Temporal Straightening as a Predictive Mechanism in Human Language Processing

Jiaming Xu (jiaming.xu@utexas.edu)
Center for Perceptual Systems, 180 E Dean Keeton St
Austin, Texas 78712, USA

Jerry Tang (jerrytang@utexas.edu)
Department of Computer Science, 2317 Speedway
Austin, Texas 78712, USA

Alexander G. Huth (huth@cs.utexas.edu)
Department of Computer Science, Department of Neuroscience,
2317 Speedway, Austin, Texas 78712, USA

Robbe L.T. Goris (robbe.goris@utexas.edu)
Center for Perceptual Systems, 180 E Dean Keeton St
Austin, Texas 78712, USA

Abstract

Predicting what comes next is central to how humans process language, and to how artificial language systems, trained on next-word prediction, learn flexible representations that support diverse tasks. Despite the importance of prediction in both systems, the neural mechanisms underlying prediction in the human brain remain poorly understood. Inspired by the temporal straightening hypothesis from vision neuroscience, we investigated predictive representations in language processing from a geometric perspective. This hypothesis proposes that the brain transforms complex inputs to follow straighter temporal trajectories in representational space, enabling prediction through linear extrapolation. Here, we tested whether a similar principle applies to the human language system. Using fMRI data from subjects listening to a natural spoken narrative, we estimated representational timescale as a proxy for trajectory straightness across regions in the language processing hierarchy. We found that timescale increased in higher-order regions, indicating that neural trajectories become progressively straighter along the hierarchy. These findings offer a new perspective on predictive mechanisms in language, suggesting that temporal straightening may serve as a general organizing principle across different systems.

Keywords: prediction; language processing; fMRI

The ability to predict upcoming phrases and sentences is central to smooth and effective communication. In natural language processing, large language models trained on next-word prediction develop general-purpose representations that support diverse downstream tasks. These findings have led many to propose that prediction is a core feature of language representations, and that the general principle of prediction offers explanatory power for language processing (Caucheteux & King, 2022; Schrimpf et al., 2021). However, it remains challenging to develop falsifiable hypotheses about the neural computation of prediction, and we still lack a clear understanding of how predictive objectives shape these representations.

Recently, in vision, Hénaff et al. (2019) introduced a theoretical framework that connects the goal of prediction to the geometry of neural population representations: the Temporal Straightening hypothesis. Visual prediction is difficult because natural input to the retinas follows complex, irregular temporal trajectories. This hypothesis suggests that our brain transforms recent inputs to follow straighter trajectories, facilitating prediction through linear extrapolation (Fig. 1a). Since its introduction, the hypothesis has been supported by empirical evidence from psychophysical, physiological, and computational studies of the visual system.

Remarkably, temporal straightening has also been observed in a fundamentally different kind of system: autoregressive transformer models (i.e., the GPT model family). In this context, straightness was quantified by analyzing the temporal trajectory of language representations in the model's activation space, where each point corresponds to the activation pattern across units in a given layer in response to a word. For a sequence of words, curvature was defined as the angle between adjacent activation vectors, and average curvature across a sentence was computed to assess changes across layers. In trained models, curvature systematically decreased from the first to the middle layers of the network (Hosseini & Fedorenko, 2023).

Motivated by this finding, we aimed to evaluate representational straightening in human brain responses to language. We hypothesized that speech and language representations in the brain similarly follow the principle of temporal straightening, and that this effect becomes more prominent along the processing hierarchy.

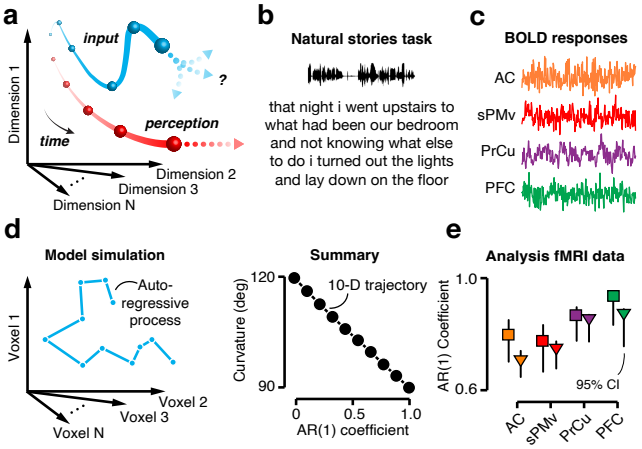

Figure 1: A new approach to estimate population trajectory curvature from fMRI data. a. Temporal straightening hypothesis. **b.** Task paradigm: fMRI responses were recorded while two subjects listened to a narrative story. **c.** Example voxel activity from four hierarchical language regions. **d.** We simulated high-dimensional population trajectories using first-order autoregressive (AR(1)) processes. The average trajectory curvature was determined by the timescale of the process (i.e., the AR coefficient). The dimensionality of the trajectory had no effect. **e.** We fit an AR(1) to averaged voxel-activity traces for two subjects. Across the language-processing hierarchy, the coefficient estimate tended to increase, corresponding to straighter trajectories. CI based on nonparametric bootstrapping.

Results

To test our hypothesis, we analyzed previously collected fMRI data from two human subjects who listened to a spoken narrative from *The Moth Radio Hour*. The story was around 10 minutes long and was repeated 10 times. We extracted data from four regions of interest (ROI) that span the language processing hierarchy: auditory cortex (AC), superior premotor ventral area (sPMv), precuneus (PrCu), and prefrontal cortex (PFC) (LeBel et al., 2023) (Fig. 1b–c). A direct approach to measuring neural curvature in fMRI data treats each voxel within a ROI as a separate axis in representational space, computes

angles between successive time points, and averages them. While effective with low-noise data, the inherent noise in fMRI renders such direct measures unreliable. To address this, we considered that temporal straightness may be related to representational timescales, which have been studied extensively in fMRI (Lerner, Honey, Silbert, & Hasson, 2011; Jain et al., 2020). When speech is processed, auditory information first reaches the primary auditory cortex (AC), which operates on sub-second timescales and encodes rapidly changing acoustic and word-level features. This information is then passed through a hierarchy of regions—including parts of the superior temporal and frontal lobes—where it is integrated over increasingly longer timescales to construct meaning. Consequently, regions with shorter timescales are expected to exhibit more dynamic changes in representational space, yielding temporal trajectories with higher overall curvature.

To verify our intuition, we used first-order autoregressive (AR(1)) processes to simulate high-dimensional population trajectories. We varied the AR coefficient, which controls the timescale of the process, and measured the resulting average trajectory curvature. This revealed a systematic relationship: higher AR(1) coefficients yielded lower average population curvature. Importantly, the dimensionality of the trajectory had no effect on this relationship. These results confirm that the AR(1) coefficient can serve as an indirect indicator of population trajectory curvature in a given brain area (Fig. 1d).

Guided by our simulation results demonstrating that AR(1) coefficients can serve as proxies for trajectory curvature and timescale, we applied the same modeling approach to our fMRI dataset to examine timescale variation across the language processing hierarchy. For each ROI, we first averaged the fMRI responses across repetitions of the narrative stimulus to reduce trial-by-trial variability. We then averaged across voxels within each ROI to obtain a single, denoised time course per region. Fitting AR(1) models to these time courses yielded coefficient estimates that reflect the intrinsic temporal integration of each region. Consistent with our hypothesis, AR(1) coefficients tended to increase along the language-processing hierarchy, indicating longer timescales and progressively straighter neural trajectories in higher-order regions. We estimated 95% CI by bootstrapping: repeatedly sampling 10 runs with replacement and applying the same averaging and AR(1) fitting procedure as in the main analysis to each sample. For both subjects, the coefficient estimates of AC are significantly different from sPMV, precuneous and prefrontal ($p < 0.05$, Fig. 1e).

While these results are consistent with our hypothesis, estimation of AR(1) coefficients from BOLD signals is susceptible to noise, and importantly, the magnitude of this noise varies across brain regions. In particular, earlier areas (e.g., AC and sPMv) tend to have higher signal-to-noise ratios (SNRs) than later areas (e.g., PrCu and PFC), raising the possibility that the observed increase in timescales across the hierarchy could be partly driven by differences in data quality. To address this, we performed a control analysis in which we compared AR(1) coefficients across ROIs while holding SNR constant. Specifically, for each ROI, we computed the SNR of each voxel and grouped the voxels into three SNR bins, using the same bin edges across all ROIs. Within each bin, we averaged the trial-averaged responses across voxels, and fit AR(1) models to these bin-wise average time courses. We found that the original pattern of increasing AR(1) coefficients across the language hierarchy held consistently across all SNR bins, suggesting that the observed timescale differences are not driven by variation in SNR (Fig. 2).

Together, our simulation and model-based fMRI analysis suggest that neural population trajectories become progressively straighter along the language hierarchy in the human brain.

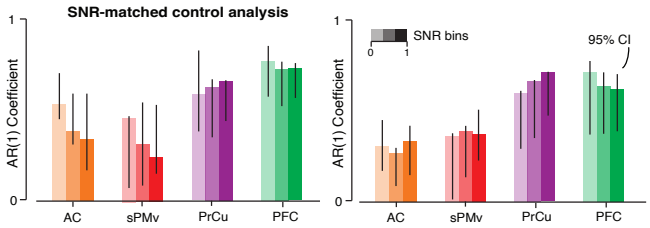

Figure 2: Controlling for SNR in timescale estimation. AR(1) coefficients estimated after binning voxels within each ROI by SNR, shown separately for each subject. The original pattern of increasing timescales across the language hierarchy remained consistent across all SNR bins, indicating that the observed effect is not driven by variation in SNR. CI based on nonparametric bootstrapping.

Discussion

Inspired by the temporal straightening hypothesis in vision, we tested whether a similar principle governs neural computations in the human language system. We showed that representational timescale, estimated via AR(1) coefficients, reflects trajectory straightness and can be recovered from noisy fMRI data. Timescale increased systematically along the language hierarchy, suggesting that higher-order regions maintain straighter trajectories to support prediction. Extending temporal straightening from vision to language points to a shared computational principle across brain systems—and more broadly, across systems optimized for prediction. Supporting this, we found striking convergence between our results and prior findings in autoregressive transformer models, suggesting common predictive strategies in biological and artificial systems. Our work also introduces representational geometry as a novel lens on predictive mechanisms in language. While our framework is grounded in geometric principles, we used timescale as an indirect proxy for straightness. Future work will clarify this relationship, helping to situate our findings within cognitive neuroscience and bridge temporal and geometric accounts of prediction.

References

Caucheteux, C., & King, J.-R. (2022). Brains and algorithms partially converge in natural language processing. *Communications biology*, *5*(1), 134.

Hosseini, E., & Fedorenko, E. (2023). Large language models implicitly learn to straighten neural sentence trajectories to construct a predictive representation of natural language. *Advances in Neural Information Processing Systems*, *36*, 43918-43930.

Hénaff, O. J., Goris, R. L., & Simoncelli, E. P. (2019). Perceptual straightening of natural videos. *Nature neuroscience*, *22*(6), 984-991.

Jain, S., Vo, V., Mahto, S., LeBel, A., Turek, J. S., & Huth, A. (2020). Interpretable multi-timescale models for predicting fmri responses to continuous natural speech. *Advances in Neural Information Processing Systems*, *33*, 13738-13749.

LeBel, A., Wagner, L., Jain, S., Adhikari-Desai, A., Gupta, B., Morgenthal, A., ... Huth, A. G. (2023). A natural language fmri dataset for voxelwise encoding models. *Scientific Data*, *10*(1), 555.

Lerner, Y., Honey, C. J., Silbert, L. J., & Hasson, U. (2011). Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *Journal of neuroscience*, *31*(8), 2906-2915.

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., ... Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, *118*(45).