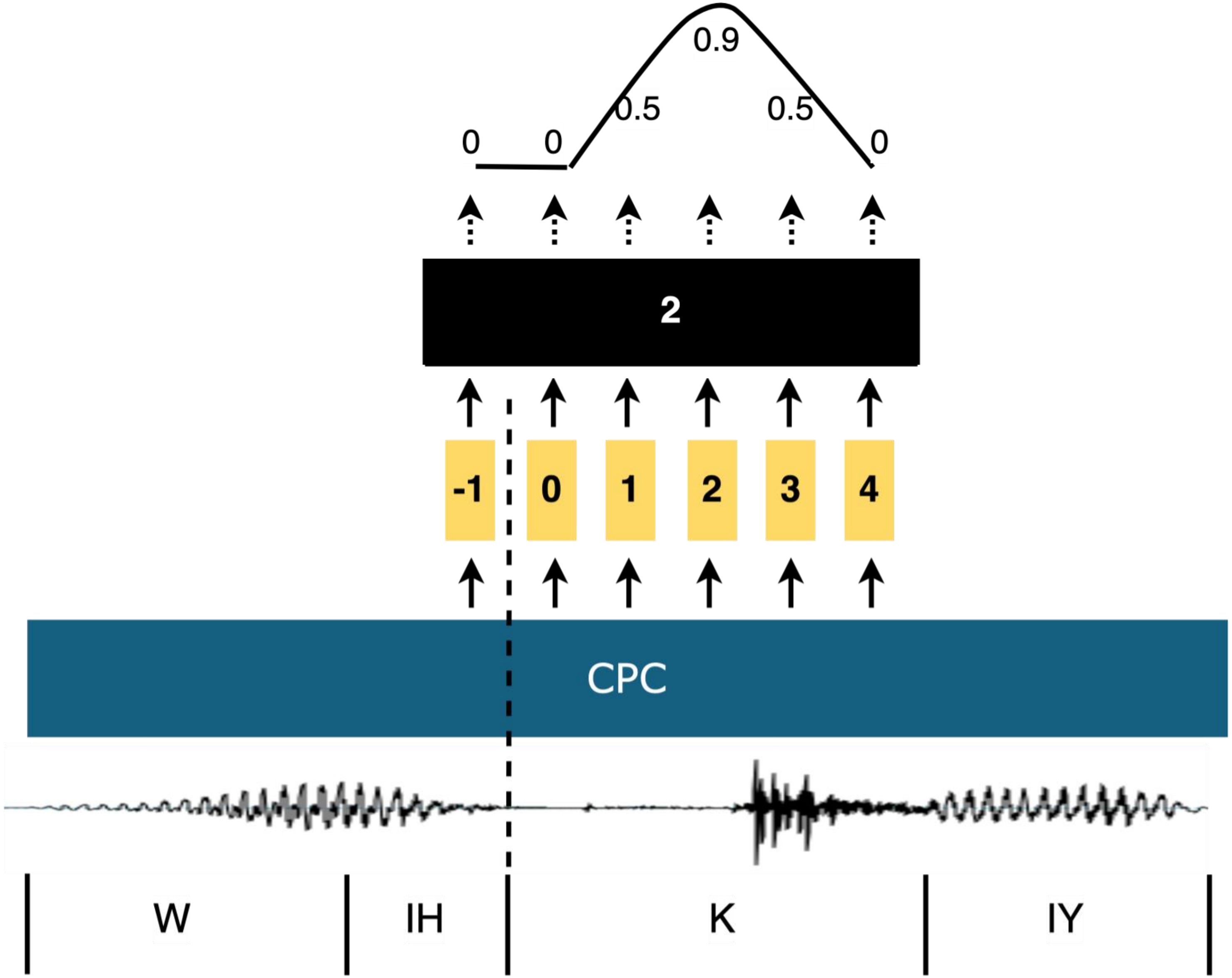
Test time

Does the encoding pattern change in this window?



Train time \ Test time	-1	0	1	2	3	4
3						
2						
1	0	0.5	0.9	0.5	0	0
0	0.5	0.9	0.5	0	0	0

Does the encoding pattern change in this window?

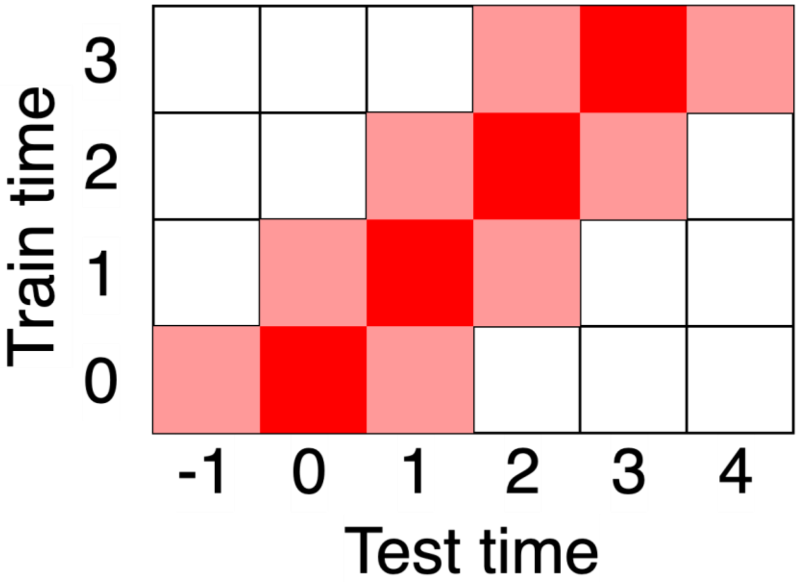


Train time \ Test time	-1	0	1	2	3	4
3						
2	0	0	0.5	0.9	0.5	0
1	0	0.5	0.9	0.5	0	0
0	0.5	0.9	0.5	0	0	0

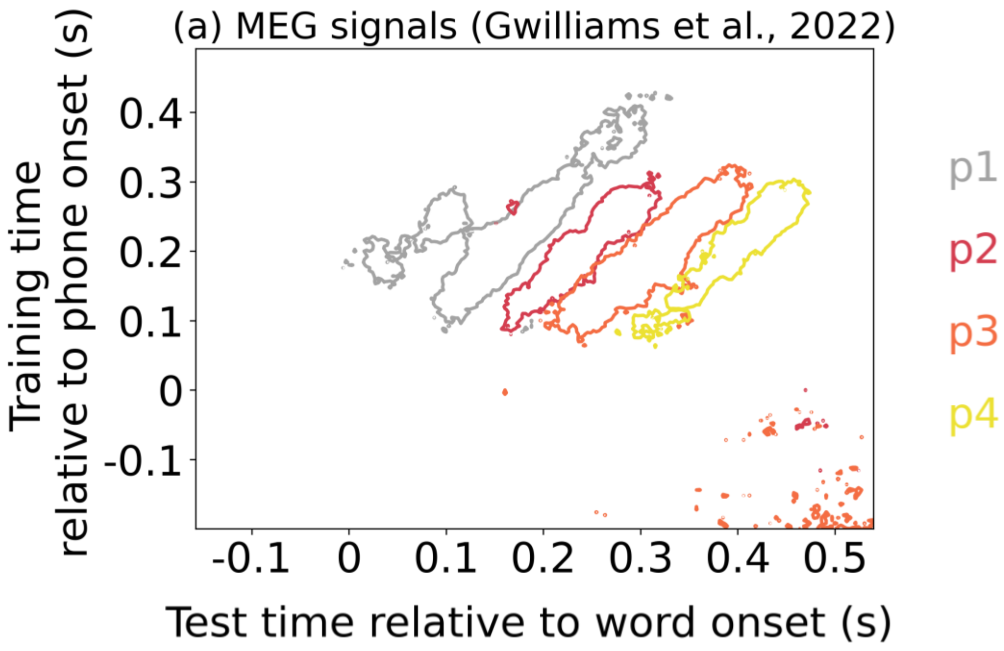
Does the encoding pattern change in this window?

Train time \ Test time	-1	0	1	2	3	4
3	0	0	0	0.5	0.9	0.5
2	0	0	0.5	0.9	0.5	0
1	0	0.5	0.9	0.5	0	0
0	0.5	0.9	0.5	0	0	0

Does the encoding pattern change in this window?

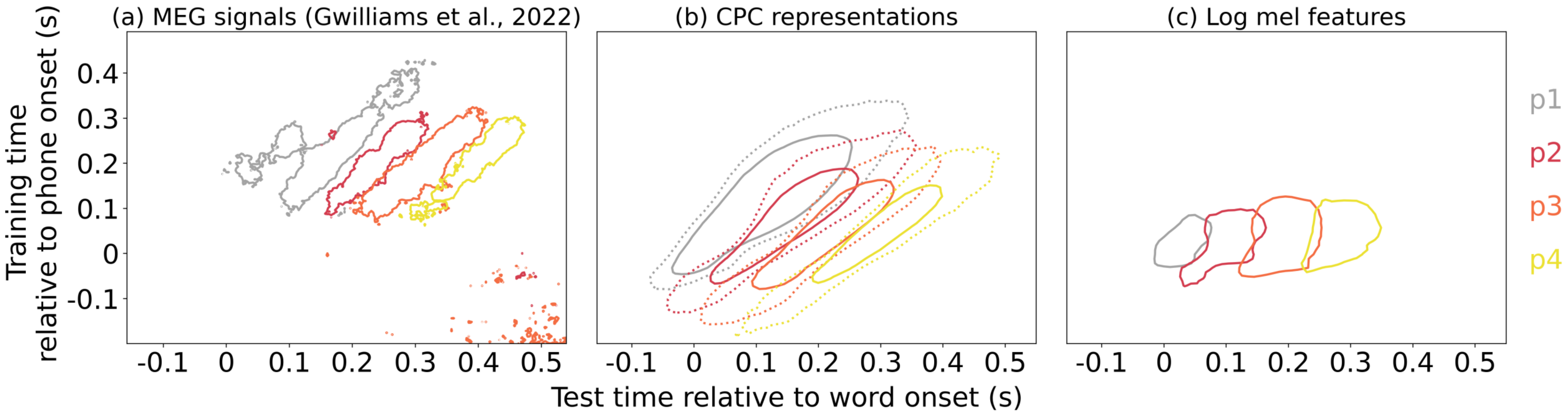


Dynamic encoding in brain signals



- Brains encode three successive phones simultaneously
- The encoding pattern evolves over time
  - Encoding temporal information

Dynamic encoding in brain signals and in model representations



Conclusions (part 2)

Dynamic encoding can be acquired through predictive learning

- Does not rely on top-down information / linguistic knowledge
- Follow-up: would we see the same pattern in the same model trained on non-speech audio scenes?

Outline

In the representation space of self-supervised learning models:

1. Speaker information is encoded orthogonally to phonetic information

2. Multiple successive phones are encoded at the same time

3. There is some extent of cross-context generalizability

Context-invariant phonemic representations

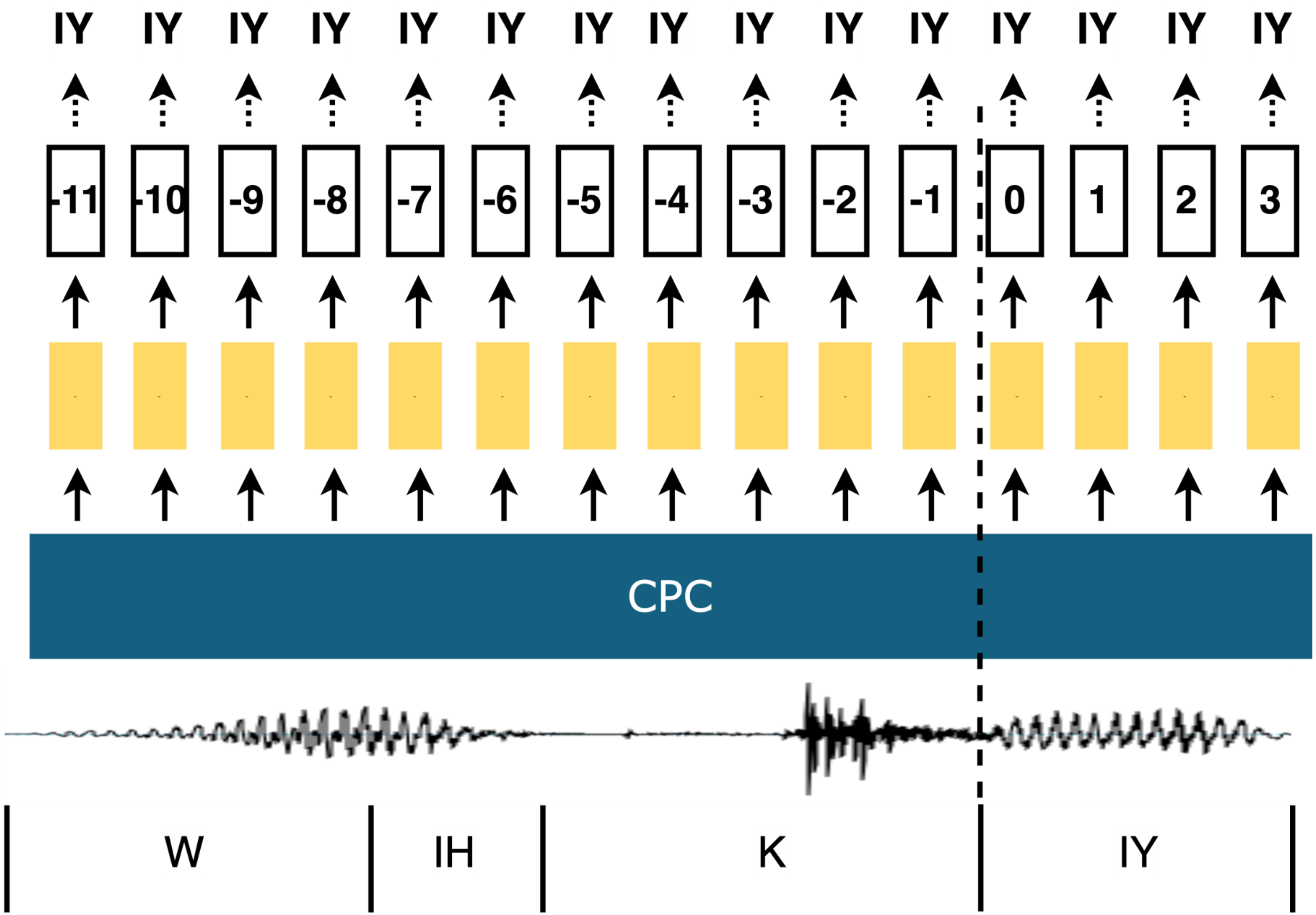
Gwilliams et al. (2022) found that the encoding patterns support some degree of cross-*position* generalization and implied there is context-invariant phonemic representations.

- Phone position conflates different contexts
- They did not report results on acoustic features

Gwilliams et al. (2022) "Neural dynamics of phoneme sequences reveal position-invariant code for content and order." *Nature communications*

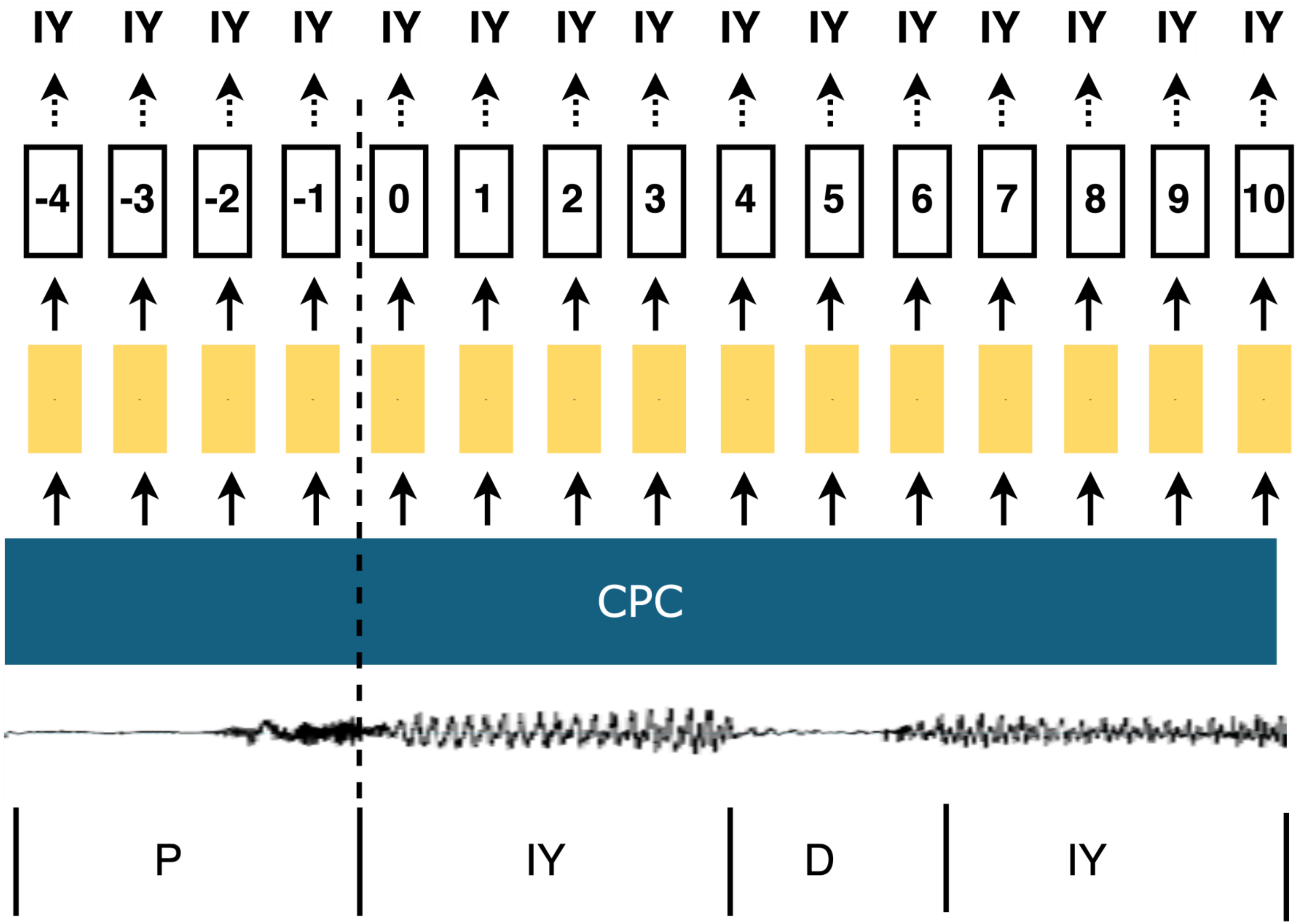
Does the encoding pattern of a phoneme generalize across contexts?

Train on vowels following /k/



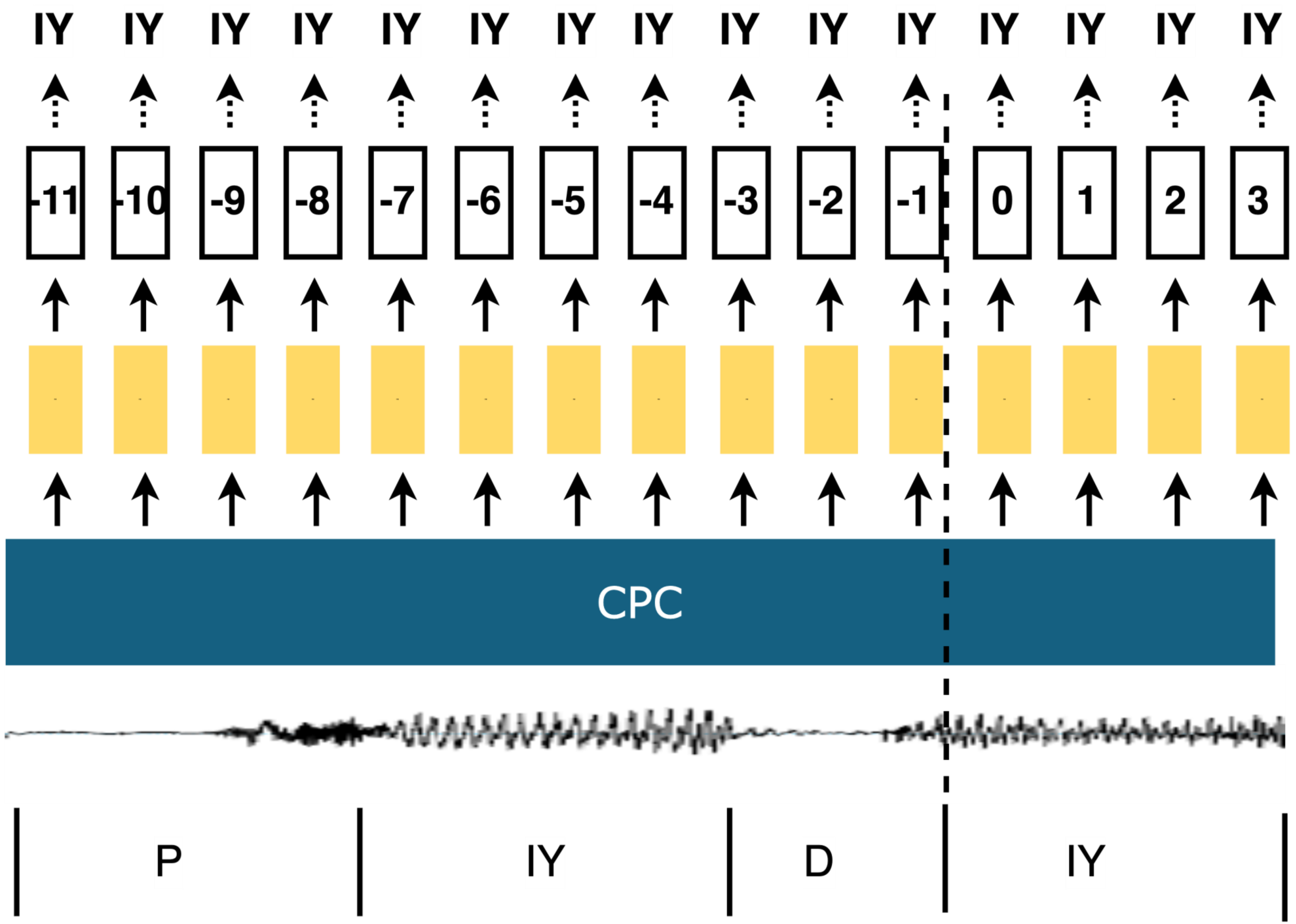
Does the encoding pattern of a phoneme generalize across contexts?

Test on vowels following /p/



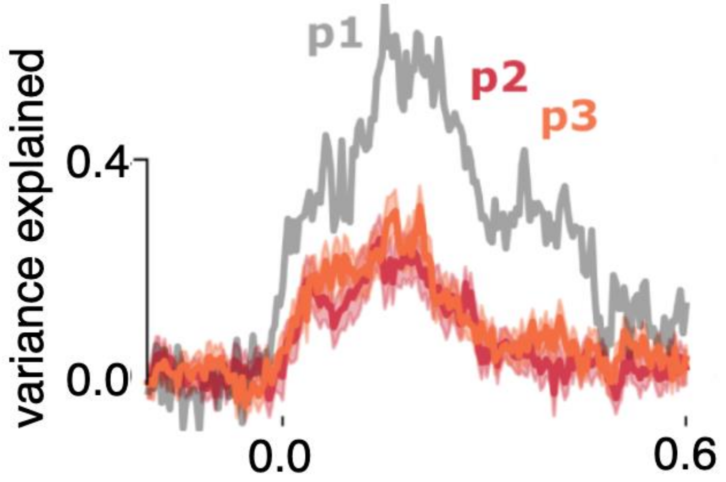
Does the encoding pattern of a phoneme generalize across contexts?

Test on vowels following /d/

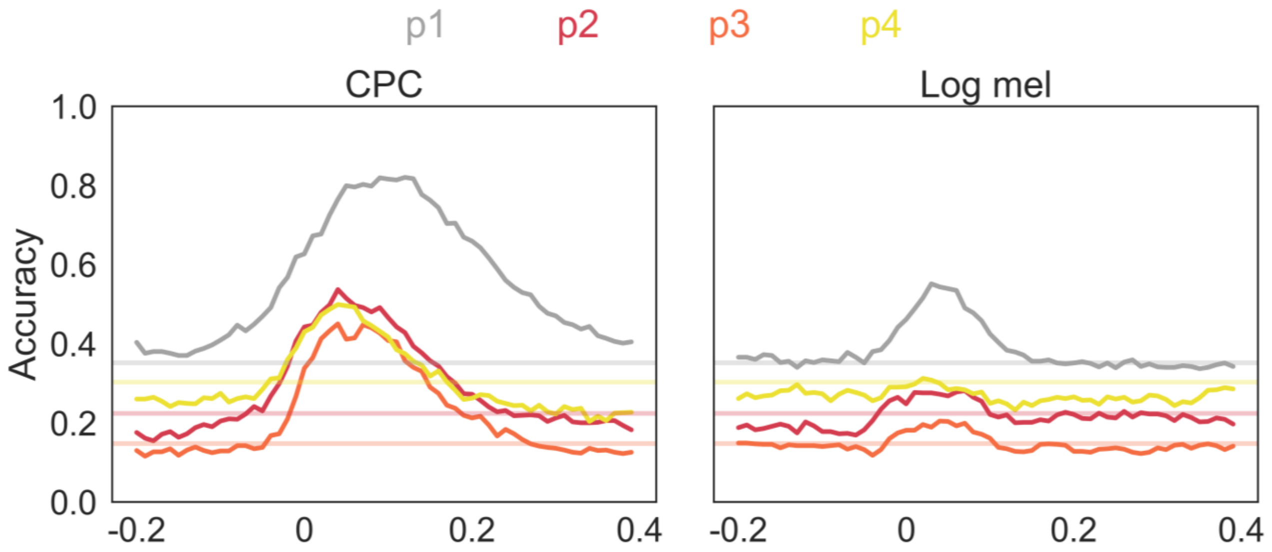


Do the encoding patterns generalize across positions?

Partial generalization in brain signals



Time relative to phone onset (s)



Time relative to phone onset (s)

And in the models, but also some generalization in acoustic features.

- Cross-context generalization tests showed similar patterns.
- The degree of generalization correlates with acoustic similarity.

Conclusions (part 3)

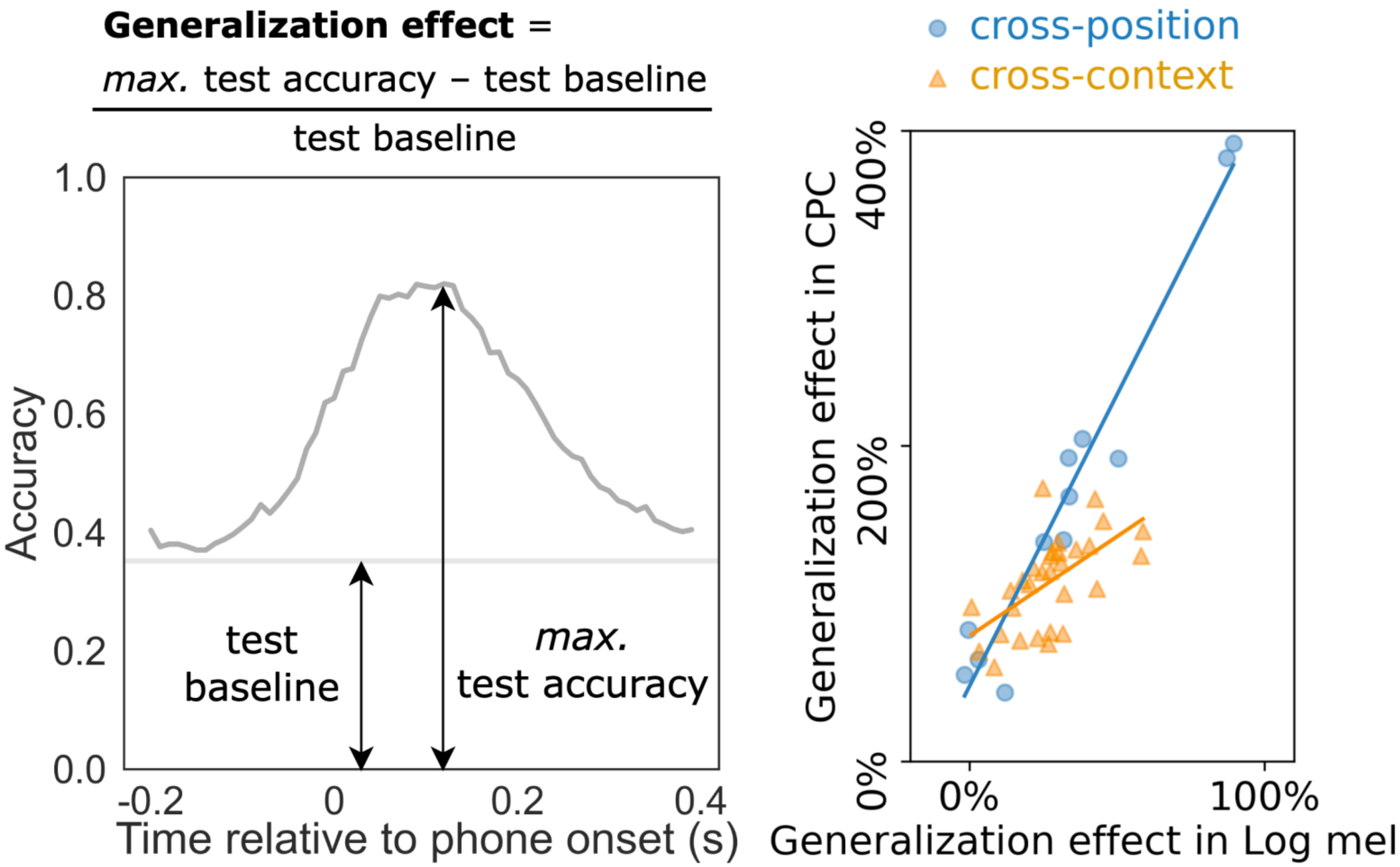
There is insufficient evidence for context-invariant phonemic encoding in either models or brains.

- Top-down information used to identify context-dependent encoding?
- Do we really need context-invariant SSL representations?

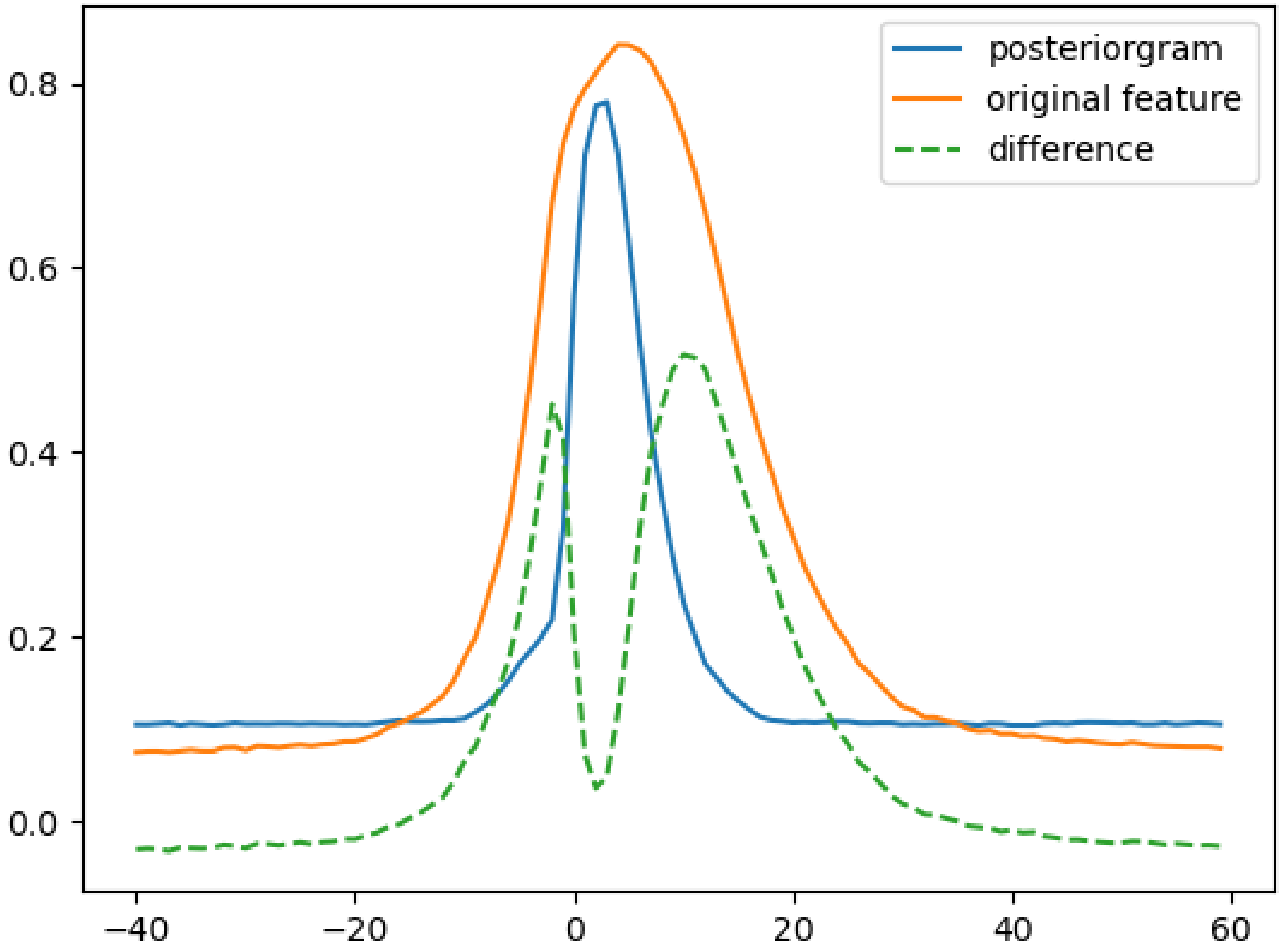
Overall conclusions

- SSL models
  - Readily disentangle speaker and phonetic information
  - Develop temporal dynamics like brains
  - Absence of fully context-invariant phonemic representation
- More broadly
  - SSL models can shed light on speech representations in humans
  - Neuroscience studies offer novel perspectives for analyzing NNs

The generalization effect is dependent on acoustic similarity



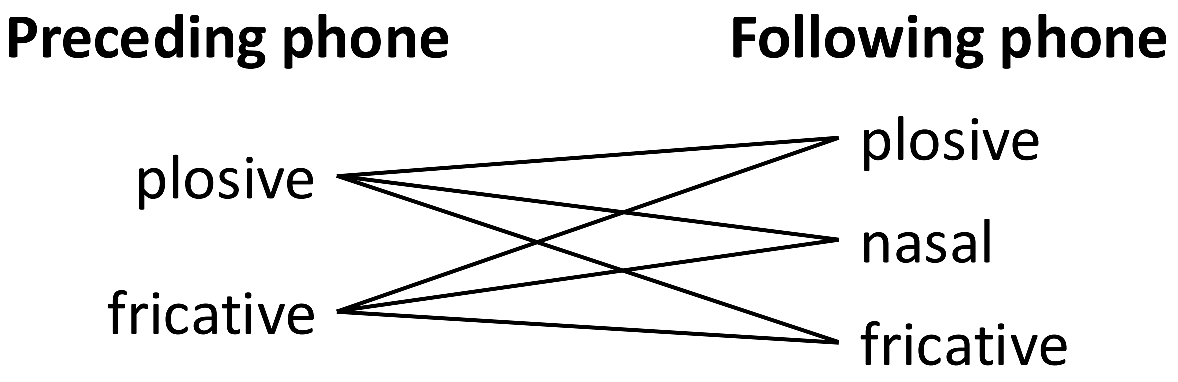
Decoding accuracy



Time relative to phone onset

Note

- Gwilliams et al. (2022) only reported cross-position generalization
- We tested both cross-position and cross-context generalization.
  - For controllability, we only considered vowel classification
  - For phonetic contexts, we only considered the manner of articulation of the preceding and following phone



Do the encoding patterns generalize across contexts?