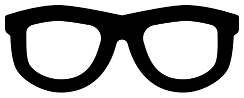
A predictive learning model can simulate ***temporal dynamics*** and ***context effects*** found in neural representations of continuous speech

Oli Danyi Liu, Hao Tang, Naomi Feldman, Sharon Goldwater

University of Edinburgh, University of Maryland College Park

Many perceptual processes involve tracking and integrating sequentially presented objects



Reading

Motion tracking

Speech

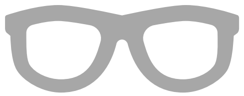
Music

Speech perception involves

~~Many perceptual processes involve~~

tracking and integrating

sequentially presented ~~objects~~ phones



Reading

Motion tracking

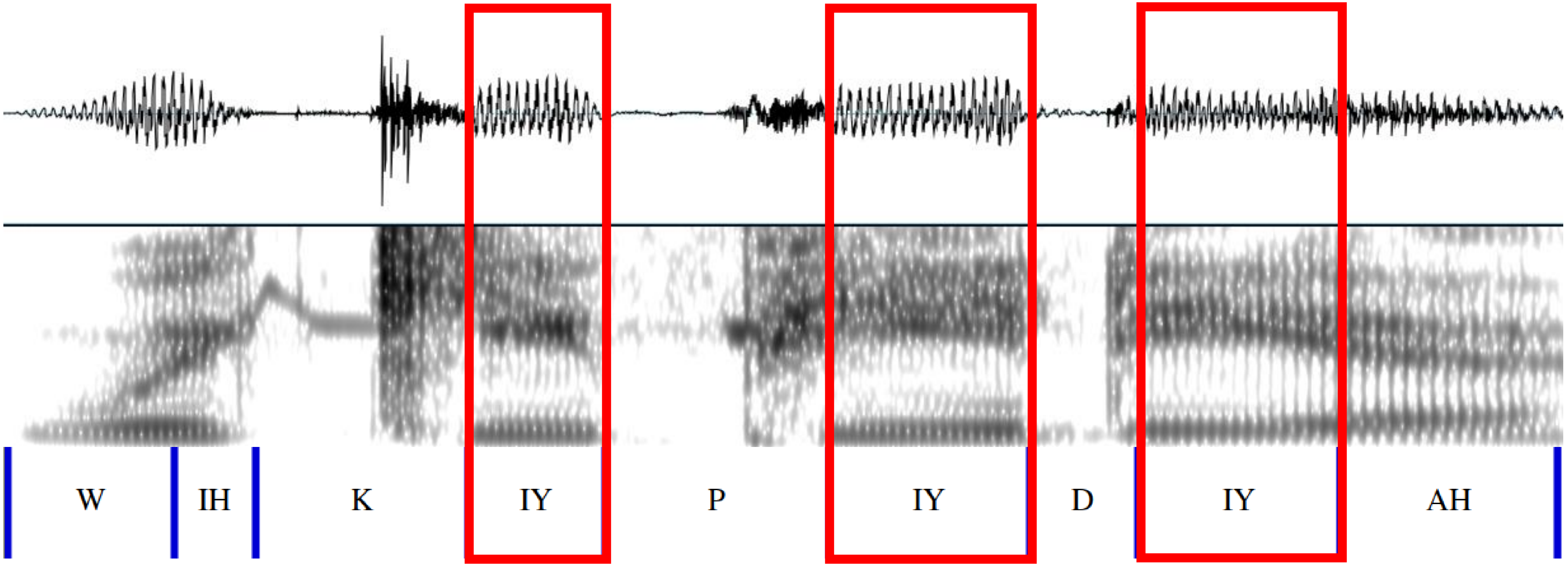
Speech

Music

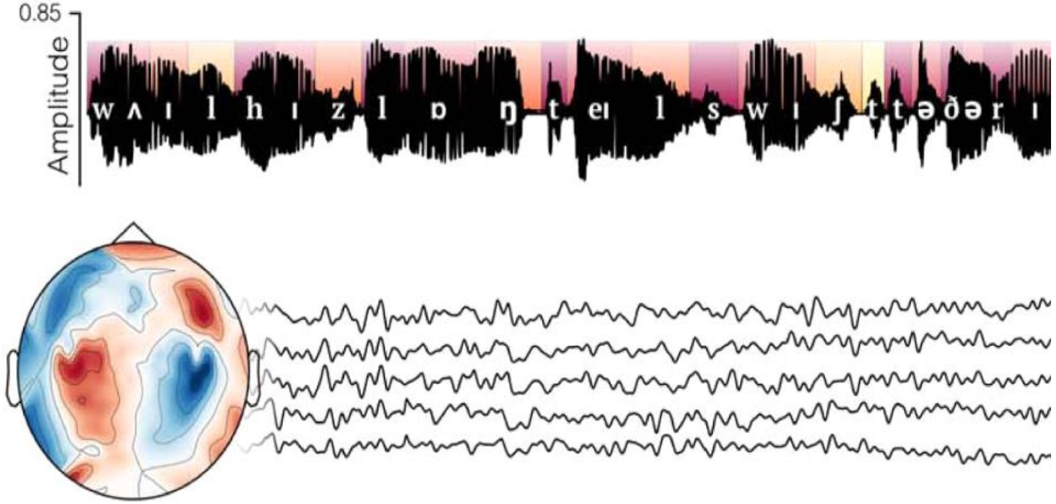
Speech perception involves tracking and integrating sequentially presented phones

“wikipedia”

The acoustic realization of a phoneme is sensitive to its surrounding context due to coarticulation



How do humans overcome these challenges?



Studies on how neural representations support this process:

Mesgarani, Cheung, Johnson, & Chang, 2014;
Khalighinejad, Cruzatto Da Silva, & Mesgarani, 2017;
Yi, Leonard, & Chang, 2019;
Hamilton & Huth, 2020…

Gwilliams, L., King, J.-R., Marantz, A., & Poeppel, D. (2022)
Neural dynamics of phoneme sequences reveal position-invariant code for content and order

Simulating properties found in neural signals

Gwilliams et al. (2022)

- analyzed MEG recordings from human listeners
- identified temporal dynamics and context effects

In this work, we simulated their analyses with *a computational model* to

- explore why or how these properties arise
    - Do we observe the same properties in the model?
- examine some open questions regarding the context effects

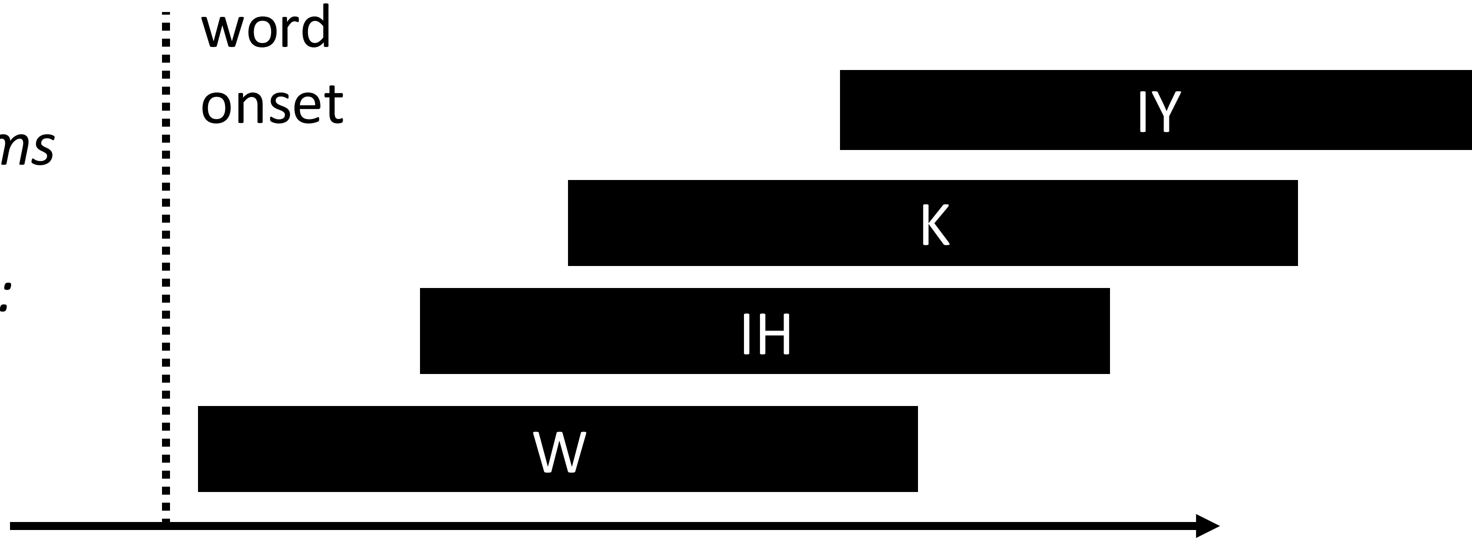
Findings from Gwilliams et al. (1)

Phones are encoded in the brain for longer than their actual durations.

Neural signals

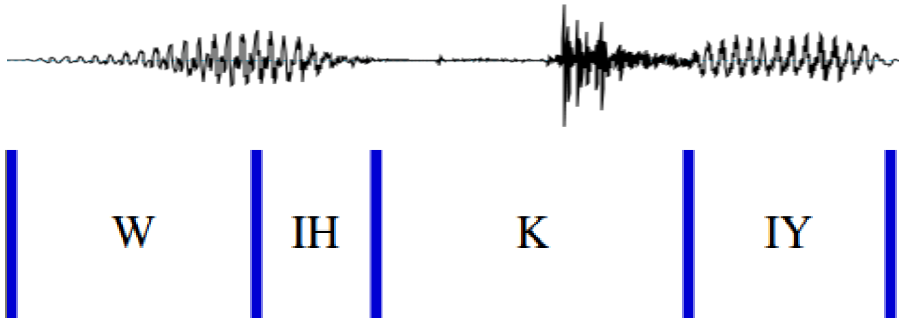
Phonetic features are maintained for up to 300ms

Khalighinejad et al., 2017: phoneme categories are encoded for up to 350ms



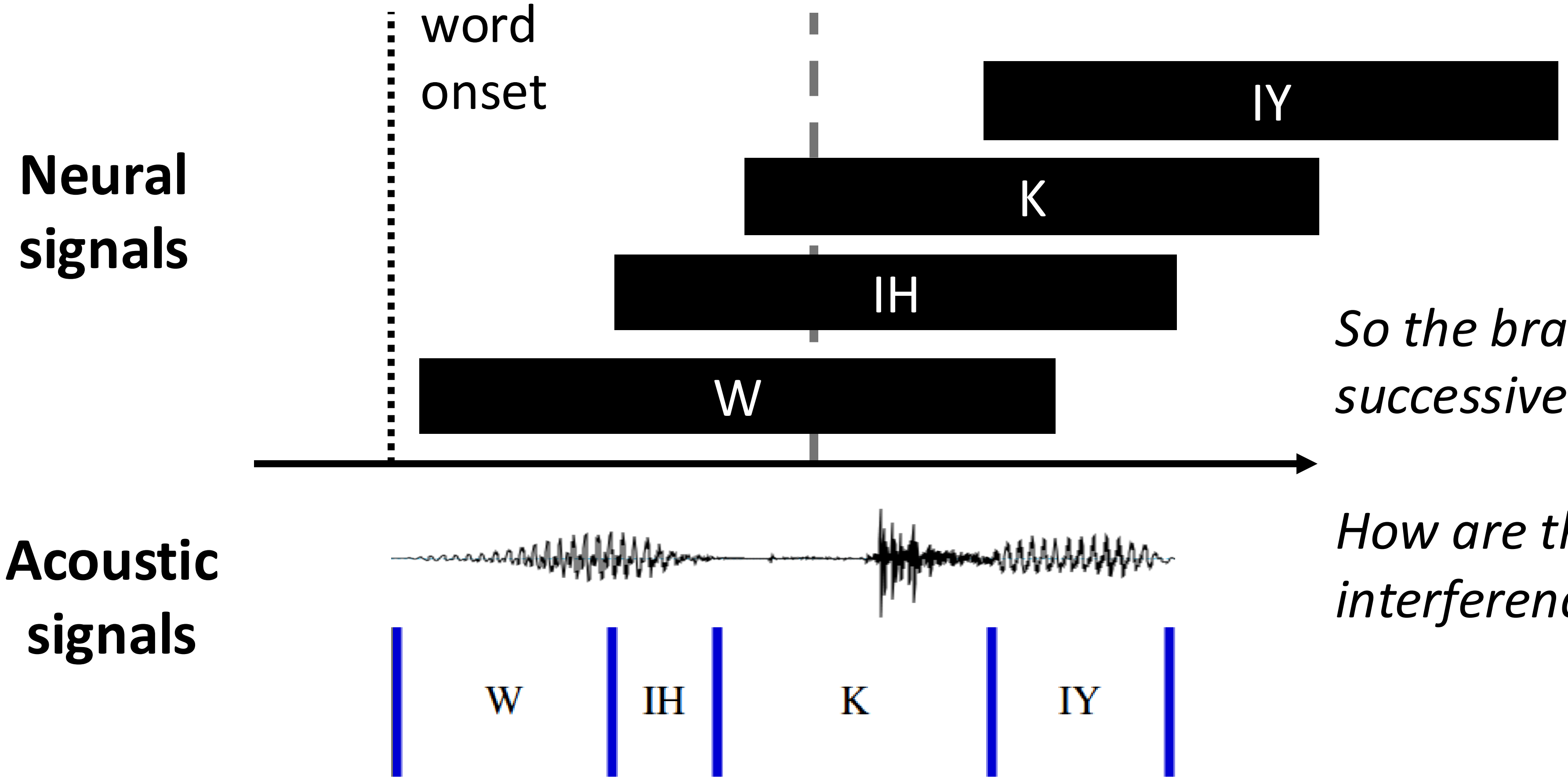
Acoustic signals

average phone duration: 80ms



Findings from Gwilliams et al. (1)

Phones are encoded in the brain for longer than their actual durations.

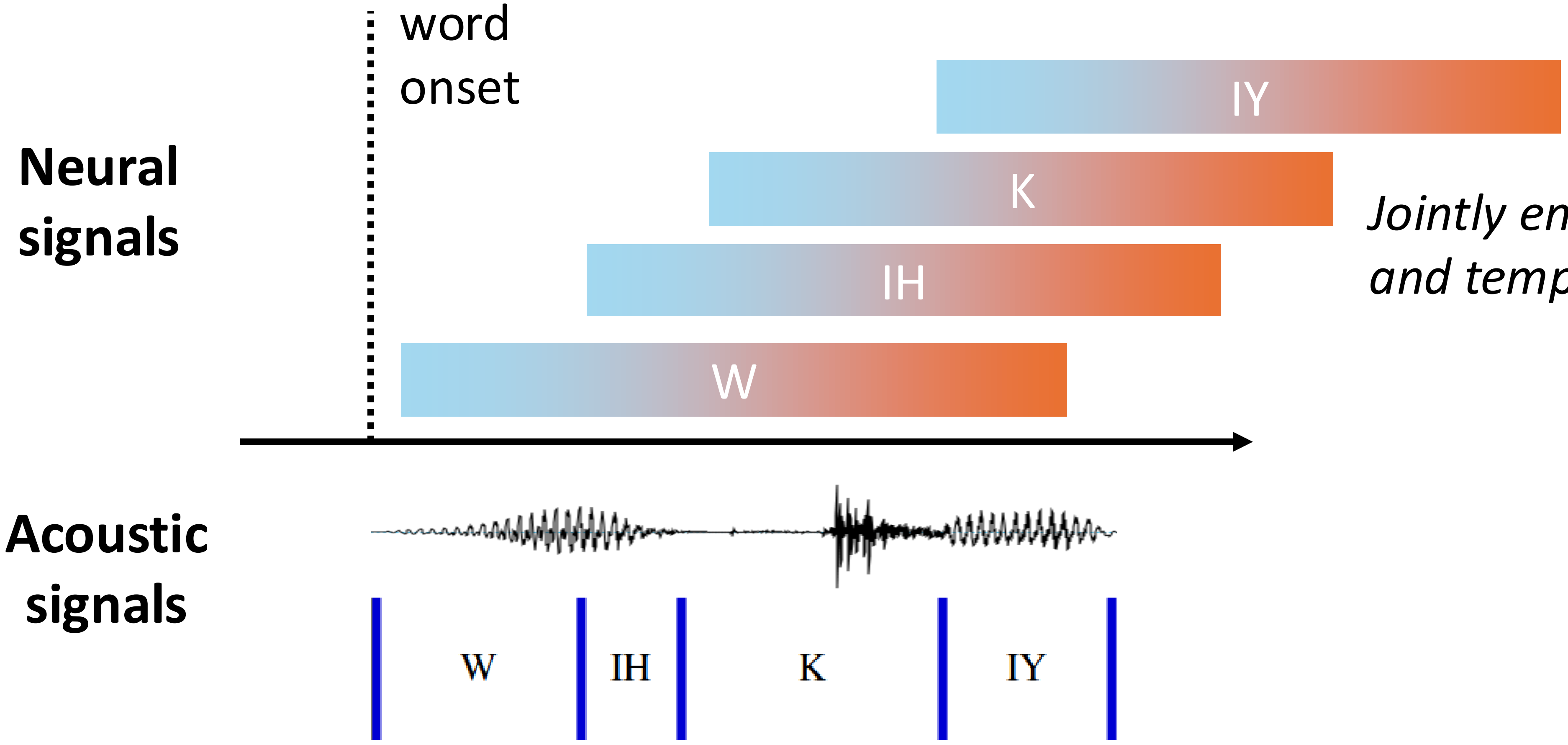


So the brain encodes multiple successive phones simultaneously.

How are they maintained without interference?

Findings from Gwilliams et al. (2)

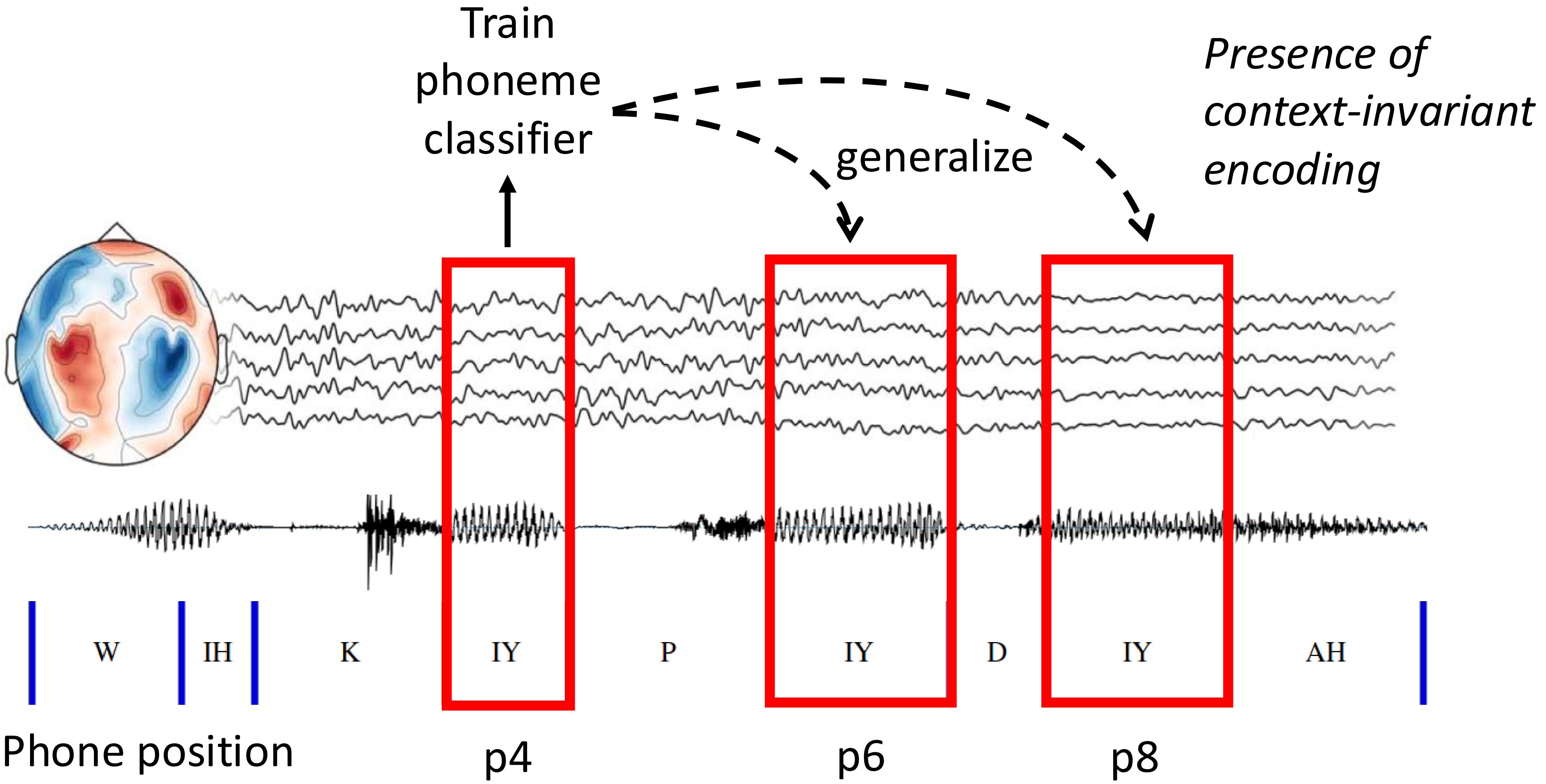
The encoding pattern of a phone evolves over time.



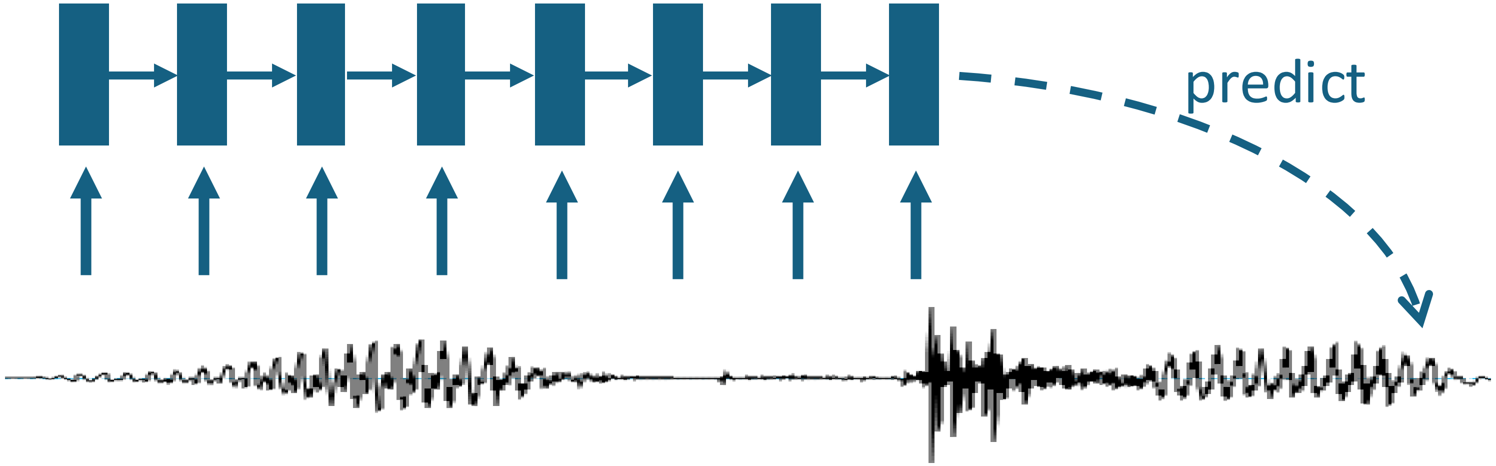
Jointly encoding phonetic and temporal information

Findings from Gwilliams et al. (3)

The encoding pattern generalizes across phone position.



The model we used

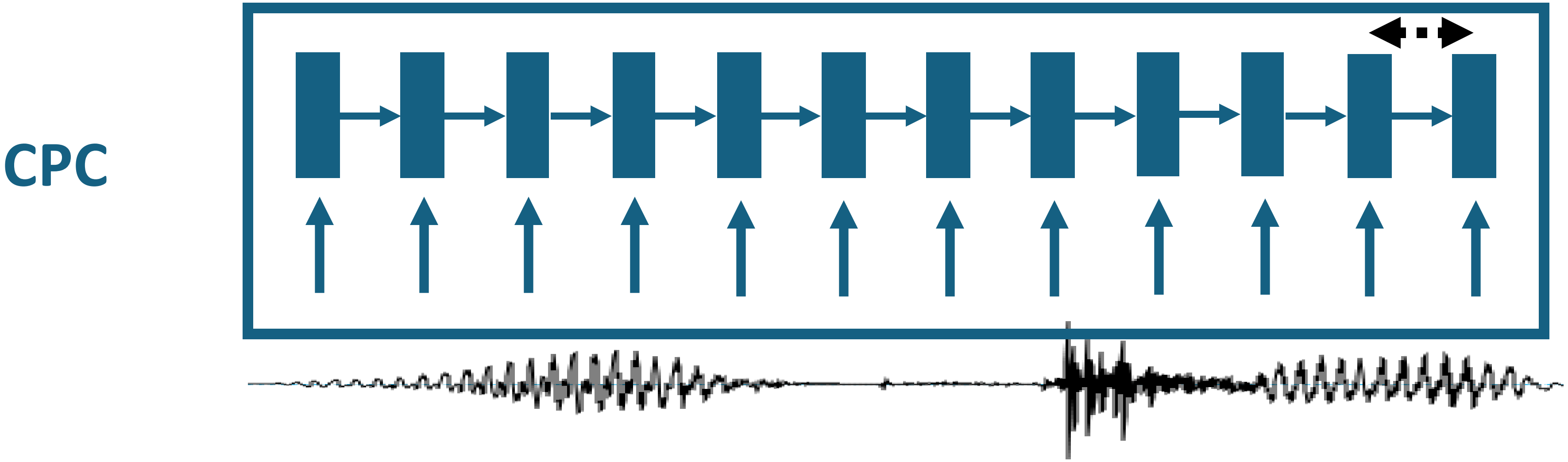


- **Architecture**: LSTM (recurrent neural network)
- **Learning mechanism:** predict upcoming acoustics based on past context in utterance
  - Trained on raw speech waveforms (audiobooks) without access to texts

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

The model we used

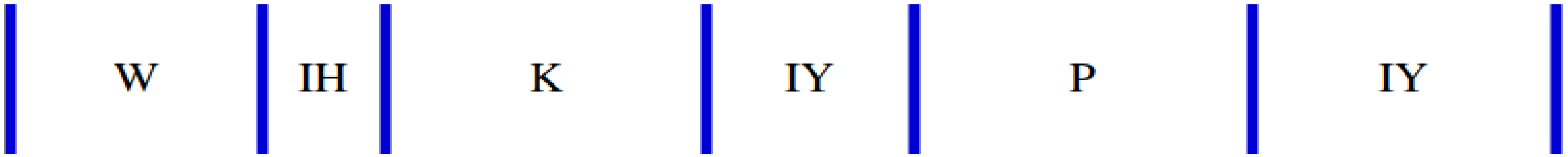
Representations: 512-dimensional vectors spaced by 10ms



- **Architecture**: LSTM (recurrent neural network)
- **Learning mechanism:** predict upcoming acoustics
  - Trained on raw speech waveforms (audiobooks) without access to text

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

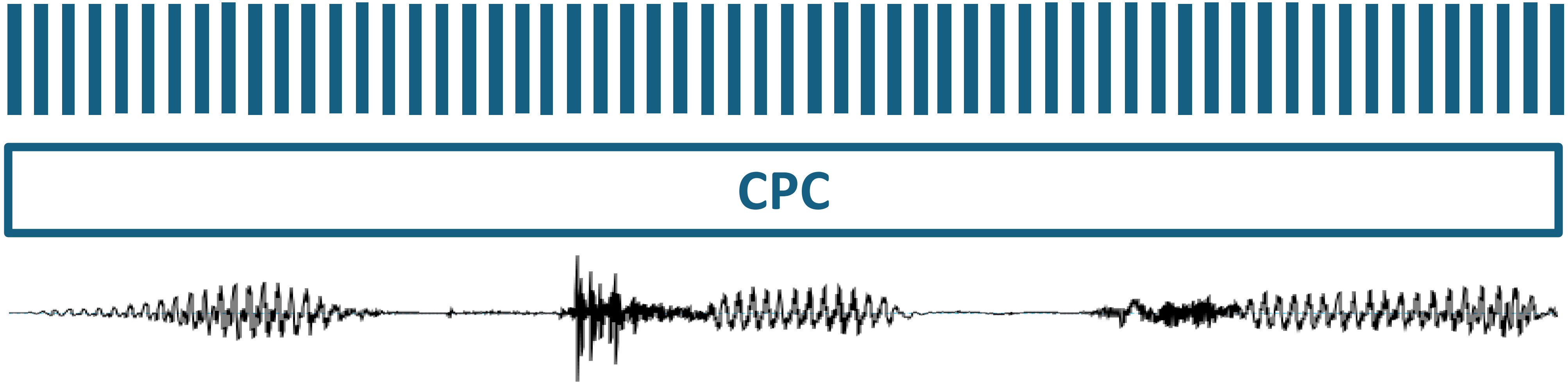
Decoding phonemes from representations



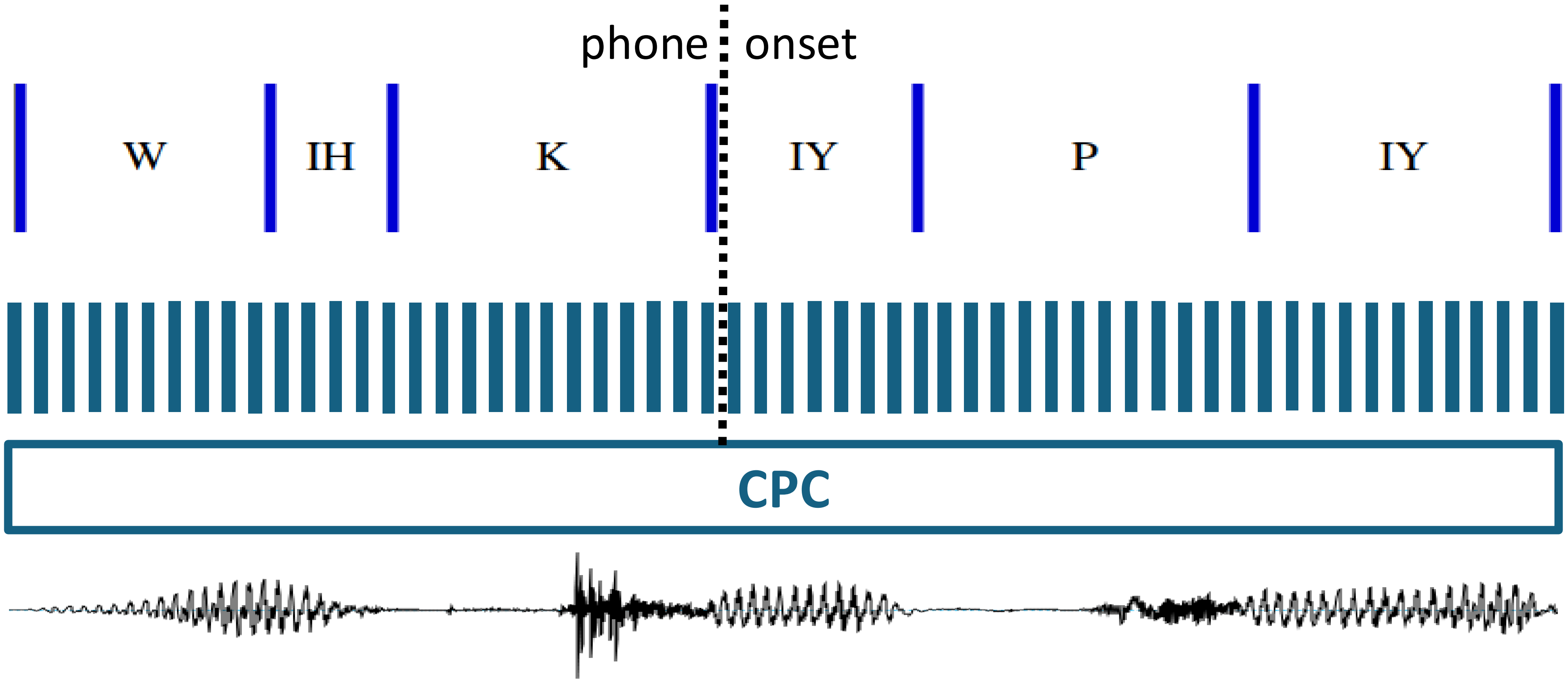
Target: phoneme category

Decoder: linear classifier

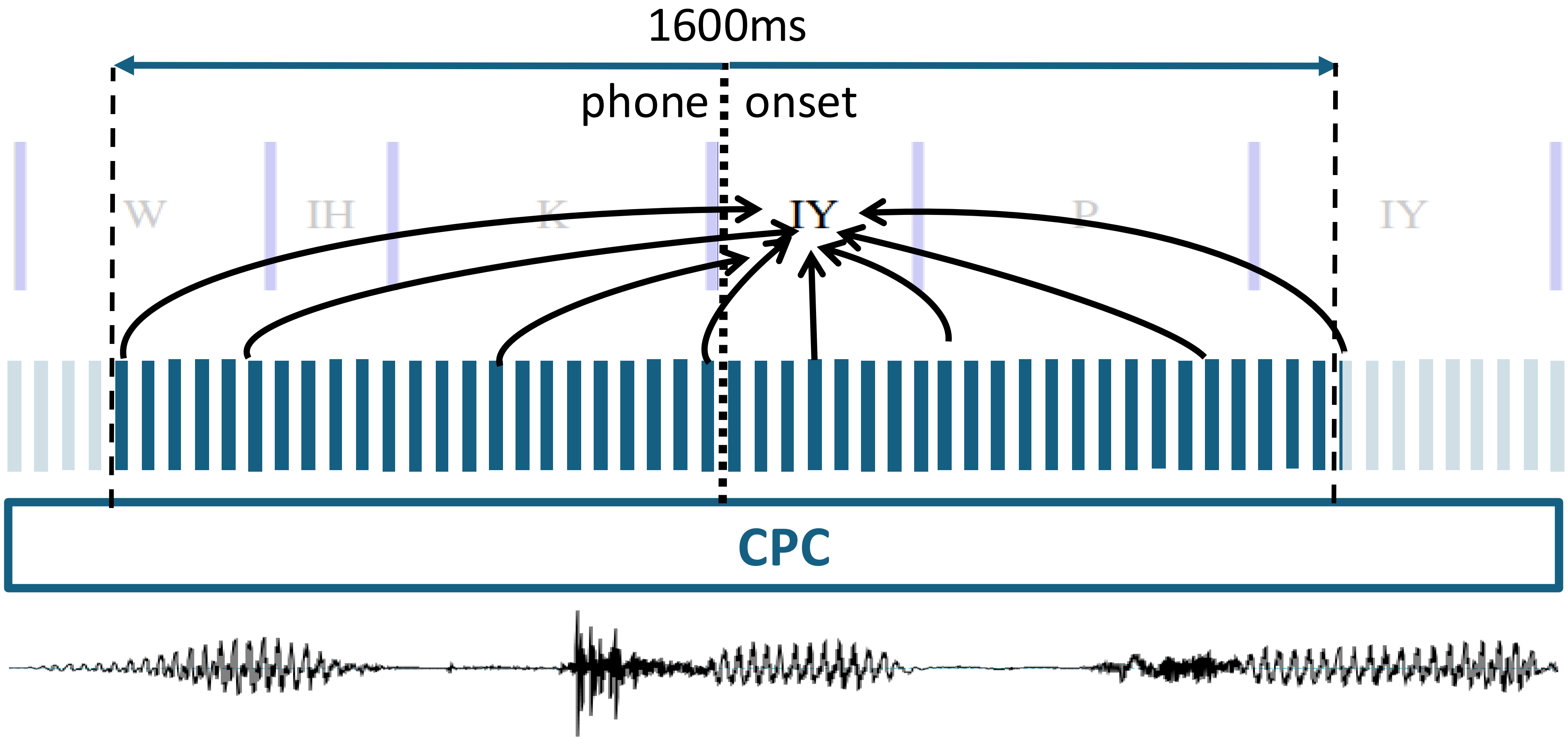
Input: one representation vector



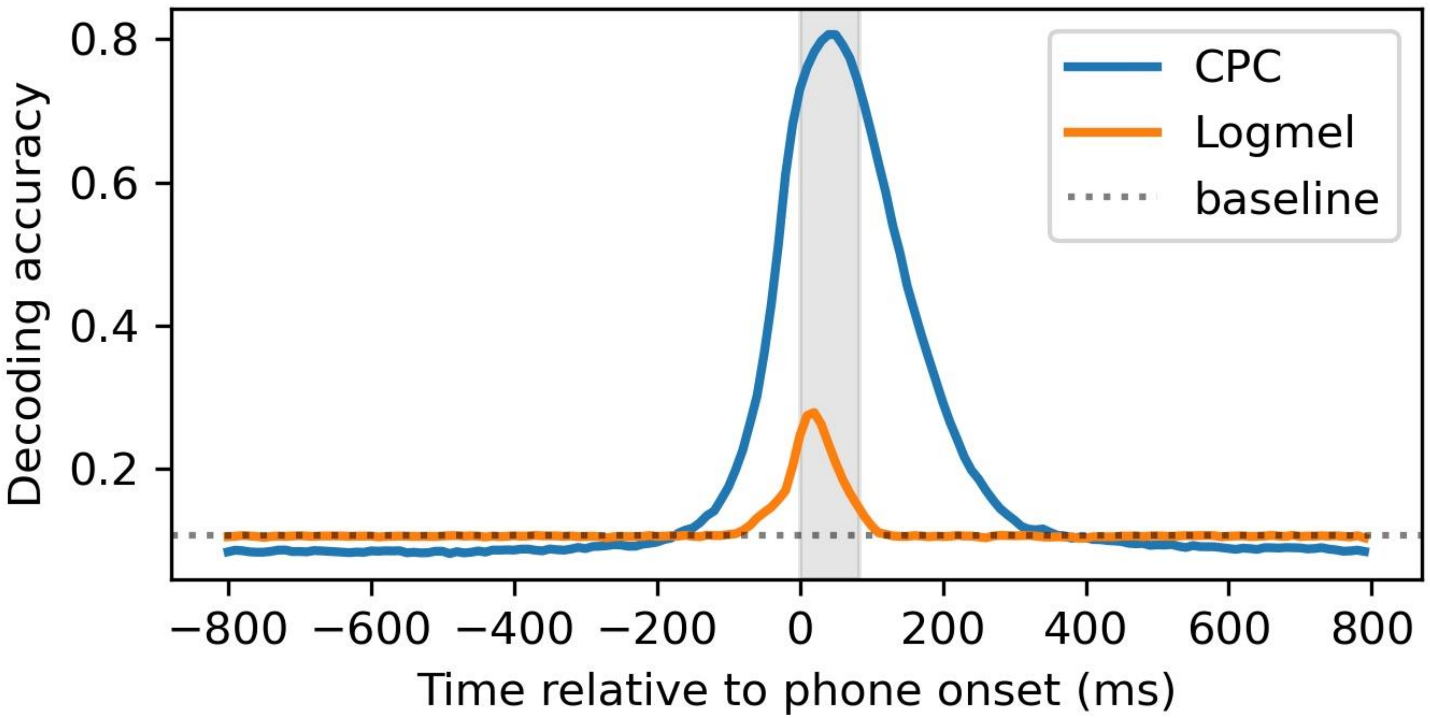
How long is a phone encoded for?



How long is a phone encoded for?



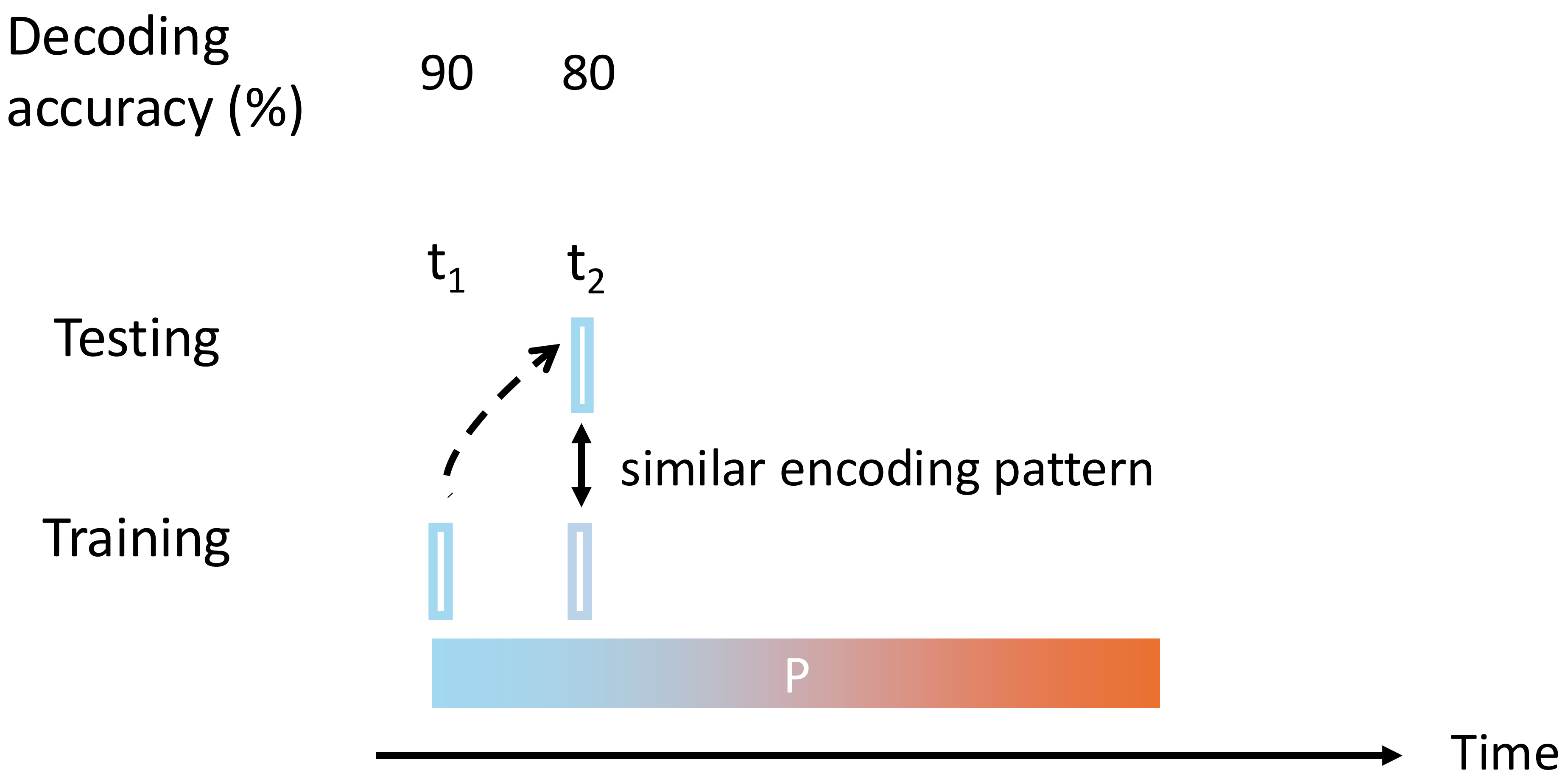
The window of phonetic decodability



Like brains, the model encode each phone for longer than its duration.

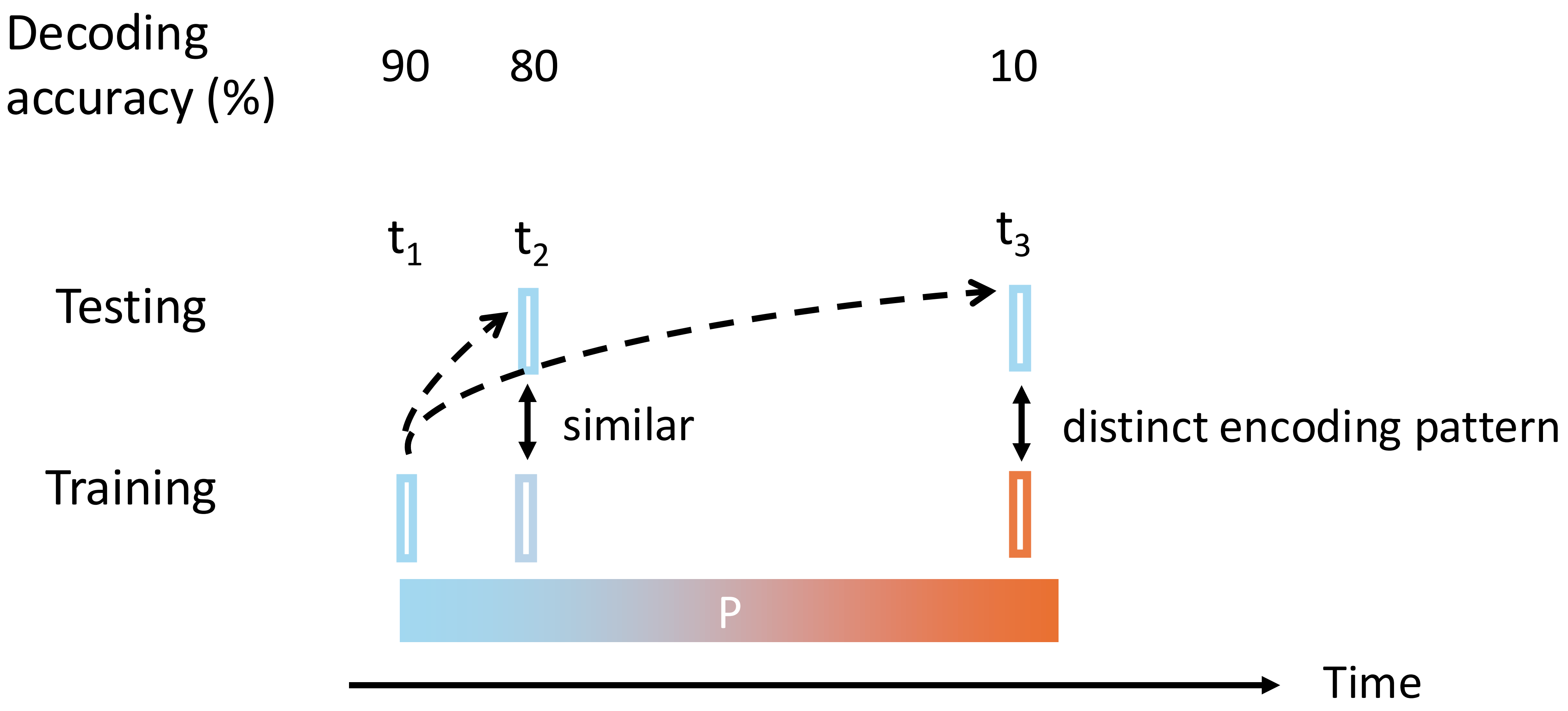
Does the encoding pattern evolve in this window?

Does the encoding pattern identified for t_1 generalize to t_2?



Does the encoding pattern evolve in this window?

Does the encoding pattern identified for t_1 generalize to t_2?



If the encoding pattern is evolving

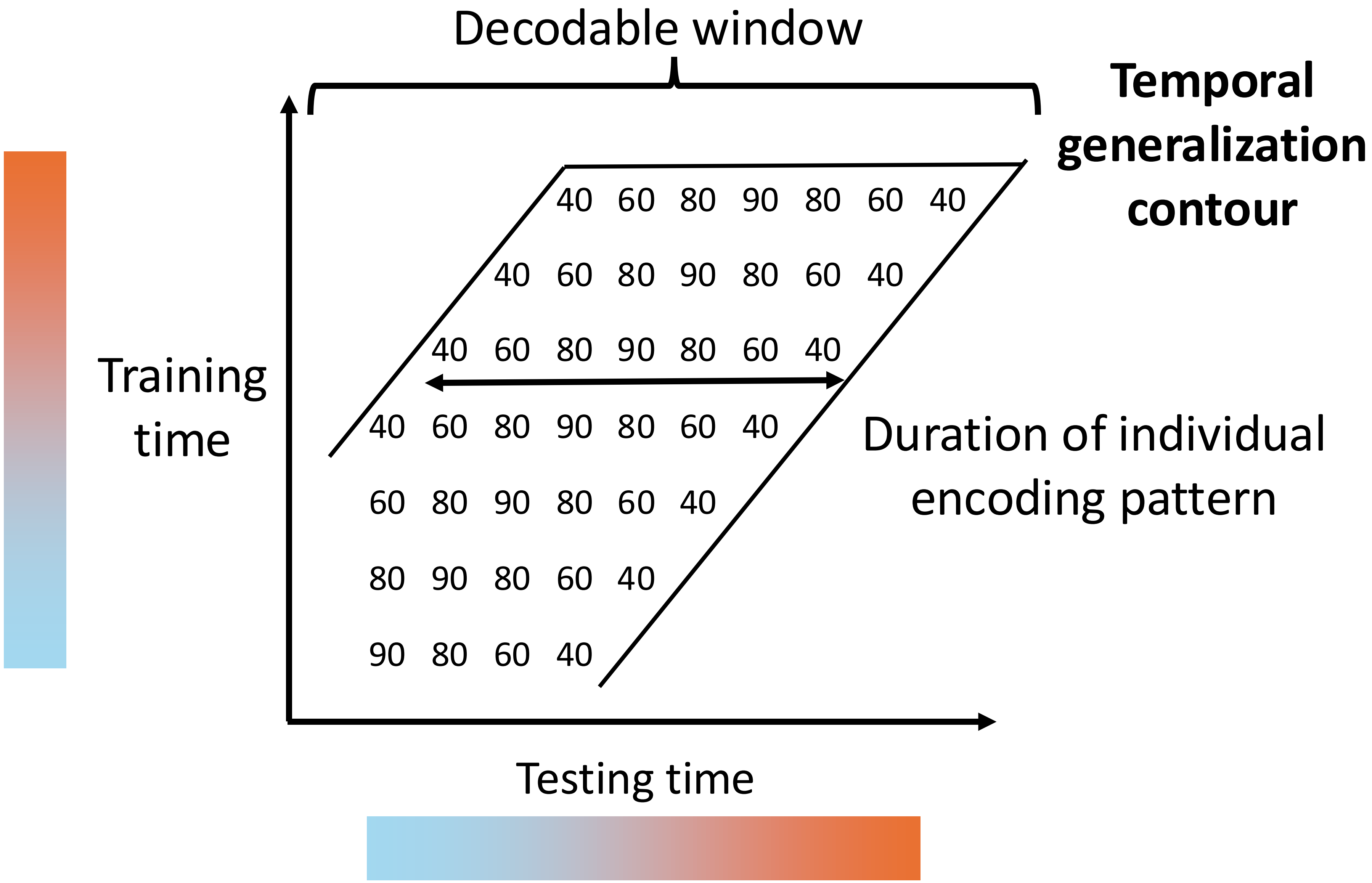
Temporal generalization matrix

Training time

10	10	20	40	60	80	90	80	60	40
10	20	40	60	80	90	80	60	40	20
20	40	60	80	90	80	60	40	20	10
40	60	80	90	80	60	40	20	10	10
60	80	90	80	60	40	20	10	10	10
80	90	80	60	40	20	10	10	10	10
90	80	60	40	20	10	10	10	10	10

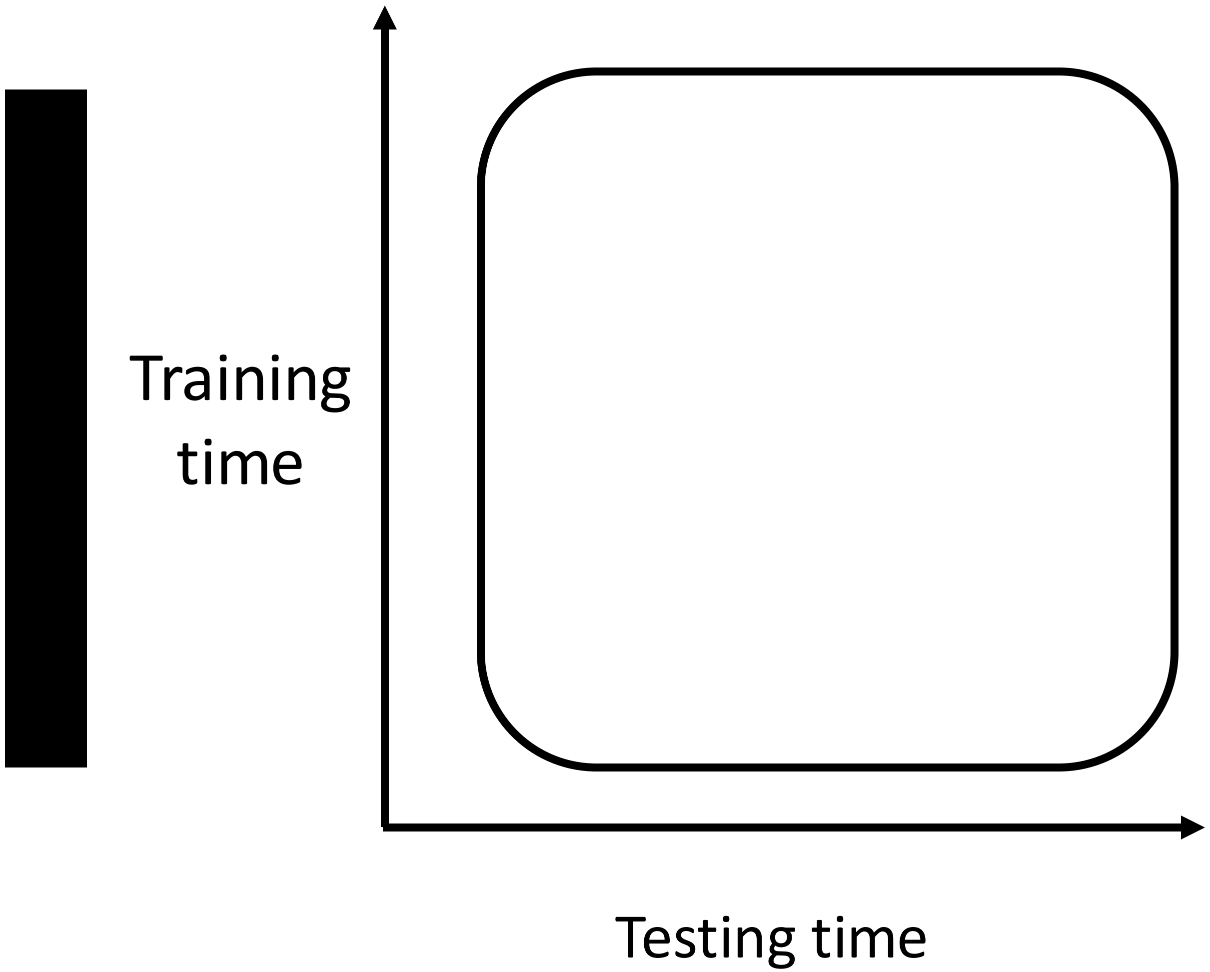
Testing time

If the encoding pattern is evolving



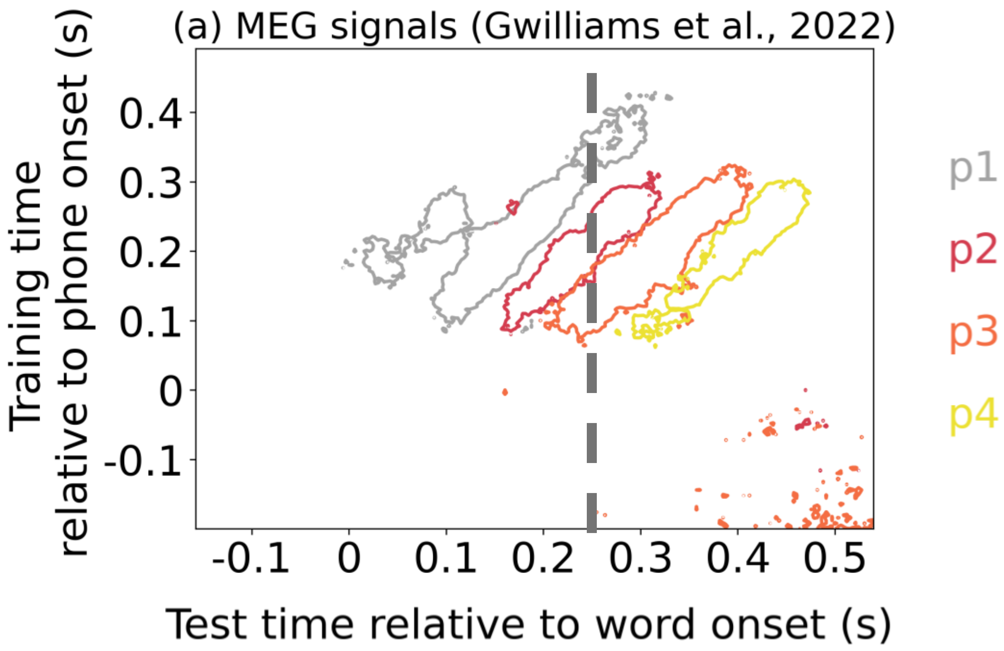
If the encoding pattern is stable

Duration of decodable window ≈ duration of individual encoding pattern



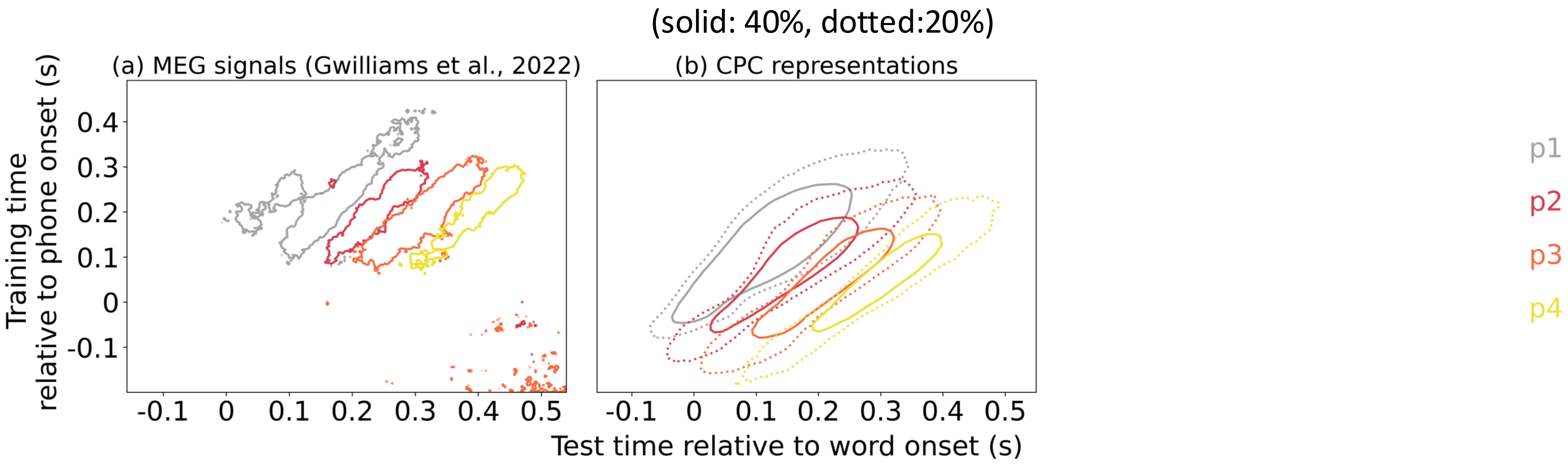
Temporal generalization contour

Dynamic encoding in neural signals



- The encoding pattern of each phone evolves over time
- The brain maintains three successive phones simultaneously

Dynamic encoding in model representations



The model exhibits similar temporal dynamics as brain signals.

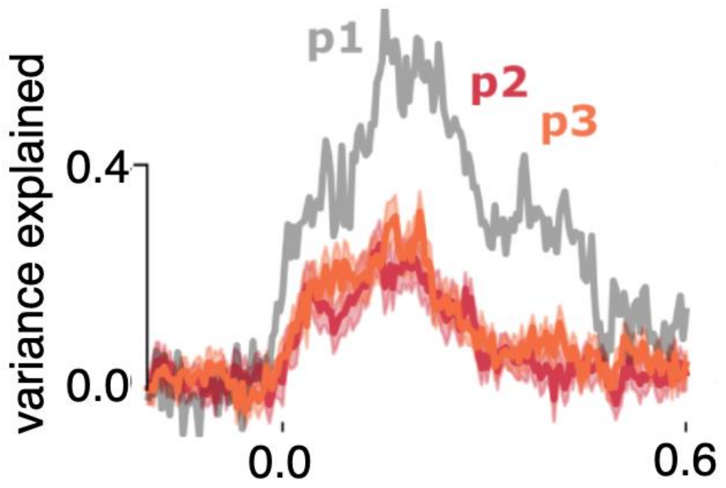
These two properties can arise without top-down information or linguistic knowledge.

Context effect

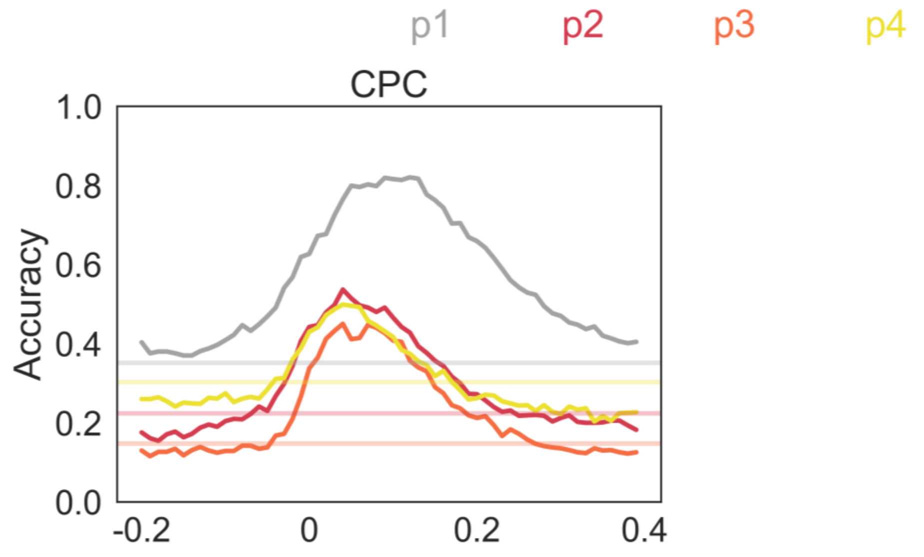
- Gwilliams et al. tested cross-position generalization
  - To more directly evaluate context-invariance, we also tested cross-context generalization
- Does the generalization effect come from acoustic similarity?
  - We compared generalization in the model against generalization with acoustic features

Cross-position generalization

Gwilliams et al. found partial generalization in brain signals



Time relative to phone onset (s)



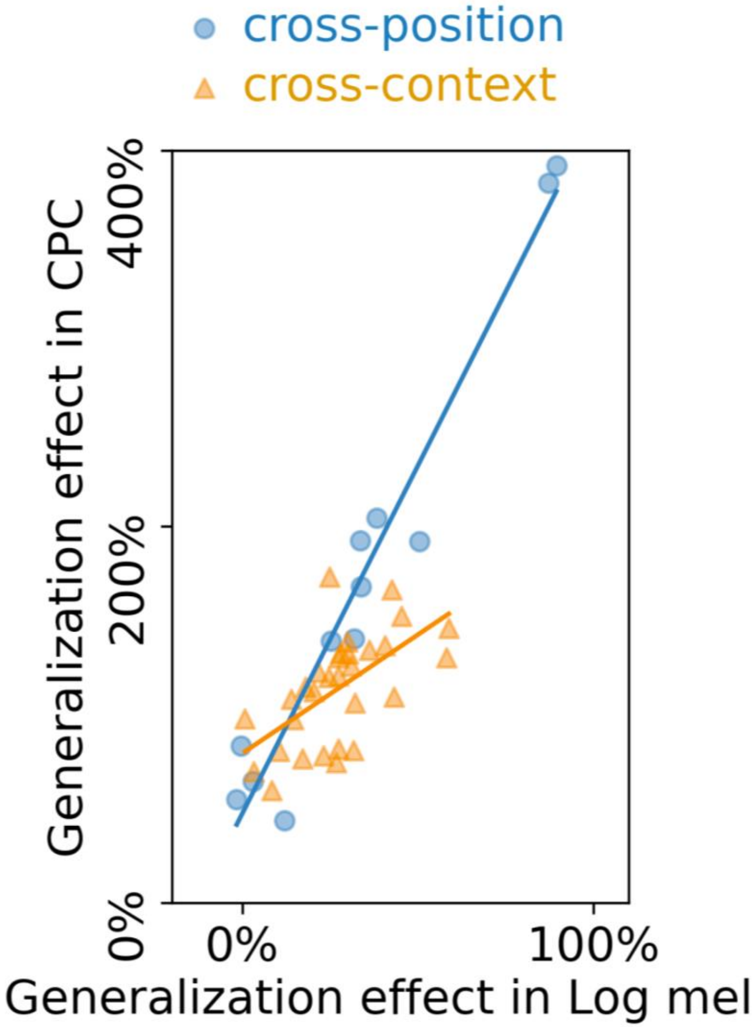
Time relative to phone onset (s)

Model representations also support incomplete generalization.

There is a small degree of cross-position generalization in acoustic features.

Similar patterns in cross-context generalizations results.

Generalization effects could depend on acoustic similarity



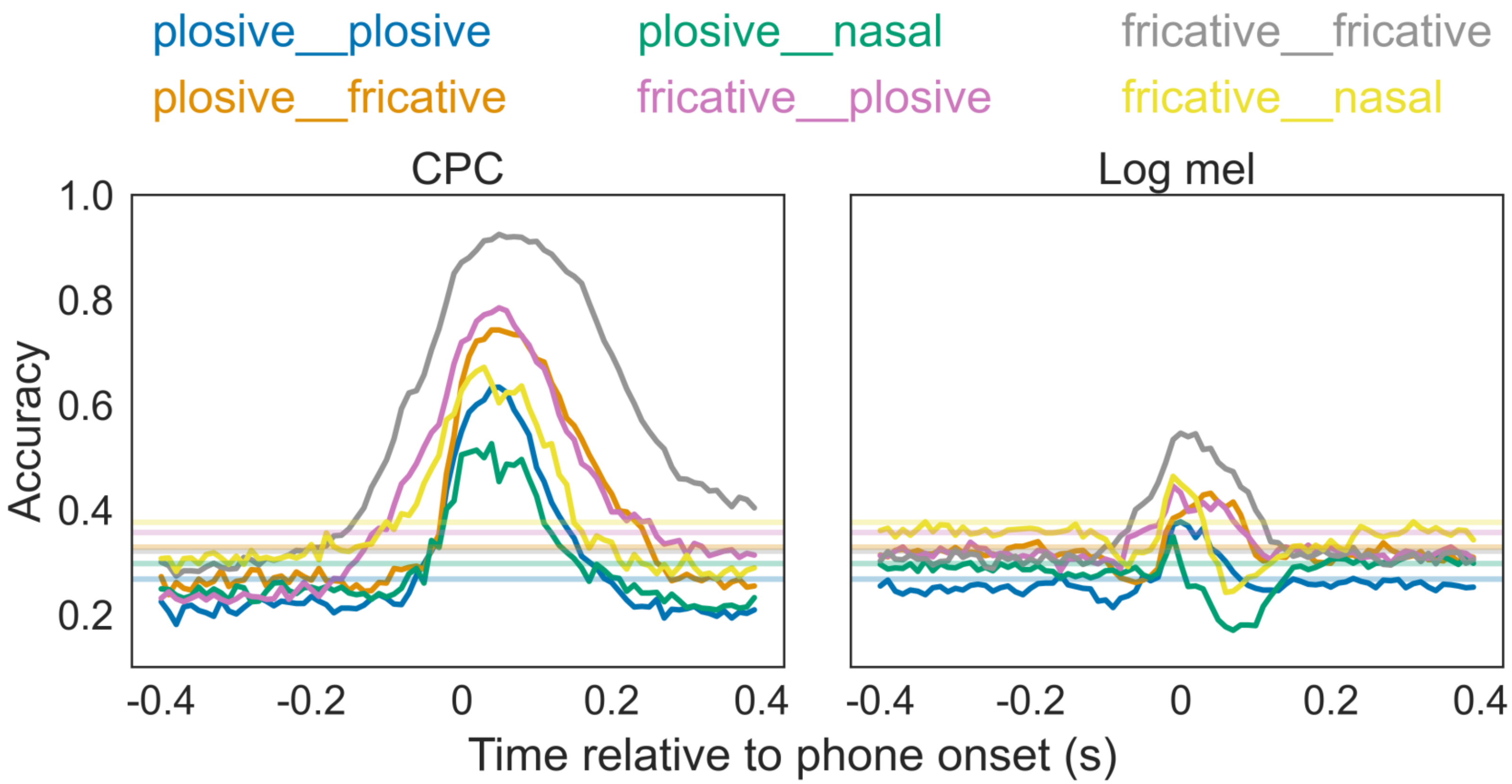
- strong positive correlation between the extent of generalization effect in model representations and in acoustic features
- generalization effects in the model depends on the acoustic similarity of the training and test contexts
- It's possible learning induces more context-invariance, but partial generalization alone does not support that

Conclusions

- We showed that a predictive learning model can simulate temporal dynamics found in neural encoding of human listeners
    - These properties can arise without top-down information or prior linguistic knowledge
- Also similar to brains, the model supports partial cross-context generalization
    - The generalization effect might be driven by acoustic similarities
    - Further studies are required to confirm the presence of context-invariant encoding

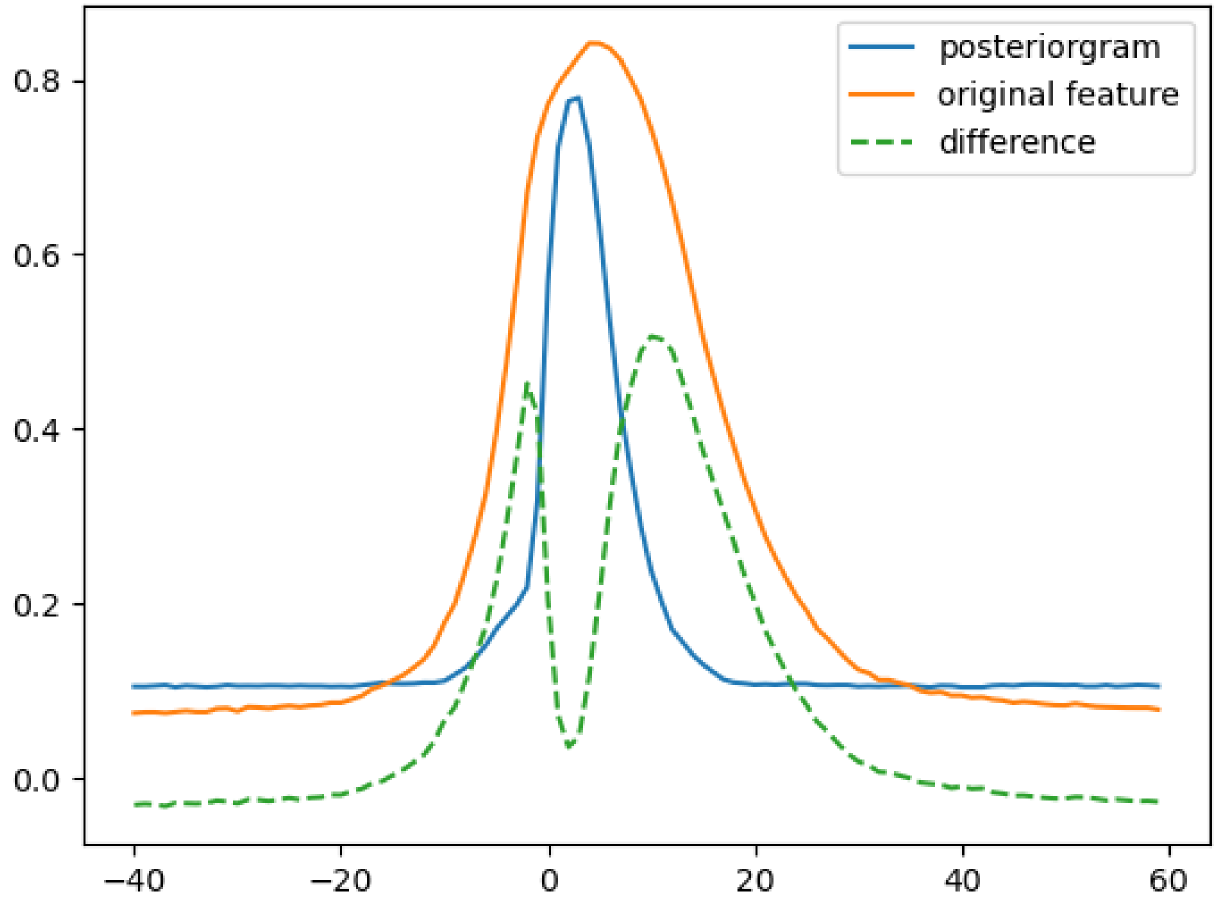
oli.liu@ed.ac.uk

Cross-context generalization



Posteriorgram baseline

Decodi
ng
accurac
y



Time relative to phone
onset

Contrastive predictive coding