Persistent and evolving encoding of phones reflects general auditory processing

Anonymous authors
Double blind review

Abstract

Speech recognition involves storing and integrating sequentially presented information. Recent work in cognitive neuroscience has identified temporal characteristics in humans' neural encoding of speech that may facilitate this process. A modeling study found similar properties in a self-supervised learning model trained on raw speech, suggesting these properties can arise without prior linguistic knowledge. In this work, we further explore the domain specificity of the same properties through testing representations of speech extracted from a model only trained on non-speech audio. The model replicated key aspects of the temporal characteristics, implying they might not be specific to speech perception, but rather features of general auditory processing.

Keywords: speech perception; predictive learning; self-supervised learning; decoding analyses

Speech recognition involves mapping continuous acoustic signals to a sequence of discrete linguistic symbols. While identifying individual phones is non-trivial in itself, extracting lexical items requires successive phones to be tracked and integrated. This means multiple phones need to be encoded simultaneously, as well as their relative order. This temporal information is crucial for differentiating between words composed of the same phonemes arranged in different orders, e.g. *tap, apt, pat*. Another challenge lies in the often ambiguous boundaries between words or morphemes, which must be resolved to partition a sequence of phones before aggregating them into meaningful units. Despite these issues, speech recognition unfolds in the human brain with little conscious effort.

Gwilliams, King, Marantz, and Poeppel (2022) analyzed MEG recordings of humans listening to short narratives, and identified several properties that are absent from the corresponding acoustic signals. Consistent with earlier findings (Khalighinejad, Cruzatto Da Silva, & Mesgarani, 2017), they found that phone representations are maintained in the brain well beyond their presence in the acoustics. During this period, the encoding pattern of a phone evolves dynamically over time, rather than remaining static. Together, these properties could allow the brain to simultaneously maintain multiple successive phones, as well as their relative order, through jointly encoding phonetic and temporal information. A subsequent modeling study (Liu, Tang, Feldman, & Goldwater, 2024) found that a predictive learning model trained on unlabeled raw speech — contrastive predictive coding (CPC) (Oord, Li, & Vinyals, 2018) — exhibits similar characteristics. Their results suggest that these properties can arise without top-down supervision from linguistic units.

In this work, we further evaluate the domain specificity of these temporal dynamics. That is, we ask whether these patterns reflect general features of auditory processing. To this end, we tested an implementation of CPC that was only exposed to non-speech audio during training. We fed the same speech stimuli into CPC-audio and CPC-speech and analyzed corresponding representations to compare the temporal dynamics of phonetic encoding in the two models.

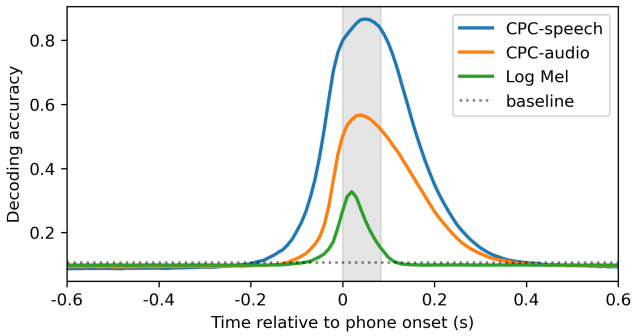


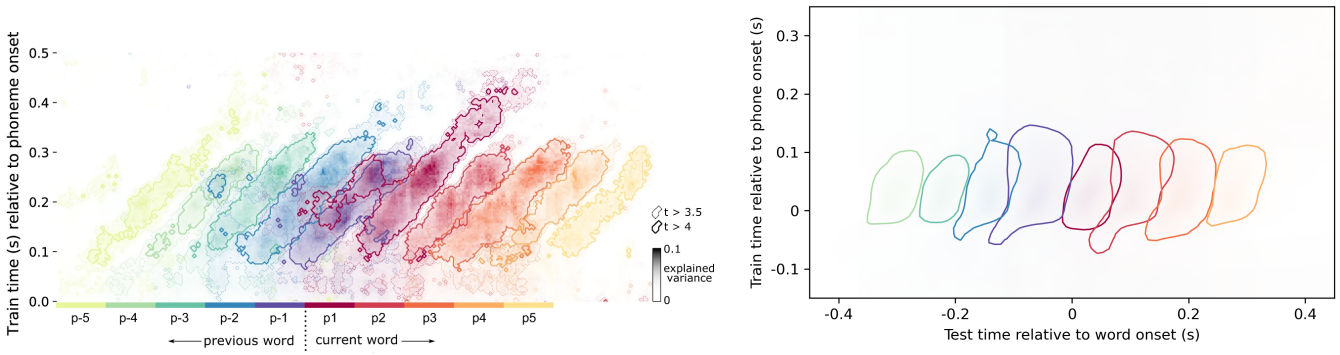
Figure 1: Average decoding accuracy for phoneme category given model representations or acoustic features for each frame within a 1200ms sliding window centered on phone onset. The shaded area shows the average duration of a phone.

Models Both models were LSTM-based recurrent neural networks trained with the objective of predicting upcoming acoustics, i.e. the next 120ms of input. The CPC-audio model we used was trained on 500 hours of animal vocalization and environmental sounds (Poli, Schatz, Dupoux, & Lavechin, 2024), and the CPC-speech model we used was the same implementation (Nguyen et al., 2020) considered in Liu et al. (2024), which was trained on 6000 hours of English audiobooks.

Decoding analyses Following previous work, we used time-resolved decoding to characterize the time course of phonetic encoding in model representations. This involves training a separate decoder (multinomial logistic regression), which takes in a single representation and predicts the phoneme label, for each time step (latency) with respect to the onset of a phone. To map out the **decodable window** of a phone, we tested each trained decoder on the same time step as it was trained on, with a held-out set of representations. To evaluate how long each encoding pattern persists within the decodable window, we performed **temporal generalization** analysis, where we tested each trained decoder on all time steps. The speech stimuli we used to train and test the decoders were drawn from a 5.4-hour subset of an English audiobook corpora (Panayotov, Chen, Povey, & Khudanpur, 2015).

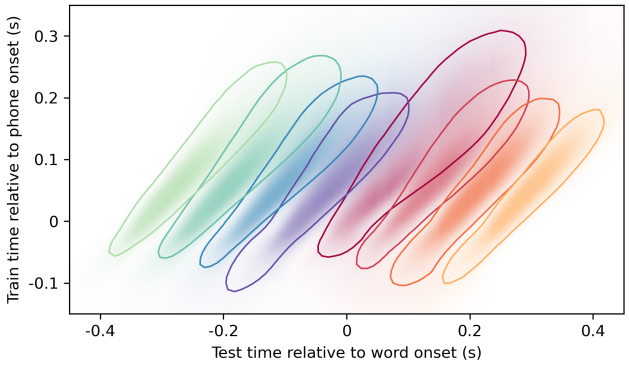
Results

Decodable window As shown in Fig 1, the phoneme category of a phone remains decodable in CPC-audio representations up until 400ms after phone onset, which is almost the

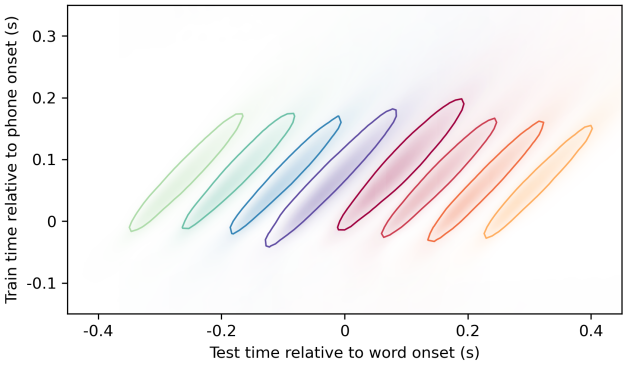


(a) MEG recordings (figure from Gwilliams et al. (2022))

(b) Log Mel spectrogram (results from Liu et al. (2024))



(c) CPC-speech (results from Liu et al. (2024))



(d) CPC-audio (our new results)

Figure 2: Temporal generalization (TG) results for (a) MEG signals, (b) log Mel acoustic features, (c) CPC-speech, and (e) CPC-audio. Contours represent decoding accuracy, within which higher color intensity indicates higher accuracy. Contours correspond to accuracy thresholds of 0.4 for CPC-speech and CPC-audio, or 0.2 for log Mel spectrogram features.

same as how long phonetic information is maintained in CPC-speech. Although CPC-audio yields a lower decoding accuracy than CPC-speech, it is still much higher than acoustic features (log Mel spectrograms), suggesting the model has acquired features useful for distinguishing phonemes despite only being trained for general auditory processing without any exposure to speech. However, CPC-audio does differ from CPC-speech in its limited support for predicting upcoming phones before their onset. This implies the predictive effect mostly comes from statistical regularities that are specific to speech.

Temporal generalization The temporal generalization (TG) analysis produces a matrix of dimension (training time step) $\times$ (testing time step), which we visualize as a contour plot. To facilitate direct comparison with Gwilliams et al.'s results, we followed them in grouping phone tokens into sets according to their position in the word, and applying TG analysis separately to each set. These sets correspond to phones in the first through fourth positions within a word (p1–p4) and phones in the final through fourth-to-last positions (p-1–p-4), yielding 8 TG matrices, visualized as 8 contours in a single plot. Each contour is shifted by the average onset latency of its corresponding phone onset relative to the word onset, ensuring that, as a whole, the plot reflects the progression of each phonetic encoding while a word unfolds.

From Fig 2c-d, we can see slanted, elongated contours in the temporal generalization results of both CPC-audio and CPC-speech, that resemble those Gwilliams et al. found in MEG recordings (Fig 2a). For all phone positions, the diagonal axis of the contour, which signifies the period that a phone is decodable from the representations, is much longer than any of its horizontal widths, which show the duration that each neural pattern persists. In contrast, the TG contours of log Mel features show stabler encoding patterns, indicating that the dynamic encoding in both models arises through learning.

At the same time, the TG results of CPC-speech exhibited subtle patterns not present in CPC-audio, which reflect speech-specific characteristics. These include incrementally downward shift of successive contours along the vertical axis, and the significantly larger contour corresponding to p1. These point to increasingly early prediction of later phones in a word and longer retention of word-initial phones, respectively, both of which were also identified in human speech processing (Saffran, Aslin, & Newport, 1996; Gwilliams et al., 2022). The absence of these characteristics in CPC-audio's TG results suggests that these properties arose specifically through learning from speech.

Despite these differences, we overall found persistent and dynamically evolving encoding of phones in a predictive learning model trained on non-speech audio, suggesting these properties might be more domain-general than previously thought.

References

Gwilliams, L., King, J.-R., Marantz, A., & Poeppel, D. (2022). Neural dynamics of phoneme sequences reveal position-invariant code for content and order. *Nature communications*, *13*(1), 6606.

Khalighinejad, B., Cruzatto Da Silva, G., & Mesgarani, N. (2017, February). Dynamic encoding of acoustic features in neural responses to continuous speech. *The Journal of Neuroscience*, *37*(8), 2176–2185.

Liu, O. D., Tang, H., Feldman, N. H., & Goldwater, S. (2024). A predictive learning model can simulate temporal dynamics and context effects found in neural representations of continuous speech. In *Proceedings of the 46th annual conference of the cognitive science society.*

Nguyen, T. A., de Seyssel, M., Rozé, P., Rivière, M., Kharitonov, E., Baevski, A., ... Dupoux, E. (2020, December). The Zero Resource Speech Benchmark 2021: Metrics and baselines for unsupervised spoken language modeling. *arXiv:2011.11588 [cs, eess]*. (arXiv: 2011.11588)

Oord, A. v. d., Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: An asr corpus based on public domain audio books. In *2015 ieee international conference on acoustics, speech and signal processing (icassp)* (p. 5206-5210). doi: 10.1109/ICASSP.2015.7178964

Poli, M., Schatz, T., Dupoux, E., & Lavechin, M. (2024). Modeling the initial state of early phonetic learning in infants. *Language Development Research*, *5*(1).

Saffran, J., Aslin, R., & Newport, E. (1996). Statistical learning in 8-month-old infants. *Science*, *274*, 1926–1928.