

Extending a Formal Model of Predictive Coding to Human Spoken Word Recognition

Bryce Rogers

bryce.e.rogers@gmail.com

School of Chemical, Biological, and Environmental Engineering, Oregon State University, Corvallis, OR 97331, U.S.A.

Monica Yi-Chen Li

monica.yc.li@gmail.com

Department of Psychological Sciences, University of Connecticut, Storrs, CT 06269, U.S.A.

Thomas Hannagan

thomas.hannagan@stellantis.com

Department of Psychological Sciences, University of Connecticut, Storrs, CT 06269, U.S.A.; and Stellantis Group, Paris 78300, France

James S. Magnuson

james.magnuson@uconn.edu

Department of Psychological Sciences, University of Connecticut, Storrs, CT 06269, U.S.A.; BCBL, Basque Center on Cognition Brain and Language, 10009 Donostia-San Sebastián, Spain; and Ikerbasque, Basque Foundation for Science, 48009 Bilbao, Spain

Kevin S. Brown

kevin.brown@oregonstate.edu

School of Chemical, Biological, and Environmental Engineering and Department of Pharmaceutical Sciences, Oregon State University, Corvallis, OR 97331, U.S.A.

There is a general consensus in theories of human speech recognition that humans engage in predictive processing during online speech processing. There are also claims that predictive processing is indicative of the operation of a predictive coding (PC) mechanism. PC is a generative, hierarchical feedback framework where feedback signals consist of input predictions, while feedforward signals consist primarily of prediction errors (PE). Some researchers have taken decreased neural signals when inputs conform to expectations as evidence for PC and claim that other possible explanatory frameworks (e.g., interactive activation) are incompatible with reported reductions in PE. However, these claims have been advanced using narrow-scope computational implementations of PC without known abilities to adequately (i.e., plausibly)

simulate broader human spoken word recognition (SWR) behavioral phenomena. Here, we present the first known neurally and behaviorally adequate, mathematically formal neural network PC model of time-dependent human SWR. After demonstrating that the new model is able to emulate a fundamental, empirical signature of human SWR behavior (time course of lexical activation and competition), we compare model dynamics to neural performance targets that have been touted as hallmarks of PC in SWR. While the new model readily exhibits predictive processing (anticipatory activation of phonemes consistent with lexical knowledge) and reduced neural activity when inputs match expectations (a necessary component of PC) more subtle patterns of neural activity under priming and noise conditions, previously proposed as diagnostic of PC do not emerge appreciably. This suggests that such patterns may not be hallmarks of PC. We discuss implications for PC-based theories of SWR.

1 Introduction

Prediction is a key aspect of human language processing. For example, readers and listeners are sensitive to cloze completion probabilities (Taylor, 1953) of words in sentences (e.g., “It was a windy day, so the boy went outside to fly his . . .”: kite [high predictability] versus airplane [low predictability]; DeLong et al., 2005), and of phonemes in words (e.g., formula versus nonword “formubo,” which violates lexical prediction at /b/; Gagnepain et al., 2012). Humans also process language using more complex predictions, for example, predictions based on an understanding of the thematic roles of nouns (e.g., Altmann & Kamide, 1999; Kukona et al., 2011). Increased neural activity has long been observed in response to unexpected linguistic input (cf. Kutas & Federmeier, 2011, who note that the N400 component of the event-related potential is proportional to word surprisal). However, over the past decade, the theory that this increase of neural activity in response to unexpected inputs is due specifically to prediction error (PE) has garnered much support and attention.

In the domain of human speech recognition, Davis and colleagues (Blank & Davis, 2016; Davis & Sohoglu, 2020; Gagnepain et al., 2012; Sohoglu & Davis, 2016) have been particularly strong proponents of the view that reduced neural activity when expectations are confirmed is evidence for reduced PE, which they posit supports a computational framework known as predictive coding (PC: Rao & Ballard, 1999; Spratling, 2008; Friston, 2005) but contradicts another framework, interactive activation (IA, as in the TRACE model of human speech perception and spoken word recognition; McClelland & Elman, 1986). In IA, there is vertical interaction between hierarchically organized layers (feedforward and feedback connections), as well as horizontal interaction within layers (via lateral

inhibition and additional connectivity in the TRACE model). Davis and colleagues argue that IA predicts that neural activity should increase when inputs are expected, because the dynamics of IA “sharpen signals,” while PC uniquely reduces activity when inputs are expected.

There are reasons to be skeptical that these characteristics can reliably differentiate between PC and IA frameworks. First, activity reduction when inputs conform to expectations is often posited to constitute evidence for PC without linkage to formal models of predictive coding. While Blank & Davis (2016) conducted simulations meant to pit PC against IA, their implemented models were conceptually based on these two frameworks but were not formally equivalent to them (Gaston et al., n.d.). Second, as we discuss below, formal specifications of PC (Rao & Ballard, 1999; Spratling, 2008) may be less different from IA than intuition initially suggests, given, for example, that the conventionally parameterized IA TRACE model exhibits signal reduction when expectations (lexically predicted phonemes) are confirmed (Luthra et al., 2021), contra predictions (Gagnepain et al., 2012).

The relation of predictive processing to different computational frameworks is one of the most interesting theoretical challenges in the field of human spoken word recognition (SWR). It connects to the long-standing debate on whether lexical-level processing has a direct influence on sublexical processing (Elman & McClelland, 1988; Magnuson et al., 2003a, 2003b, 2018; McQueen, 2003; McQueen et al., 2006, 2009; Norris et al., 2000, 2003, 2016; Pitt & McQueen, 1998; Samuel & Pitt, 2003; Magnuson et al., 2024). PC would seem to imply such direct influence, but Norris et al. (2016) claim that the feedback in IA and PC would be qualitatively distinct. Thus, exploring whether the IA framework is qualitatively distinct from PC will have implications for multiple theoretical concerns, but it will require comparable implementations of each framework.

Our focus in this study will be strong claims in the literature regarding the relative compatibility or incompatibility of different frameworks with extant data. Specifically, Blank and Davis (2016) asserted that IA is incompatible with behavioral and neurobiological data that they further claim are only consistent with PC. However, these claims are founded on simulations with incomplete models that simulate only selected aspects of PC and IA (e.g., simple, quasi-hierarchical models that employ either additive or subtractive feedback, where the former is meant to stand in for IA and the latter for PC, though neither serves as a totally plausible implementation of its respective framework). To date, no formal, neurally and behaviorally adequate, neural network PC model has been implemented for speech that would permit evaluation of SWR task response properties that could differentiate the PC model from competitor frameworks.¹

¹ Note that Yildiz et al. (2013) created a PC model for speech. While it operated on real speech, their model was limited to recognizing the names of the digits 0 to 9, and was not

There has been a proliferation of narrow-scope models in SWR: models that are used for proof-of-concept demonstrations of principles without being tested for their ability to reproduce fundamental human behavior in the domain. The ability to simulate human behavior, also known as *behavioral adequacy*, is (alongside neural adequacy) one of the five primary criteria for model realism and utility proposed by Magnuson et al. (2020), extending McClelland and Elman's (1986) distinction between behavioral and computational adequacy.

Indeed, models intended to advance understanding of perception and cognition in biological organisms that fail to emulate basic, empirically observed behaviors are necessarily limited in their ability to provide reliable domain insight, much less advance theoretical understanding. Consider the mathematical model used by Gagnepain et al. (2012) and the 24-word sharpened signal and prediction error models of Blank & Davis (2016). In addition to architectural limitations, none have, to our knowledge, been assessed for behavioral adequacy (e.g., Do they exhibit robust word recognition? Do they emulate the time course of phonological competition observed in human SWR?). Both the Gagnepain et al. (2012) and Blank and Davis (2016) models were first created to test for specific experimental patterns they had observed in SWR and hypothesized to be signatures of PE and thus PC. However, we hold that a model must also be assessed for its ability to simulate fundamental, benchmark human SWR data patterns such as the time course of phonological competition (Alloppenna et al., 1998) to be considered viable.

We address the lack of such models by implementing, to our knowledge, the first formal neural network PC model of human SWR that operates on over-time speech-like input and exhibits robust behavioral adequacy. This model, as well as its static input precursor presented in section 2.1 (created to demonstrate that we can extend the PC framework to perform a classification task), are direct extensions of the formal variant of PC developed by Rao and Ballard (1999). The models we consider in this letter are shown in Figure 1. At top is the original Rao and Ballard (1999) model (sPC, for static predictive coding). In the middle is our modification of the sPC model (called sPCC, for static predictive coding classifier) that extends the sPC model to perform a classification task; we describe this extension in detail and verify that sPCC can classify words when input as static images (pseudospectra). The bottom model (rPCC: recurrent predictive coding classifier) is our sequential, over-time SWR model; it is a modification of the dynamic (Kalman-filter based) model of Rao (1999), with the classification layer of

compared to human SWR. Another notable PC approach to speech is the active listening model of Friston et al. (2021). This is an innovative generative account with several appealing features, but it differs substantially from the aims of this work. In particular, it is not a learning model and is oriented more toward inference with a specified generative architecture. It addresses speech recognition at a different level of analysis from ours, making direct comparison difficult. For these reasons, we do not consider it further here.

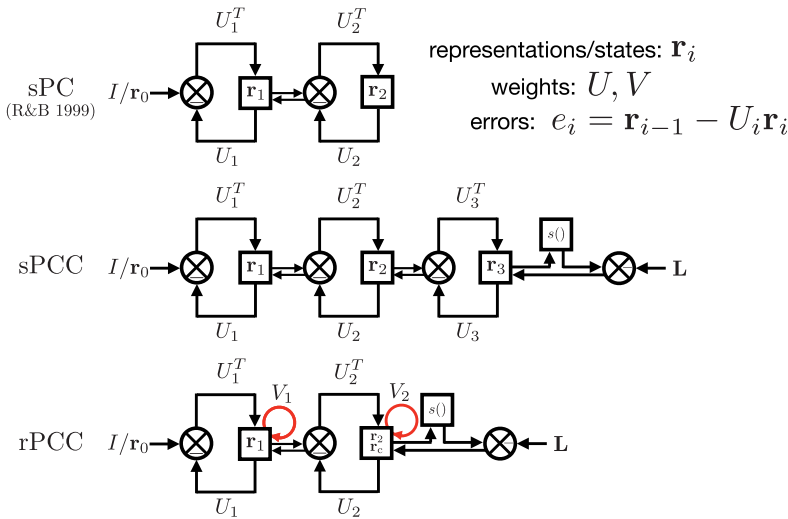


Figure 1: The Rao and Ballard (1999) formal predictive coding (PC) model (top) and extensions to create predictive coding classifiers (PCCs). Here, we depict three models, beginning with the static PC (sPC) model of Rao and Ballard (1999) (top). The static PC classifier (sPCC) model adds a final localist node layer ($\mathbf{r}_3$), with one node per label (L), to adapt the model to classification tasks (static and simplified image-like representations of speech). Finally, the recurrent PC classifier (rPCC) model extends sPCC (though it contains one fewer layer) to process sequential inputs by adding recursive weights (red arrows). We note that the rPCC model admits an additional error term not shown here, but shown in Figure A.1, in the appendix and discussed in detail in the text. Circled crosses with a single minus sign are subtraction blocks for the quantities that feed into them.

sPCC and a set of context units ($\mathbf{r}_c$) that transmit loss information during training. We note that while a similar model was proposed by Rao (1999), the dynamics (V_1 and V_2 connections) were turned off in that implementation, making the model equivalent to a static model.

Once we establish that the rPCC model can perform word recognition (i.e., given sublexical inputs, that the model can settle on the correct lexical output node), we assess the model's ability to simulate other fundamental aspects of human SWR (in particular, the time course of phonological competition, such as the complex cohort and rhyme effects observed by Allopenna et al., 1998). Next, we perform a critical test of the model: simulating a paradigm used by Gagnepain et al. (2012) where lexically based predictions (e.g., given the input "formu" the English lexicon strongly predicts that the next syllable will be *-la*) are either confirmed (the input is indeed formula) or violated (the input is a nonword such as "formubo"). Human subjects are exquisitely sensitive to lexically based probabilities in

both behavior and brain (where activations in lexically associated cortical regions are reduced when predictions are confirmed versus violated; Gagnepain et al., 2012; Sohoglu & Davis, 2020).

Luthra et al. (2021) applied TRACE to this paradigm and found (contra predictions of Gagnepain et al., 2012) that TRACE in fact shows reduced activity (diminished lexical inhibition and feedback and diminished phonemic inhibition) given valid lexical input like “formula” versus a nonword like “formubo.” We present similar pairs of stimuli to our rPCC model, allowing us to begin to assess how qualitatively different IA and PC truly are. Then, since our rPCC model does not have phonemic readouts of any sort, we use a decoding technique to discover rich, emergent anticipatory representations within the model: gradiently increasing predictive activation related to upcoming phonemes and syllables. Finally, we assess whether a pattern of model activity that has been proposed as a neural hallmark of PC in human speech processing (Blank & Davis, 2016) emerges in our rPCC model.

2 Extending Predictive Coding

2.1 The Static Predictive Coding Model. The first challenge in extending PC is that formal models do not typically perform tasks like object recognition or classification. The model described by Rao and Ballard (1999), for example, builds two levels of progressively larger visual receptive fields, but only operates over small patches of images and does not “do” anything with the output (though it crucially simulates emergence of plausible receptive field properties). Extending PC to perform object classification is an interesting challenge in its own right, and we present a detailed report on classification elsewhere (Rogers et al., n.d.). We give a brief description here to set the stage for the recurrent model described in the next section.

Rao and Ballard’s (1999) original model has two hierarchical levels and is trained on small patches of natural images. The receptive fields of each processing level increase in size as the hierarchy ascends, such that a higher level (i.e., level 2) can make spatial predictions (through top-down connections) about the immediately lower level (level 1), given an input image. The original model in Rao and Ballard (1999) has only two levels; however, we can extend this to an arbitrary number of levels as follows, where the model minimizes the following cost function during training:

$$\begin{aligned}
 J = \sum_{i=1}^{n-1} & \left[\frac{1}{\sigma_{i-1}^2} (\mathbf{r}_{i-1} - f(U_i \mathbf{r}_i))^T (\mathbf{r}_{i-1} - f(U_i \mathbf{r}_i)) \right. \\
 & \left. + \frac{1}{\sigma_i^2} (\mathbf{r}_i - f(U_{i+1} \mathbf{r}_{i+1}))^T (\mathbf{r}_i - f(U_{i+1} \mathbf{r}_{i+1})) \right] \\
 & + \sum_{i=1}^n g(\mathbf{r}_i) + \sum_{i=1}^n h(U_i).
 \end{aligned} \tag{2.1}$$

In this expression, $\mathbf{r}_0$ is taken to be the input I , with the understanding that $\mathbf{r}_0$ (unlike the other $\mathbf{r}_i$) is static and never updated. f is an optional squashing function, such as $\tanh(x)$. Defining f as the identity function produces a linear model. Training adjusts the $\mathbf{r}_i$ (representation at each level) and U_i (between-level connections) in order to make a given level's representation agree with the top-down predicted representation of the level above it in the hierarchy.

To extend this approach to classification, we must add a classification layer; we will simply use a standard one-hot output layer with one node per possible label class. Then we must extend the cost in equation 2.1 (which tries to match representations at each level) by adding a binary cross-entropy cost to the J in equation 2.1,

$$J_C = -L^T \log [s(\mathbf{r}_n)], \quad (2.2)$$

in which L is a one-hot label vector and $s(\mathbf{x}) = e^{x_i} / \sum_i e^{x_i}$ (the softmax function). A signal processing diagram of the model (which we call sPCC) is shown in the middle panel of Figure 1.

Gradient updates to adjust both the $\mathbf{r}_i$ and U_i can be derived by computing derivatives of $J + J_C$ with respect to those variables. The $\mathbf{r}$ update equations, for $i = 1, \dots, n-1$, are

$$\begin{aligned} \mathbf{r}_i \rightarrow \mathbf{r}_i + \frac{k_r}{\sigma_{i-1}^2} U_i^T F_i (\mathbf{r}_{i-1} - f(U_i \mathbf{r}_i)) \\ - \frac{k_r}{\sigma_i^2} (\mathbf{r}_i - f(U_{i+1} \mathbf{r}_{i+1})) - \frac{k_r}{2} g'(\mathbf{r}_i). \end{aligned} \quad (2.3)$$

Because of J_C , the update for the top-most layer $\mathbf{r}_n$ is slightly different:

$$\begin{aligned} \mathbf{r}_n \rightarrow \mathbf{r}_n + \frac{k_r}{\sigma_{n-1}^2} U_n^T F_n (\mathbf{r}_{n-1} - f(U_n \mathbf{r}_n)) \\ - \frac{k_r}{\sigma_n^2} (s(\mathbf{r}_n) - \mathbf{L}) - \frac{k_r}{2} g'(\mathbf{r}_n). \end{aligned} \quad (2.4)$$

In all cases, F_i is a diagonal matrix. If the function f is the identity function, then F_i is an identity matrix of the appropriate size. If $f = \tanh$, then

$$(F_i)_{vv} = 1 - \tanh^2 [(U_i \mathbf{r}_i)_v]. \quad (2.5)$$

The σ terms are noise variances (which can be different for different levels), and the k are learning rates. The term g (see equation 2.1) is a prior term; for a gaussian prior with negative log likelihood, $g(\mathbf{x}) = \alpha \mathbf{x}^T \mathbf{x}$, $g'(\mathbf{x}) = 2\alpha \mathbf{x}$. For a sparse prior with negative log likelihood, $g(\mathbf{x}) = \alpha \sum_i \log(1 + x_i^2)$, $g'(\mathbf{x}) = \alpha \sum_i 2x_i / (1 + x_i^2)$.

Table 1: Feature Map of Phonemes in the rPCC Lexicon.

Features	/b/	/k/	/s/	/r/	/a/	/˘/	/i/	/u/
consonantal	1	1	1	1	0	0	0	0
sonorant	0	0	0	1	1	1	1	1
syllabic	0	0	0	0	1	1	1	1
nasal	0	0	0	0	0	0	0	0
voiced	1	0	0	1	1	1	1	1
continuant	0	0	1	1	1	1	1	1
labial	1	0	0	0	0	0	0	0
alveolar	0	0	1	1	0	0	0	0
palatal	0	0	0	0	0	0	0	0
anterior	1	0	1	0	0	0	0	0
velar	0	1	0	0	0	0	0	0
coronal	0	0	1	1	0	0	0	0
sibilant	0	0	1	0	0	0	0	0
high	0	0	0	0	0	0	1	1
mid	0	0	0	0	0	0	0	0
low	0	0	0	0	1	1	0	0
back	0	0	0	0	0	1	0	1
central	0	0	0	0	0	0	0	0
round	0	0	0	0	0	0	0	1
tense	0	0	0	0	0	0	1	1
reduced	0	0	0	0	0	0	0	0

The updates for U for $i = 1, \dots, n$ are as follows:

$$U_i \rightarrow U_i + \frac{k_U}{\sigma_{i-1}^2} F_i(\mathbf{r}_{i-1} - f(U_i \mathbf{r}_i)) \mathbf{r}_i^T - \frac{k_U}{2} h'(U_i), \tag{2.6}$$

in which the priors on U (h in equation 2.1) can take either of the forms (gaussian or sparse) that g takes, with U substituted for $\mathbf{x}$, U_{ij} for x_i , $\sum_{ij}$ for $\sum_i$, and precision parameter λ for α . Here, we use primes on g and h to indicate derivatives of the prior functions with respect to their arguments, that is, $\mathbf{r}$ for g and U for h .

We trained the sPCC model to classify the five natural images from Rao and Ballard (1999) and 212 pseudo-spectrograms of the words from the standard TRACE lexicon (McClelland & Elman, 1986).

TRACE uses “pseudo-spectral” inputs, where phonemes are represented by sets of articulatory features that ramp up to a maximum value (1) and then ramp back down to a minimum value. Phoneme centers are spaced such that the ramping up of one phoneme overlaps with the ramping down of the previous phoneme, providing a coarse analog of coarticulation. Rather than the original TRACE set of seven phonetic features with nine levels each, we used a set of 21 simple features from Magnuson, Tanenhaus et al. (2003) (see Table 1). An example pseudo-spectrogram is shown in Figure 2.

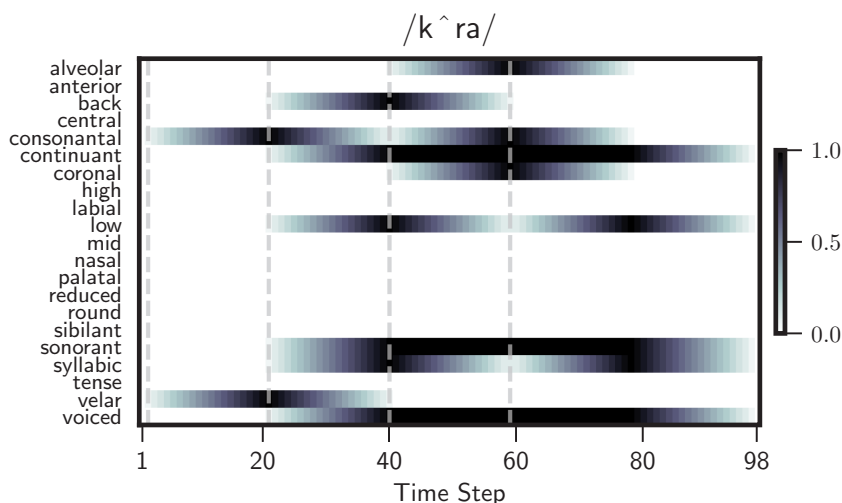


Figure 2: Example pseudospectral (TRACE-like) input used for sPCC and rPCC models. This shows the pseudospectrum for 1 of the 12 artificial CVCV words that we used to train the rPCC model. The sPCC model was trained on the 212-word TRACE lexicon using the same pseudospectral format. Phoneme onsets here appear as vertical, dashed grey lines. The artificial words all had the same length, consisting of 98 time steps, with 21-element phonetic vectors at each step.

Accuracy for the five natural images from Rao and Ballard (1999) reached (and remained at) 100% after approximately 300 epochs. The classification accuracy for the 212 pseudospectra gradually increased and plateaued at around 80% over the course of 1,000 training epochs. Full details on parameter values, training regimes, and so on can be found elsewhere (Rogers et al., n.d.); these two simulations were summarized here to establish that our classification framework is valid and can be used in the recurrent model. We did not perform a comprehensive search for parameters that could increase accuracy as our interest is in developing a model that can process over-time signals, to which we turn now.

2.2 Recurrent Predictive Coding for Over-Time Signals. Extending PC to a recurrent form is not trivial, but it is necessary for developing a plausible PC-based model of human SWR. As we will see in this section, it is possible to extend sPCC to a recurrent version that achieves high classification accuracy and demonstrates human-like phonological competition over time.

Rao and Ballard (1997) and Rao (1999) discuss in detail how one might extend their basic PC approach (Rao & Ballard, 1999) to a recurrent version

(though they did not in fact test the recurrent model). Their formulation casts the model as a Kalman filter (Kalman, 1960). The major difference between the static model (sPCC) and the recurrent model is that the states $\mathbf{r}$ now have their own internal dynamics (encoded in a matrix V), as opposed to only updating when new data arrive. Because of this, we have multiple estimates of the model parameters (we present full details in the next section): $\bar{\mathbf{r}}$ and $\bar{U}$ denote updates due to the internal model dynamics, and $\hat{\mathbf{r}}$ and $\hat{U}$ represent the blended state estimate that combines model predictions with input data. The state estimate provided by the model is thus a blend of input from the world (encoded in U) and an internal model of the world (encoded in V). Our recurrent PCC (rPCC) model also includes an additional localist output layer for classification, similar to the static PCC (sPCC) model described previously.

2.3 rPCC Model Architecture and Dynamics. The architecture of the rPCC model is illustrated in the bottom panel of Figure 1. Our rPCC model consists of a hidden layer ($\mathbf{r}_1$) and an output layer ($\mathbf{r}_2$), which contain the model's sublexical and lexical representations, respectively, given the input (I). However, as in the previous section, we will write the updates for a model of arbitrary size. Parameter updates can be written as before, assuming again that $I = \mathbf{r}_0$. First, the internal updates for $i = 1, \dots, n$ are computed, in which time is now explicit:

$$\bar{V}_i(t-1) = \hat{V}_i(t-2), \quad (2.7)$$

$$\bar{U}_i(t) = \hat{U}_i(t-1), \quad (2.8)$$

$$\bar{\mathbf{r}}_i(t) = \bar{V}_i(t-1)\hat{\mathbf{r}}_i(t-1), \quad (2.9)$$

$$\bar{\mathbf{r}}_c(t) = \bar{V}_n(t-1)\hat{\mathbf{r}}_c(t-1), \quad (2.10)$$

and in which V is a matrix. Once the internal (barred) estimates have been computed, the data-based (hatted) estimates for $i = 1, \dots, n-1$ are computed as follows:

$$\hat{\mathbf{r}}_i(t) = \bar{\mathbf{r}}_i(t) + \frac{k_{\mathbf{r}_i}}{\sigma_{i-1}^2} \bar{U}_i^T (\bar{\mathbf{r}}_{i-1}(t) - \bar{U}_i(t)\bar{\mathbf{r}}_i(t)) \quad (2.11)$$

$$- \frac{k_{\mathbf{r}_i}}{\sigma_i^2} \bar{U}_{i+1}^T (\bar{\mathbf{r}}_i(t) - \bar{U}_{i+1}(t)\bar{\mathbf{r}}_{i+1}(t)),$$

$$\hat{U}_i(t) = \bar{U}_i(t) + \frac{k_{U_i}}{\sigma_{i-1}^2} (\bar{\mathbf{r}}_{i-1}(t) - \bar{U}_i(t)\hat{\mathbf{r}}_i(t)) \hat{\mathbf{r}}_i(t)^T, \quad (2.12)$$

$$\hat{V}_i(t) = \bar{V}_i(t) + \frac{k_{V_i}}{\sigma_i^2} (\hat{\mathbf{r}}_i(t) - \bar{V}_i(t)\hat{\mathbf{r}}_i(t-1)) \hat{\mathbf{r}}_i(t-1)^T, \quad (2.13)$$

in which $k_{\mathbf{r}_i}$, k_{U_i} and k_{V_i} are (layer-dependent) learning rates and the σ^2 terms are noise variance and covariance parameters. The top-most layer update equations ($\hat{\mathbf{r}}_n$, $\hat{U}_n$, $\hat{V}_n$) differ from their lower-layer equivalents only in the

manner by which top-down feedback is incorporated:

$$\hat{\mathbf{r}}_n(t) = \bar{\mathbf{r}}_n + \frac{k_{\mathbf{r}_n}}{\sigma_{n-1}^2} \bar{U}_n^T (\bar{\mathbf{r}}_{n-1}(t) - \bar{U}_n(t) \bar{\mathbf{r}}_n(t)), \quad (2.14)$$

$$\hat{U}_n(t) = \bar{U}_n(t) + \frac{k_{U_n}}{\sigma_{n-1}^2} (\bar{\mathbf{r}}_{n-1}(t) - \bar{U}_n(t) \hat{\mathbf{r}}_c(t)) \hat{\mathbf{r}}_c(t)^T, \quad (2.15)$$

$$\hat{V}_n(t) = \bar{V}_n(t) + \frac{k_{V_n}}{\sigma_{n'}^2} (\hat{\mathbf{r}}_c(t) - \bar{V}_n(t) \hat{\mathbf{r}}_c(t-1)) \hat{\mathbf{r}}_c(t-1)^T. \quad (2.16)$$

Here, $\hat{\mathbf{r}}_c$, a set of context units similar to those used in simple recurrent networks (Elman, 1990), is the sum of $\hat{\mathbf{r}}_n$ and a classification cost term analogous to J_C in equation 2.2 (sPCC). The complete $\hat{\mathbf{r}}_c$ update equation is

$$\hat{\mathbf{r}}_c(t) = \hat{\mathbf{r}}_n(t) - \frac{k_{\mathbf{r}_n}}{2\sigma_n^2} (s_k(\bar{\mathbf{r}}_n(t)) - \mathbf{L}), \quad (2.17)$$

where $\mathbf{L}$ is the label vector corresponding to the input.

Updates to $\hat{\mathbf{r}}_c$ occur distinctly from updates to $\hat{\mathbf{r}}_n$ in order that label information be incorporated into the model during training using a term that can be silenced during testing, when U and V updates cease during model testing, $\hat{\mathbf{r}}_c$ and associated label information no longer contribute actively to model dynamics, yet the model's top-most layer is still able to propagate predictions of the input ($\bar{U}_2 \bar{\mathbf{r}}_2$) into layer 1, and its hidden representation of the input ($\hat{\mathbf{r}}_2$) still updates using bottom-up error ($\bar{\mathbf{r}}_1 - \bar{U}_2 \bar{\mathbf{r}}_2$). This bidirectional information flow is fundamental to hierarchical PC. Again we note that we have described the linear version of the model, in which the function f (see equation 2.1) is the identity function rather than $\tanh$. The nonlinear version is a straightforward generalization of these equations. An expanded version of the rPCC schematic (see Figure 1, bottom row) in which the time-dependent update scheme of rPCC is shown in greater detail is in Figure A.1 in the appendix.

2.4 Training and Testing Procedures.

2.4.1 Inputs. The model was trained on a set of 12 consonant-vowel-consonant-vowel (CVCV) "words" made up of combinations of the consonants /b/, /k/, /s/, and /r/ and the vowels /a/, / $\wedge$ /, /i/, and /u/. The consonants were chosen such that each consonant had at least one unique phonetic feature (see Table 1, where the unique consonant features are italicized). The 12-word lexicon consisted of /b $\wedge$ si/, /b $\wedge$ ra/, /k $\wedge$ si/, /k $\wedge$ ra/, /kibu/, /kisa/, /ribu/, /risa/, /suk $\wedge$ /, /suki/, /ruk $\wedge$ /, and /ruki/. This strategy was modeled after Magnuson, Tanenhaus et al. (2003); it provides a compact lexicon with tight control over phonological similarity. Each word

Table 2: CVCV Lexicon.

Word	Cohort	Rhyme
/b ^ˆ si/	/b ^ˆ ra/	/k ^ˆ si/
/b ^ˆ ra/	/b ^ˆ si/	/k ^ˆ ra/
/k ^ˆ si/	/k ^ˆ ra/	/b ^ˆ si/
/k ^ˆ ra/	/k ^ˆ si/	/b ^ˆ ra/
/kibu/	/kisa/	/ribu/
/kisa/	/kibu/	/risa/
/ribu/	/risa/	/kibu/
/risa/	/ribu/	/kisa/
/suk ^ˆ /	/suki/	/ruk ^ˆ /
/suki/	/suk ^ˆ /	/ruki/
/ruk ^ˆ /	/ruki/	/suk ^ˆ /
/ruki/	/ruk ^ˆ /	/suki/

in the lexicon had one cohort (sharing the same initial two phonemes as the target word) and one rhyme (sharing all but the first phoneme as the target word), and all other words in the lexicon are considered unrelated to the target word (even though they may share phonemes, such overlap does not appear to drive robust competition in human listeners). For example, the word /kibu/ had /kisa/ as its cohort and /ribu/ as its rhyme. Other constraints included (1) no repeated phonemes within a given word, (2) no two words are syllabic transpositions of each other (e.g., /kibu/ and /buki/ are not both in the lexicon), and (3) each phoneme appeared no more than four times at a given phoneme position. See Table 2 for each lexical item, along with its single cohort and rhyme. Each word also has nine unrelated words, which we do not show in the table.

2.4.2 Parameters. As shown in the bottom panel of Figure 1, we consider a model with two noninput layers. Our sublexical layer $\mathbf{r}_1$ consists of 10 units. This was established by trial-and-error; performance improved and stabilized as we increased the size of $\mathbf{r}_1$ from 5 to 10. Both $\mathbf{r}_2$, $\mathbf{r}_c$ and L are of size 12 (the size of the CVCV lexicon). The model was trained for 5,000 epochs. In each epoch, each word in the lexicon was presented once in random order. Each word was presented as a series of 98 (word duration) vectors of length 21 with values ranging from 0 to 1. Once again, weight updates occurred after loss calculations at each time step, representing a form of stochastic gradient descent. Free parameters were determined empirically:

- The noise variances were $\sigma_0^2 = 10$, $\sigma_1^2 = 10$, $\sigma_{1'}^2 = 1$, $\sigma_2^2 = 5$, and $\sigma_{2'}^2 = 1$.
- The learning rates for state unit and weight updates were $k_{\mathbf{r}_1} = 0.1$, $k_{\mathbf{r}_2} = 5$, $k_{U_1} = 0.1$, $k_{U_2} = 0.1$, $k_{V_1} = 0.01$, and $k_{V_2} = 0.01$.

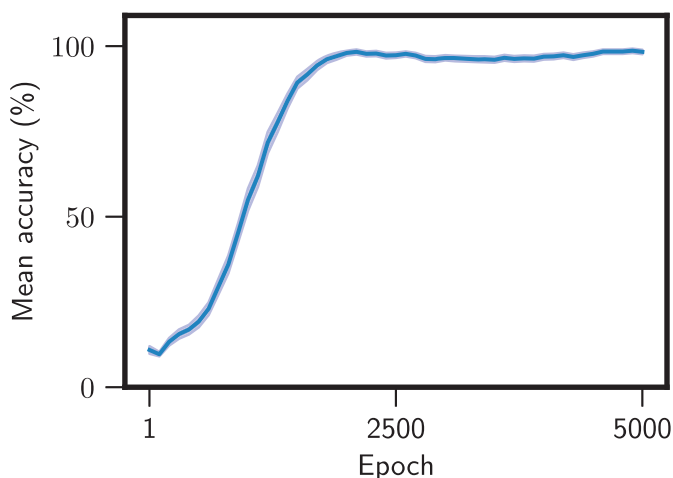


Figure 4: Training the rPCC model. The rPCC model correctly recognizes all 12 CVCV items in its training lexicon after around 2,000 epochs of training. The curve shows accuracy averaged over 100 unique training runs. Error bands indicate 95% empirical confidence intervals.

& Peters, 1998). A word was considered recognized when the activity of the lexical node corresponding to the input word was larger than the activity of all other lexical nodes. Accuracy was averaged over 100 unique training runs. As shown in Figure 4, mean accuracy gradually increased over the first 2,000 epochs and then plateaued, hovering around 100%.

3.2 Time Course. A critical empirical target of human SWR is phonological competition, which can be measured using eye-tracking data from the visual world paradigm (VWP; Allopenna et al., 1998), in which fixation behavior in response to spoken words demonstrated that word recognition is modulated by phonological relationships. Models producing or approximating this particular pattern in their recognition dynamics would be considered more behaviorally plausible than those that do not. Human data from the VWP experiment are shown in the left panel of Figure 5. Eye fixations to targets (e.g., beaker) and onset cohorts (e.g., beetle) separate from unrelated items (e.g., carriage) early on. As the target diverges from the cohort, fixations to the target increase while fixations to the cohort decrease. At the same time, as the input becomes more similar to the rhyme (e.g., speaker), rhyme fixations increase. The canonical, replicated target pattern for models is thus early cohort competition and later rhyme competition.

To generate the right panel of Figure 5, we presented each of the 12 CVCV words in turn to a trained rPCC model and recorded the (softmax-transformed) dynamics of the lexical layer, $\hat{\mathbf{r}}_2$. Here, we use a one-parameter

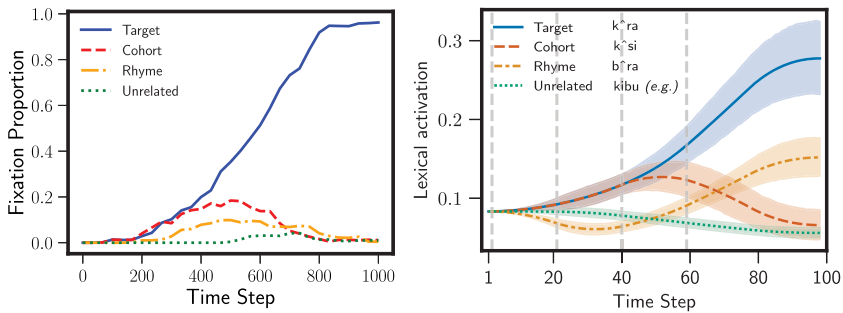


Figure 5: Human phonological time course and rPCC model lexical activation time course after 5,000 epochs of training. (Left) Human data: Fixation proportions over time; data from Allopenna et al. (1998) replotted. Proportions do not sum to 1.0 because fixations to a central cross are not included. (Right) Model: Since the outputs are one-hot encoded, each curve is the time evolution of the corresponding element of the softmaxed $\hat{\mathbf{r}}_2$ vector. Specifically, in the CVCV lexicon, for each input there is one vector element corresponding to the target word—one for its (single) cohort and one for its (single) rhyme. All other elements are averaged to produce the Unrelated (green dots) curve. These values are averaged over all targets that the model correctly identifies. An example target-cohort-rhyme triplet (and one of nine possible unrelated words) is shown in the legend. Phoneme onsets appear as vertical, dashed grey lines. Note that activations sum to 1.0 once we account for the nine unrelated items per stimulus set (see Table 2). Error bands show 95% empirical confidence intervals for the model.

version of the softmax function ($[s_k(\mathbf{x})]_i = e^{kx_i} / \sum_j e^{kx_j}; k = 20$): this function is equivalent to the Luce (1959) choice rule, where k is an additional squashing parameter (equivalent to inverse temperature) and is typically applied because the transformation better predicts human choice behavior. Activity was conditionally averaged over all targets that the model correctly recognized (in our case, all 12 words in the lexicon). We estimated empirical 95% confidence intervals using bootstrap resampling by repeatedly resampling the observed data with replacement 10,000 times and deriving the interval from the resulting distribution of estimates (this method is used for empirical CIs throughout the publication).

Average target activation increased and eventually dominated as the time-varying inputs unfolded, and both cohort and rhyme effects were present (see Figure 5). Segment (phoneme onset) markings (vertical dashed lines) in Figure 5 along with transcriptions of sample items make the details of these activations clear. For example, cohorts show identical activation patterns to targets until after time step 40, which is sensible given that this is the first segment (i.e., /r/ vs /s/) in which the two differ. After this

point, cohorts are suppressed relative to the target, echoing results found in human subjects (left panel). Rhymes are suppressed early because they differ in the initial segment, but then begin to rise sometime after the second segment is presented due to emerging overlap with the target. We take these results as a good indicator of the behavioral adequacy of the rPCC model.

We note a few final details about our operationalization of recognition. First, while it may not be intuitively obvious, the activations in the right panel of Figure 5 do indeed sum to 1.0 and can be interpreted as conditional probabilities. Roughly, at the end of the time course, the target has activation of approximately 0.26, the rhyme 0.15, and the cohort 0.07. The mean unrelated activation is just below 0.06. Add these numbers together (with nine unrelated items) and we obtain a sum of 1. It should also be noted that a softmax activation of 0.3 does not imply a recognition accuracy of 0.3; we can apply a decision policy to these activations. The operationalization of accuracy we have chosen based on these activations leads to high accuracy, since the targets are sufficiently dominant, regardless of their absolute activation value, to support robust word recognition. What these relative activations imply for more complex decision policies (like eye movements, for example) is a question beyond the scope of this letter, but see Magnuson (2025) for treatment of such details.

3.3 rPCC Model Validation Summary. By extending the Kalman-filter-based PC model developed by Rao (1999) to a multilevel, recurrent PC classification model (rPCC), we demonstrated that the model can achieve robust classification accuracy and human-like SWR behavior on a small but phonologically diverse lexicon. Specifically, the time course of lexical node activations mimics empirical observations in human eye-tracking data (see Allopenna et al., 1998). To our knowledge, this is the first time a PC model has been shown to demonstrate a human-like SWR time course, which we hold to be a critical baseline for models of human SWR. With the validated rPCC model in hand, we now turn to additional crucial questions: (1) whether the rPCC model demonstrates reduced responses to familiar versus novel items as observed by Gagnepain et al. (2012), (2) whether it exhibits anticipatory activation (i.e., evidence of predictive processing), and (3) whether it demonstrates putative hallmarks of PC that have been observed in human neural activity (Blank & Davis, 2016).

4 Predictive Processing and Putative Neural Hallmarks of PC _____

In what follows, we test the rPCC model on a lexical prediction paradigm that isolates lexically based phoneme predictions (Gagnepain et al., 2012), and then a representational similarity paradigm aimed at identifying putative hallmarks of PC in patterns of neural responses to speech (Blank & Davis, 2016). A particular pattern of similarity (assessed via

representational similarity analysis, or RSA; Kriegeskorte, 2008) has been proposed to be indicative of PC (Blank & Davis, 2016), but this has not yet been tested in a model that plausibly implements formal PC.

4.1 Lexical Prediction Paradigm. Gagnepain et al. (2012) presented listeners with words and word-like strings that either conformed to lexical expectations (e.g., “formula”) or violated those expectations (e.g., “formubo”). They used a two-day design, where listeners were familiarized with one set of nonwords on the first day and a second set introduced on the second day. After the second-day exposure session, participants did a pause-detection task while their brain activity was monitored using MEG (magnetoencephalography). The critical question is what would happen at the deviation point (DP) in their items. For the word/nonword pair, “formula/formubo” the deviation point was immediately after the /u/ sound (the point where the nonword deviated from the word). The critical question was whether, on day 2, the day 1 nonwords would pattern with familiar words or with the novel nonwords presented only on day 2. If the day 1 nonwords patterned with words, that would be consistent with those items beginning to be “lexicalized” (likely due to the opportunity for sleep-based consolidation).

Gagnepain et al. (2012) found an elevated neural response in the superior temporal gyrus (STG) at the deviation point to novel nonwords that were first encountered on day 2 (relative to responses to corresponding true words at the deviation point). No such difference was observed for nonwords to which listeners had been first exposed on day 1, suggesting that the day 1 nonwords had indeed begun to be consolidated into listeners’ lexicons. Gagnepain et al. (2012) interpreted this result as consistent with PC: once a nonword like “formubo” began to be consolidated into lexical knowledge (after an opportunity for sleep-based consolidation), the phoneme at the deviation point became less surprising because it was becoming lexically consistent.

Luthra et al. (2021) questioned the interpretation that this necessarily confirmed the operation of a formal PC mechanism and tested whether the TRACE model (McClelland & Elman, 1986) could simulate their results by exhibiting predictive activation of anticipated elements and showing model-internal activation reductions for expected versus unexpected inputs at the deviation point (analogous to signal reductions in human MEG). Given that TRACE is an interactive activation model, without any explicit prediction mechanisms, it seemed plausible that it would not show predictive processing or activity reduction for expected inputs compared to unexpected inputs. In fact, TRACE exhibited both: lexically expected phonemes were anticipatorily activated due to top-down lexical feedback, and there was lower model activity given expected (lexically consistent) phonemes than unexpected phonemes (specifically, the amount of lexical-to-phoneme feedback was lower, as were total lexical inhibition and phoneme inhibition;

when inputs were words, the model was able to settle on stable activation patterns at the phoneme and word levels, resulting in lower total activity).

While these TRACE results demonstrate that predictive processing and lower activity given expected inputs may not be unique to PC models, a PC model that failed to show these patterns would be obviously flawed. We now therefore examine whether the rPCC model exhibits predictive processing in this paradigm, as well as whether system-internal activity reduction for expected inputs emerges in rPCC. If this reduced internal activity is related to reduced prediction error, we should observe this specifically in the error signals in the rPCC model. In the next section, we assess whether the model exhibits reduced activity for familiar (lexically consistent) items relative to unfamiliar (novel nonword) items, and in the subsequent section, we return to the question of anticipatory activation.

4.1.1 Lexical Prediction Paradigm Simulations with rPCC.

Simulations. We followed Gagnepain et al.'s 2012 approach and created novel words based on the trained lexicon (see section 2.1) by replacing the last two phonemes of a given word with other trained phonemes while maintaining the CVCV word structure. We created six novel strings, each related to a cohort pair from the trained lexicon. The six novel forms were /b^hk^h/, /k^hbu/, /kir^h/, /riki/, /sura/, and /rusa/, designed to match the input intensity (i.e., number of features and therefore total summed amplitude) of the trained words.

Before discussing our simulations of the lexical prediction paradigm, we first show $\hat{r}_2(t)$ activation curves for a word, /b^hra/, and a similar nonword, /b^hk^h/. Figure 6 shows the activity over time of each of the 12 words in the lexicon in response to the word (top panel) and nonword (bottom panel). Cohorts to each input item are colored red and rhymes blue. The target (input) word is colored black in the top panel; there is no black curve in the bottom panel because the nonword is not in the lexicon. For both the word and nonword, we see sensible competitor activation and general suppression of unrelated words.

To simulate the lexical prediction paradigm, each word and nonword was presented to the trained rPCC model, one at a time. Note that no model weights are updated during these simulations, and as each item is presented, state vectors are initialized with rPCC's quasi-sparse kurtotic $\mathbf{r}$ distribution (all zeros), and inputs are presented without noise. During run time, we monitored the L_1 norm (summed absolute value) of six signals in the model: the norms of the sublexical ($\hat{\mathbf{r}}_1$) and lexical ($\hat{\mathbf{r}}_2$) level representations, along with two kinds of model prediction errors per level (four prediction errors total), totaling six monitored values. The first set of prediction errors we consider are the differences between bottom-up information (hidden representations) and top-down information (predictions of the input),

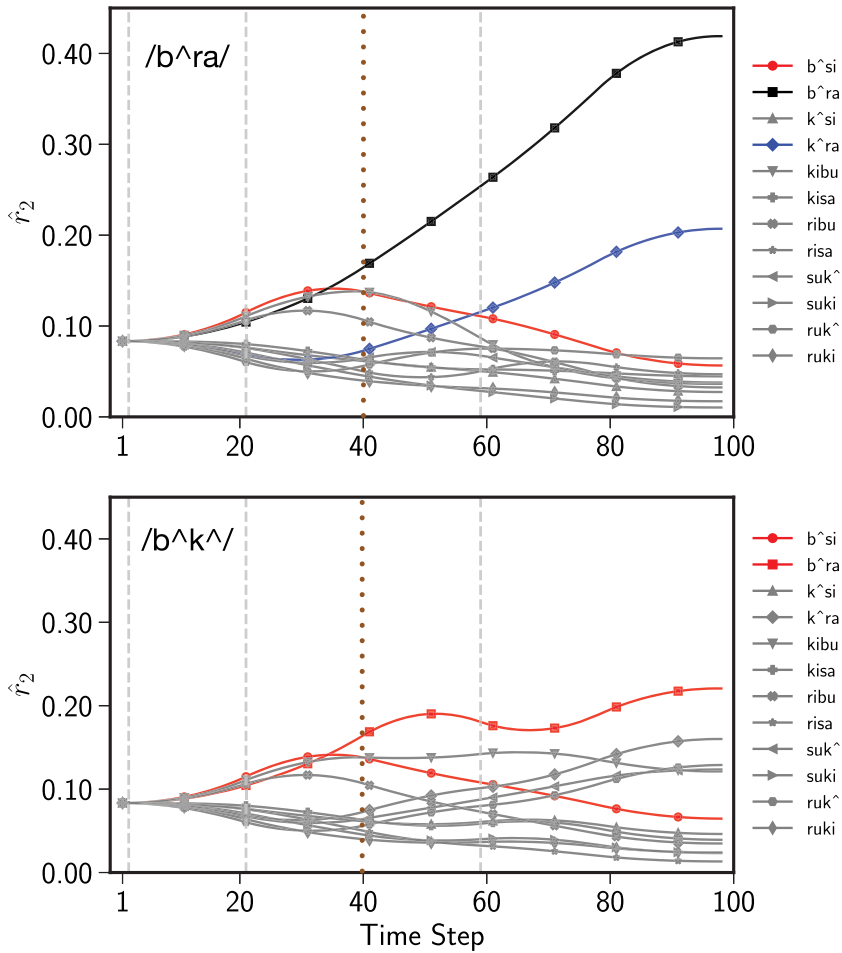


Figure 6: Time course of word and nonword activation. The top panel shows the model's softmax response to a word input ($/b^ra/$) and the bottom to a nonword input ($/b^k^/$). The target (input word) is colored in black, cohorts in red, and rhymes in blue. There is no black curve in the black plot because $/b^k^/$ is not in the lexicon. Phoneme onsets are marked with vertical gray lines, and the deviation point (also an onset) is marked with a brown dotted line.

$$\mathbf{e}_i(t) = \hat{\mathbf{r}}_{i-1} - U_i \hat{\mathbf{r}}_i, \quad (4.1)$$

where $i = 1, 2$ and as before $\mathbf{r}_0$ is understood to be the input I itself. We also monitor the within-layer errors,

$$\mathbf{e}'_i(t) = \hat{\mathbf{r}}_i(t) - V_i \hat{\mathbf{r}}_i(t-1), \quad (4.2)$$

which index the size of the correction made to the model state once the data have been seen. These $\mathbf{e}'$ errors will be small when the model's prediction of the data (prior to seeing the next sample) is good.²

Results. Figure 7 shows the results of these simulations; here, lines are activation means by word category (trained or novel) and error bands are standard error. While one can see a difference between the trained and novel words (mean-only or mean plus bands) in all panels, the effect is pronounced in $\mathbf{e}_1(t)$ (the sublexical to input error) and $\mathbf{e}'_1(t)$ (the difference between the sublexical activity before and after the data has been seen). These differences replicate the findings of Gagnepain et al. (2012), in which the responses to the trained words are reduced compared to novel words. Moreover, these response differences consistently appear where expected, starting at the onset of the second syllable (novel word deviation point: the vertical, dash-dotted brown line at time step 40). Note that reduced prediction error by definition demonstrates predictive processing in the model, though we will probe this issue more deeply in the following section.

Unexpectedly, there are also quantities where values are greater for at least some time steps for trained items than novel items (notably in the top and bottom right panels, $\hat{\mathbf{r}}_2$ and $\mathbf{e}'_2$). This discrepancy specifically occurs in the top-most layer (layer 2) of the model. Adding the classification task to PC model training breaks the hierarchical symmetry of layer objectives (see sPCC schematics, as well as equations 2.1 and 2.2): in the classifying model, the top-most layer must simultaneously form a dense representation that can be used (in tandem with U_2) to reconstruct the contents of the first layer, but must also try to produce a sparse representation that matches a one-hot vector from one of the training words. It is possible that these two goals, which are partially at odds, cause the second layer to show a sharpening behavior in this case. An alternate way to implement classification would be to have a final set U_o of projection weights such that $\mathbf{r}_2$ would not be directly coupled to the label information. Alternatively, one could also extend the model to three layers to see if the top-most layer still behaves

² Our primed error measures have a relation to prior work in sentence processing (Rabovsky et al., 2018; Rabovsky & McClelland, 2020). Those authors introduce the concept of "instantaneous prediction error," which they then calculate in their sentence-gestalt model. This model computes only a single measure of change (which they call the "semantic update"), which is the total change in hidden-unit activation following each word (Rabovsky et al., 2018, and see also Rabovsky & McClelland, 2020). Their semantic update corresponds conceptually to our $\mathbf{e}'_i$ error terms, which quantify the mismatch between the predicted internal state and the updated state (at a given layer). The sentence-gestalt model does not include an analog of our $\mathbf{e}_i$ term (unprimed error terms), which is the sensory residual comparing bottom-up input to top-down predictions within the hierarchical rPCC model (again, at a given layer).

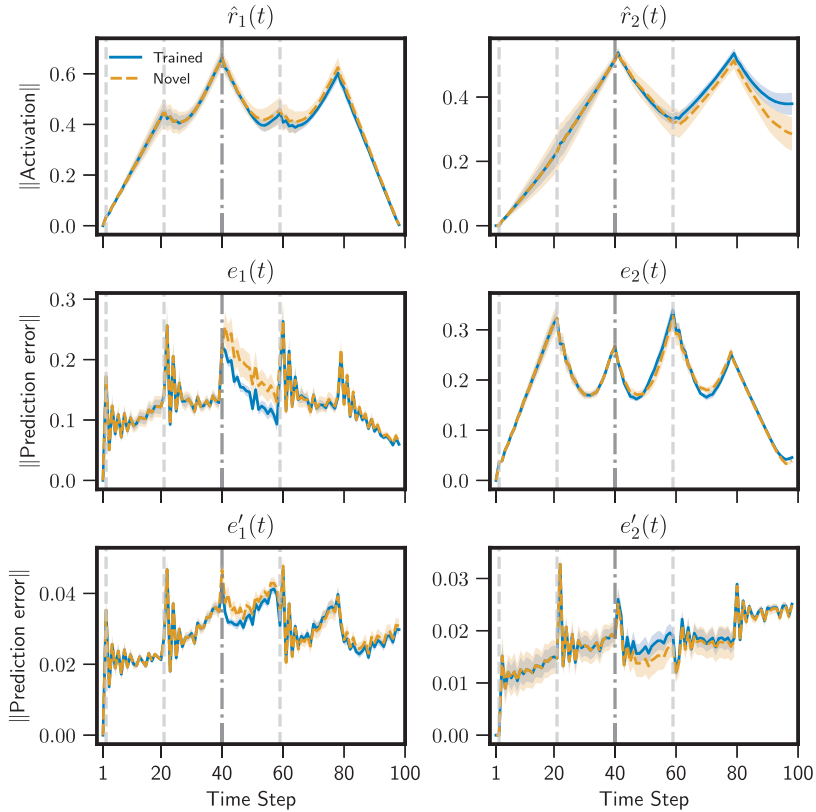


Figure 7: rPCC model performance in the lexical prediction paradigm. $\|\hat{\mathbf{r}}_1\|$ (top left), $\|\hat{\mathbf{r}}_2\|$ (top right), $\|\mathbf{e}_1\|$ (center left), $\|\mathbf{e}_2\|$ (center right), $\|\mathbf{e}'_1\|$ (bottom left), $\|\mathbf{e}'_2\|$ (bottom right), for both trained and novel lexical items as described in the text. Lines are respective averages over 12 trained or 6 novel words; 1 simulation (deterministic). Vertical lines are phoneme onsets, and the dot-dashed brown line is the deviation point (DP) between trained and novel items. Error bands are $\pm$ SE.

differently from the first two with regard to expected and novel items. We leave exploration of these possibilities for future work and note that a PC system need not exhibit reduced signal (when inputs are predictable) in every node in the system, only those elements representing prediction error. Thus, reduced signal is not necessarily expected in the novel classification layer.

We also note that the granularity of the representations formed in $\hat{\mathbf{r}}_1$ and $\hat{\mathbf{r}}_2$ is interesting. Recall that both the trained and novel items are two-syllable

CVCV patterns, and the second syllable, while it occurs in the training set, is paired with a different initial CV pair in the novel items. The rPCC model has no explicit representation of syllables or phonemes, of course, only phonetic features and words. Nonetheless, in the sublexical layer ($\hat{\mathbf{r}}_1$), the norm of the representation has four evenly spaced peaks, corresponding to the phoneme peaks (where they ramp to maximum feature intensity) in the input items (see time steps 21, 40, 59, and 78). Conversely, the norm of the lexical layer ($\hat{\mathbf{r}}_2$) has only two evenly spaced peaks, which line up with the two vowel peaks in the input item (see time steps 40 and 78). Accordingly, it seems as though the lexical layer of the model has abstracted away from the phoneme-level information in favor of larger functional units, corresponding to syllables. In future work, we aim to investigate whether this phenomenon is robust and would persist with a larger, more naturalistic lexicon.

The rPCC model thus exhibits signal reduction for expected versus unexpected inputs (in layers where prediction error is encoded). We emphasize that while these reductions in activity in the model are intuitive, this has not been tested previously in a plausible implementation of predictive coding for speech-like over-time inputs. As we have mentioned, by definition, reduced prediction error constitutes predictive processing. That said, it would be even more compelling if we could demonstrate anticipatory activation of lexically predicted phonemes, as Luthra et al. (2021) did. Given that the model does not have any explicit phonemic readout, this is a challenge. In the next section, we address this by decoding hidden states to assess to what degree lexically predicted upcoming phonemes and syllables are represented in internal model activations as stimuli are presented.

4.2 Decoding Predictive Processing. As noted above, accurate prediction error is by definition an index of predictive processing. But prediction error is not an explicit index of step-by-step anticipation. It does not reveal, for example, to what degree the model may be anticipating richer and further-in-time events. Here, we aim to address this by decoding to what degree internal model states can predict upcoming lexically consistent inputs. Specifically, we assess whether there is any evidence for anticipatory activation: activation of phonemes or syllables prior to their onsets. Because the rPCC model has no localist nodes for syllables or phonemes, we adopt a decoding approach to address this question.

We select the 12 CVCV training items from the lexical prediction paradigm, present them to the model, and collect hidden state (activations $\hat{\mathbf{r}}_1(t)$ and $\hat{\mathbf{r}}_2(t)$) over time. We then assign true cluster labels to each of the four phonemes in each pattern, based on which phonemes appear at those positions. For example, if we look at just the first phoneme, it is either /r/, /k/, /b/, or /s/. We convert these to cluster labels 0,1,2,3, and then populate the 12-element set of true labels appropriately. We then repeat this process for the other three phonemes in each pattern, as well as a separate

set of syllable labels for both the first and second syllables. Again, as an example, the initial syllable is either /ru/, /ki/, /k̂/, /b̂/, /su/, or /ri/, so we would need six labels in that case.

We then cluster the hidden state activations using K-means (Forgy, 1965) for the two model layers separately at each time point. In every clustering, we set K to be equal to the true number of clusters (phonemes or syllables, respectively). At each time point, we compute the adjusted mutual information (AMI: Vinh et al., 2009) between the appropriate set of true labels and the labels provided by K-means. The AMI ranges between 0 and 1 (though in small samples, it can be negative) and is corrected for unbalanced class labels. A value of unity represents perfect decoding accuracy. The groups provided by the hidden states map exactly onto the known groups (e.g., the fourth phoneme in the input patterns), and a value of 0 indicates the complete absence of decoding. We would take as evidence of anticipatory processing the ability of the model to at least partially decode a phoneme or a syllable before its onset occurs.

Figure 8 shows the result of this cluster-based decoding analysis. In the top row (panels A and B) we show results for phoneme decoding by both $\hat{\mathbf{r}}_1(t)$ and $\hat{\mathbf{r}}_2(t)$, while the bottom row (panels C and D) shows syllable decoding. Figure 8A shows clear evidence of anticipatory processing, for all phonemes with the possible exception of P1, whose onset is extremely early. We see clear anticipation of phonemes P2 and P3 during the early portion of P1, as well as increasing anticipation of those phonemes and clear anticipation of P4 as the input for P2 is encountered. While maximal decoding accuracy tends to occur slightly after peak input activity (see Figure 2 for an example of pseudospectral input), there is clearly partial decoding accuracy (AMI ~ 0.4 - 0.5) prior to the phoneme onsets. (Onsets here correspond to the beginning of each phoneme or syllable's "ramp-up," prior to any coarticulatory or other cues to the element's identity; these are dotted vertical lines in Figure 8.)

We observe anticipatory activation but only in $\hat{\mathbf{r}}_1(t)$. Figure 8C shows that the activity in $\hat{\mathbf{r}}_1(t)$ can also be used to accurately decode syllables, though they are not decoded as early as phonemes (decoding occurs one to two time steps before syllable onset). Figures 8B and 8D show that $\hat{\mathbf{r}}_2(t)$ does not exhibit meaningful anticipatory processing of either phonemes or syllables. Panel B shows the AMI of phonemes 1 to 3 remaining fixed from P1 onset time until around time step 50 (well past P2 and 3 onset). Panel D reveals no anticipatory syllabic processing at all.

In sum, $\hat{\mathbf{r}}_2(t)$ is generally poor at decoding anything beyond the first phoneme or syllable, but this is not entirely unexpected, since $\hat{\mathbf{r}}_2$ is the layer coupled to label information; it is much more important (and part of the training objective) that $\hat{\mathbf{r}}_2$ recognizes the word rather than any of its subconstituents. This analysis complements the evidence for predictive processing that we obtained via prediction error, with explicit measures of model-internal activation patterns demonstrating clear anticipation of

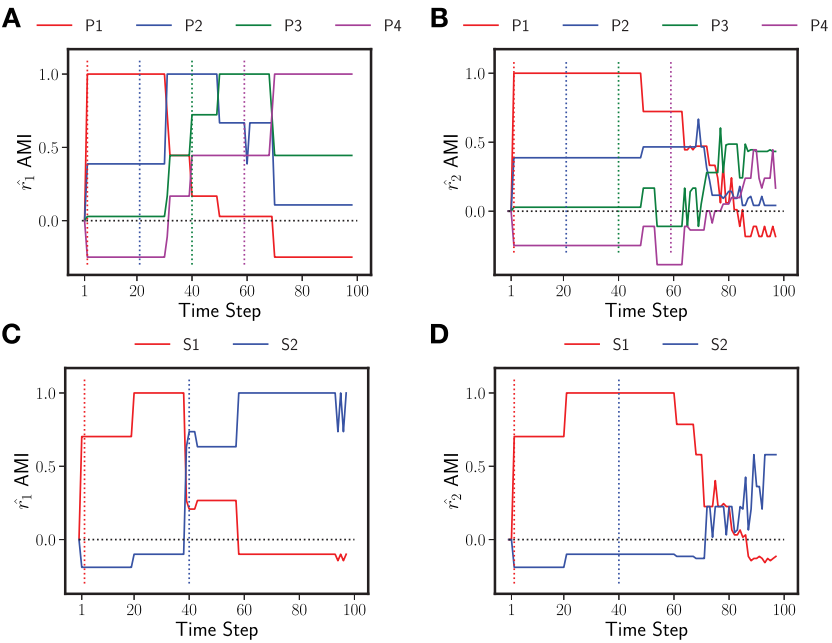


Figure 8: Cluster-based decoding of rPCC hidden states from the lexical prediction paradigm CVCV input patterns. (A, B) Phoneme decoding by $\hat{r}_1$ and $\hat{r}_2$. (C, D) Syllables. In panels A and B, phoneme onsets are indicated by dotted lines, color coded to match the decoding curves for the first phoneme (P1), second (P2), third (P3), and fourth (P4) in each CVCV input patterns. In panels C and D, the syllable onsets are indicated with color-coded dotted lines.

lexically predictable phonemes and (to a lesser degree) syllables, and with layer 2 predictions revealing a top-layer asymmetry presumably caused by the purpose of its representation.

Finally, there is a notable difference between decoding phonemes and syllables. Syllabic decoding is much weaker and later than phonemic decoding; this is sensible, given that the model does not contain any explicit syllable units. Syllable identity must be inferred indirectly from patterns distributed across both levels 1 and 2. This is in contrast to phonemic decoding. While the model also lacks explicit phoneme units, phonemic decoding must depend on layer 1 since layer 2 explicitly represents words. Moreover, AMI measures cluster-level discriminability, so it only rises above baseline when the representations are sufficiently stable, which for syllables occurs just before syllable 2 onset. What is interesting is that we get any syllable-level information at all, as seen in Figures 7 and 8, despite no explicit encoding of, or training on, syllabic units.

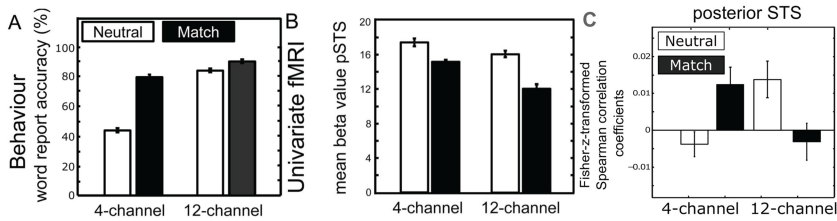


Figure 9: Key behavioral (A) and neural (B, C) results from Blank & Davis (2016). They observed an additive effect of clarity and prime in behavior (A) and (in a complementary fashion) univariate fMRI results (B), but a surprising interaction in their RSA analysis (C). Adapted from their Figures 3 (for our panels A and B) and 4 (for our panel C).

Next, we consider a proposed signature of PC in human cortical responses to speech inputs.

4.3 Representational Similarity Analysis. In another study aimed at distinguishing between IA and PC frameworks, Blank and Davis (2016) investigated how simplified models for each framework could simulate multivariate functional magnetic resonance imaging (fMRI) data from the left posterior superior temporal sulcus (pSTS, associated with sublexical representations) during speech perception.

In their experiment, participants heard spoken words degraded with two levels of clarity using vocoding: 12-channel (high clarity/low noise) and 4-channel (low clarity/high noise). They also had two levels of prior: vocoded spoken words preceded by a matching visual prime (the word itself, in the match condition) or a neutral prime (XXXX), in the neutral condition. Their behavioral results (see Figure 9A) showed additive influences of clarity and prior on accuracy. Their univariate fMRI results showed a nearly complementary additive pattern, with main effects of clarity and prior (see Figure 9B).

Crucially, they then used representational similarity analysis (RSA, details below; Kriegeskorte, 2008; Popal et al., 2020; Blank & Davis, 2016) to examine how much activation patterns in pSTS in response to a given word (e.g., “sing”) resembled activation patterns in response to a related word (differing at offset or onset, e.g., “sit” or “thing”) versus an unrelated word (e.g., “bath”). The crucial pattern Blank & Davis (2016) observed is a crossover interaction between priming and signal clarity. Specifically, RSA correlations to pSTS fMRI data in the neutral prime (XXXX) condition increased when moving from low clarity (high noise) to high clarity (low noise), but decreased from low to high clarity in the match prime condition, the printed form of the word itself (see Figure 9C).

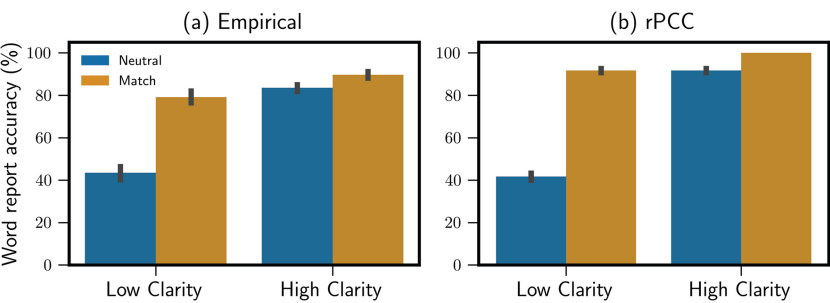


Figure 10: Word recognition accuracy in humans and the rPCC model. (a) Human behavioral data shown from Blank & Davis (2016), averaged over 21 subjects. (b) Classification accuracy of the rPCC model, averaged over 100 unique simulations. Noise levels of 0.55 and 0.275 correspond to low and high clarity (see text for details). Error bars in both plots indicate 95% empirical confidence intervals.

Blank and Davis posited that this outcome followed naturally from a system organized around prediction error. More specifically, they posited that lexical priming and signal clarity would have differential effects on two types of models: an IA-based sharpened signal model versus a PC-based prediction error model. Representations in IA models were assumed to be sharpened (i.e., enhanced) by matching lexical priors via positive feedback from lexical to sublexical representations (enhancing the activation of lexically consistent sublexical units) regardless of input clarity, such that signal clarity and priming should have the same impact: increasing relative similarity of related versus unrelated inputs.

In contrast, according to Blank & Davis’s account, since the primary representations in a PC model are prediction errors, their prediction error model’s internal representations would be more similar to input patterns when the prediction errors are informative (i.e., with large and variable values). Specifically, prediction error should be low when the mismatch between the prior and signal is low. This should obviously be true for high clarity and a matching prior (i.e., when the spoken word in a low noise corresponds to the previously presented visual prompt). Counterintuitively, prediction error should also be low for low clarity and a neutral prior because the discrepancy between prior and signal will be relatively low (Blank & Davis assumed the XXXX neutral visual prime should yield a uniformly low prior expectation over all words).

Their human subjects’ behavior indicated facilitating influences of both prime (better performance for match than neutral) and clarity (better performance for high than low clarity; see panel A of Figure 10). Crucially, using an RSA approach, they found their predicted crossover interaction in their fMRI data, with a correlation pattern consistent with higher activity

in the conditions expected to generate higher prediction error (neutral prime + high clarity and match prime + low clarity) and lower activity in the other two combinations.

Both their sharpened signal model and their prediction error model generated human-like simulations of behavior (with additive effects of clarity and prior, as in Figure 9A) and univariate fMRI responses (with responses approximately complementary to the behavioral pattern, as in Figure 9b). However, their sharpened signal model did not show the crossover pattern in their RS analysis; instead, it showed the same trends as in the behavioral simulations (consistent with both the match prime and high clarity driving higher activity). However, the prediction error model did show the crossover, with higher RSA correlations for neutral prime+high clarity and match prime+low clarity. Blank and Davis (2016) concluded that the crossover interaction is uniquely diagnostic of PC-like mechanisms.

In the next section, we simulate the manipulations from Blank and Davis (2016) in an effort to evaluate whether a formally specified, neurally and behaviorally adequate predictive coding model (our rPCC model) reproduces these patterns when tested under appropriate conditions.

Simulations. To simulate the clarity manipulation in the Blank and Davis (2016) study, we added gaussian noise (mean zero, standard deviation 0.275 for high clarity, and mean zero, standard deviation 0.55 for low) to the 12 CVCV words in the model lexicon (with values ≤ 0 clipped to 0, and values ≥ 1 clipped to 1). These noise variances were determined to produce model accuracies similar to human accuracy in the Blank and Davis (2016) study (see Figure 10). The resulting input values were clipped at a minimum of 0 and maximum of 1. We created 10 variants of each of the 12 CVCV words by adding random noise to the original items, resulting in a total of 120 inputs. Priming with neutral priors was simulated by initiating the r_2 nodes with all zeros, whereas the match priming condition was simulated by initiating the target word node with a small positive value (0.02, which led to accuracy patterns similar to those of Blank & Davis, 2016) and the other nodes with zeros (see Figure 10). Here, 100 unique simulations were performed.

Linkage to Univariate fMRI. Using the same simulation data that produced the rPCC behavioral results in Figure 10, we then explored whether we could find a quantity in the rPCC model that could replicate the univariate fMRI patterns that Blank and Davis (2016) observed (see the middle panel of Figure 9). The pattern they observed can be characterized by three relationships: (1) neutral prior, low clarity (NL) > neutral high (NH) > match(ing) prime, low (ML) > match high (MH); (2) main effect of prior: Neutral (nL + nH) > match (ML + MH); and (3) main effect of clarity: low (nL + ML) > high (nH + MH). We reasoned that univariate pSTS activity reflects how much total processing the system has to do to settle on a word

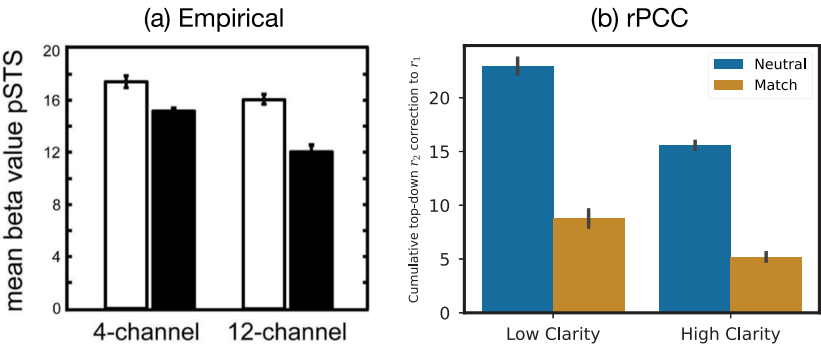


Figure 11: Univariate fMRI analog in the rPCC model. Here we show (b, right) the AUC of the absolute magnitude of the top-down correction from r_2 to r_1 until the model settles to a stable softmax prediction of the target word side-by-side (a, left) with the univariate pSTS measurement previously shown in Figure 9.

rather than momentary costs of processing. If this is the case, then quantities related to model settling time and cumulative processing burden should provide a reasonable analog to the univariate results. Clearer speech (high clarity) and informative priors (match) reduce processing cost, so the system should be able to settle more quickly and accumulate a lower total processing burden. We tested several measures that tracked the accumulation of activity (akin to area under the curve, AUC) of a given quantity until the model’s word prediction stabilizes. All provided robust matches to the univariate fMRI results. Here, we show just one representative model quantity: the cumulative (absolute) magnitude of the top-down correction signal from r_2 to r_1 until the model settles.

Figure 11 shows this model quantity alongside the original experimental result (shown also in Figure 9). We observe the correct trend in both noise and prior conditions. A similar pattern was seen in AUC measures until settling of the magnitudes of four other quantities: low-level (sensory) error, cross-level (r_2 to r_1) mismatch, r_1 within-level mismatch, and the bottom-up correction to r_2 from r_1 . Even simple settling time alone showed the correct pattern for clarity, though with a weaker effect of prior. Thus, various measures of processing burden in the rPCC model all replicate the univariate pattern observed by Blank & Davis (2016) quite well.

Representational Similarity Analysis. Our RSA followed the procedures used by Blank & Davis (2016). Figure 12 (top panel) shows our hypothesized representational dissimilarity matrix (RDM) based on phoneme similarity among words, such that words were defined as similar (with a value of 0) to their cohort and rhyme counterparts and dissimilar (with a value of 1) to other words in the lexicon. A word’s comparison to itself was

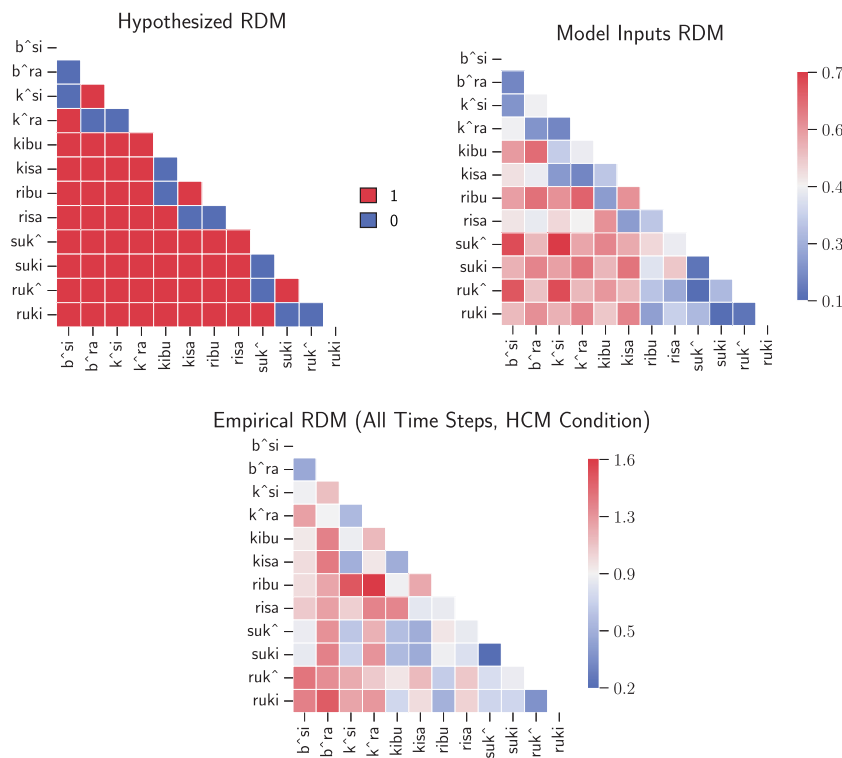


Figure 12: Representational dissimilarity matrices. (Top left) Theoretical RDM in which each target word’s rhyme and cohort are similar to the target, and all unrelated words perfectly dissimilar. (Top right) Dissimilarity matrix for the model inputs, on a scale of [0, 2] (calculated as $1 - r$, where r is the Pearson correlation). (Bottom) Empirical RDM from the rPCC model, for the high clarity match condition. This was generated from a single simulation, averaged over all 98 input timesteps. The dissimilarity scale is the same as in the middle panel.

excluded from the RDM, as in the original study. Observed RDMs were measured by computing the dissimilarity (one minus Pearson correlation coefficient) between words for the same six quantities shown in Figure 7.

Due to the time-varying nature of the inputs and model dynamics, one can compute an RDM of a given model quantity at a specific time step or concatenated across multiple time steps. For the results we present here, we computed RSA separately for each time step and then averaged the resulting correlations. In all cases we computed the Spearman correlation between the hypothesized and observed RDMs. In Figure 12, we also show the RDMs for the model input (middle panel) and an example of an empirical (model-derived) RDM (bottom panel) from a single simulation in the

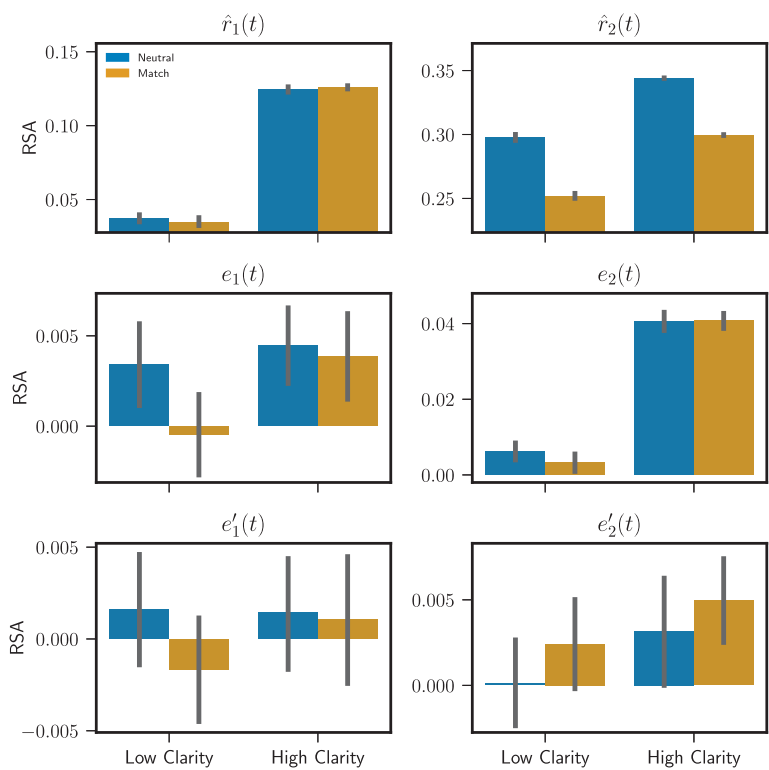


Figure 13: Effects of prior and clarity on simulated RSA, time steps 1–98, Averaged. RSA results (Fisher-z-transformed Spearman correlation coefficients) for each of the model quantities shown in Figure 7: representations ($\hat{r}$), across-layer errors (e), and within-layer errors (e'). Though robust prior and clarity effects are present (e.g., in $\hat{r}_1$ and $\hat{r}_2$), no quantity shows the crossover effect observed in Blank and Davis (2016). Bars represent, for each input condition, the signal average across all processing time steps, averaged over 100 unique simulations. Error bars show 95% empirical confidence intervals.

high clarity match condition. Note that the feature-based inputs reflect the theoretical (binary) RDM quite well, though we emphasize we did *not* use the input RDM in our analysis, only the theoretical RDM.

Results and discussion. Figure 13 shows the results of the simulated RSA analysis. Overall, the effects of signal clarity are clearly observed in each panel, though in three cases, e_1 , e'_1 , and e'_2 , the variance in outcomes is quite large. We also measure priming effects in the means of every panel, but the only panel where the result is robust (distinguishable given the error bars) is for $\hat{r}_2$.

Crucially, and contra to the findings of Blank and Davis (2016), we do not observe a crossover pattern in any model component. This pattern, consistent with the original study, would be blue bar heights increasing from low to high clarity, while orange bar heights decrease. Thus, this putatively PC-diagnostic pattern does not emerge from our formal PC model.

The failure of rPCC to produce the reported crossover does not in itself rule out predictive processing accounts of these empirical results. This divergence between rPCC and the neuroimaging pattern therefore highlights an important theoretical question rather than resolving it: Which specific architectural or computational ingredients are necessary for a predictive coding model to generate the observed interaction? Below, we speculate on possibilities for the discrepancy: testing these possibilities will be reserved for future work.

Model discrepancies. What might account for the ability of the prediction error model of Blank and Davis (2016) to simulate the RSA cross-over interaction but the absence of this pattern in the rPCC model? One key difference is that the rPCC model processes time-varying inputs, while Blank and Davis's models treat each word as a static, image-like vector with features from all phonemes presented all at once. We assume that the aggregated RSA reflects the model dynamics holistically and is an appropriate simulation for fMRI-based RSA (with its low temporal resolution), but we note that Blank and Davis (2016) computed model RSA by running the models for only a single iteration/time step (for other simulations, they let the models run for hundreds of iterations). We therefore inspected the rPCC RSA results during early simulation time. Specifically, we focused on time step 2 in the rPCC model, because this (phoneme 1 onset) is the first time step where all observed RDMs have valid values (see Figure 14). At time step 2, mean effects of both priming and signal clarity are evident in all components. Additionally, in keeping with the core finding from Figure 13, five of the six model components fail to exhibit the RSA interaction pattern observed by Blank & Davis (2016). While the pattern can be seen in one component, e_2 , the effect is mean-only and highly sensitive; taken together with the results in Figure 13, crossover appears, at best, transient, and unreliable as a diagnostic of PC architecture.

rPCC Mechanism. It is possible that the rPCC model lacks mechanisms (specific forms of hierarchical precision-weighting, nonlinear gating, or representational granularity) that would be required to capture the crossover interaction. For example, the rPCC model (like the original Rao & Ballard, 1999, model) allows an arbitrary nonlinear transformation of the predicted representations (the f in the $f(Ur)$ terms in equation 2.1. We used the identity function, but it is possible that using some nonlinear function here could yield a model with similar accuracy that displays the crossover.

fMRI. The spatial and temporal integration properties of fMRI may interact with predictive processing in ways that are not yet fully captured by current neural network predictive coding formulations.

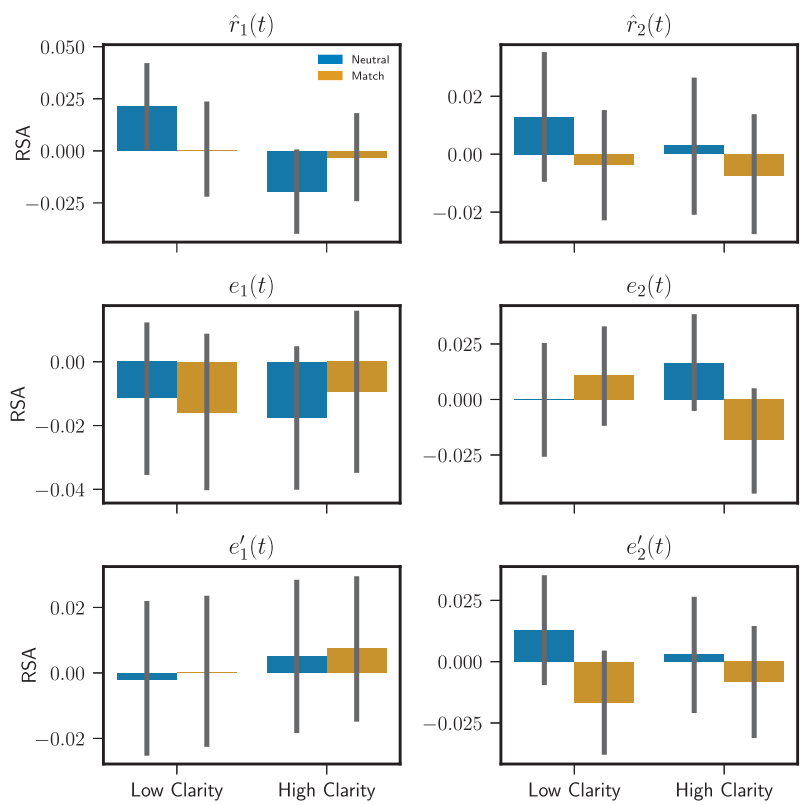


Figure 14: Effects of prior and clarity on simulated RSA, time step 2 only. Time step 2 (as opposed to time step-averaged) RSA results (Fisher-z-transformed Spearman correlation coefficients) for each of the model quantities shown in Figures 7 and 13: representations ($\hat{r}$), across-layer errors (e), and within-layer errors (e'). Most important, supporting the key finding of the bar plots in Figure 13, five of the six quantities do not reveal the crossover effect observed in Blank & Davis (2016). In the one quantity where this is present, e_2 , the effect is mean-only and highly sensitive. Additionally, and in contrast to Figure 13, no robust prior or clarity effects are seen in any component. Here again, bars represent averages per input condition over 100 unique simulations. Error bars show 95% empirical confidence intervals.

Other mechanisms. The discrepancy may reflect additional mechanisms operating alongside prediction error minimization; that is, perhaps prediction error is not the only signal driving both learning and the subsequent dynamics.

While the definitive answer to the relationship between the observed crossover and predictive coding processes may not yet be evident, the possibility we set out to test remains: the observed RSA crossover Blank

and Davis (2016) may not be a hallmark, or reliable indicator, of predictive coding of speech recognition in the brain.

While our analysis closely parallels the original RSA calculations (Blank & Davis, 2016), it is not identical due to differences in items. However, both analyses focus on contrasts over item pairs that differ in exactly one segment. The computational logic of the RSA is identical. However, Blank and Davis (2016) used CVC words that differ at the vowel, and we used CVCV words that differ at the second consonant. How important is it that our cohort pairs diverge at the second consonant position in a CVCV rather than the final consonant in a CVC or that our rhyme pairs overlap in three segments instead of two? Our view is that if the RSA analysis is capturing segment-level predictive coding mechanisms, it should not be critically dependent on the CVC vs. CVCV templates.

Finally, we note a feature of the rPCC model unrelated to testing whether it is consistent with the RSA crossover. It is notable that in a manner parallel to the lexical prediction paradigm results from Figure 7, we see RSA responses in the $\hat{r}$ and $\mathbf{e}$ components of rPCC that strongly match the timescales of phonemic and syllabic input peaks when we expand the averaged data from Figure 13 across all of its original time steps (see Figure 15). Of particular interest, the sublexical layer representation $\hat{r}_1$ changes over time in a way that closely matches phonemic input peaks (see vertical, dashed grey lines at time steps 21, 40, and 59), while the behavior of lexical layer representation $\hat{r}_2$ is more consistent with consonant input peaks alone (sublexical information; see time steps 21 and 59). Again, it appears as though the model has abstracted away from phonemic information to form functional units more closely matching syllables (still sublexical). This layer-dependent representational distinction is something we might expect in human neurobiology if hierarchical, neural PC were at play in speech perception, with progressively larger units encoded as we progress through the hierarchy.

5 Discussion

We began by extending a static predictive coding model (sPCC) to a classification task, which was prerequisite to training a recurrent classifier. We then built such a recurrent classifier (our rPCC model) using the classification layer approach tested in the sPCC model. We demonstrated that it could recognize a small lexicon of CVCV words, with time series of pseudospectra as input. When testing lexical competition in our model, we obtained the crucial empirical target previously demonstrated in a visual world paradigm study (Allopenna et al., 1998): human-like phonological competition over time, with early competition between onset competitors and later competition from rhymes, with both types of competitors more active than phonologically unrelated lexical items.

We then used the rPCC model to simulate two studies of human SWR (Gagnepain et al., 2012; Blank & Davis, 2016), to determine whether a

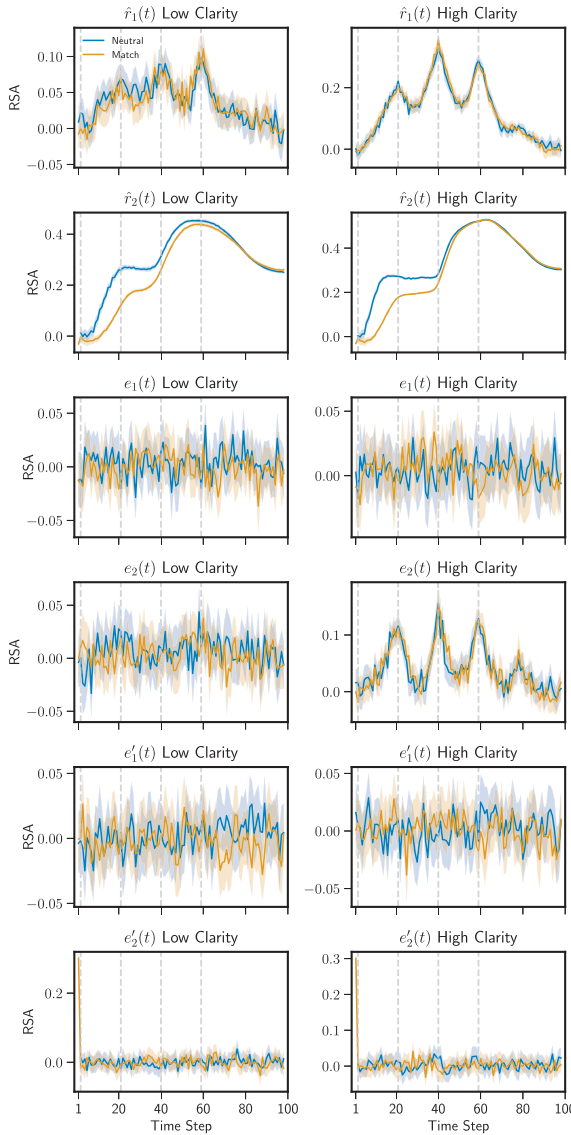


Figure 15: Effects of prior and clarity on simulated RSA, time steps 1–98, Linear. These results represent the RSA data from Figure 13 expanded across all original time steps, instead of averaged to render Figure 13’s bar heights, and confidence intervals (here, intervals are bands). Reinforcing the results from the lexical prediction paradigm (see Figure 8), we see strong responses in $\hat{r}$ and $\mathbf{e}$ components which match the timescale of phonemic and syllabic inputs (vertical, dashed grey lines at time steps 21, 40, 59).

plausible extension of PC to a recurrent form for over-time, speech-like signals would demonstrate effects that support or contradict proposed diagnostic hallmarks of PC. We began with the lexical prediction paradigm (Gagnepain et al., 2012). Given that greater signal responses to unexpected than to expected inputs have been widely accepted as a necessary consequence of PC, we assessed whether the rPCC model would produce stronger segmental prediction error responses to novel words than to trained words in some model components—and confirmed that it does. We also showed, using the same lexical prediction paradigm of Gagnepain et al. (2012), that the rPCC model shows anticipatory activation to upcoming phonemes, and thus engages in true predictive processing.

Using the same lexical prediction paradigm, we further examined the state-unit activations and information flows of the rPCC model in analogy to similar work with TRACE by Luthra et al. (2021). Notably, TRACE and rPCC both demonstrated reduced internal signals given expected versus unexpected inputs in model components that were not explicitly encoding prediction errors. Similar results between the two models suggest that (1) such signal reduction may not necessarily indicate prediction error, either in the brain or in a computational model, and/or that (2) the two model dynamics might be functionally equivalent despite the lack of explicit implementation of prediction error computation in TRACE. Of course, the extent to which TRACE and rPCC are similar will require further investigation (and complementary possibilities should also be considered, such as whether rPCC demonstrates signal sharpening in its state units).

Then, by introducing input noise and lexical priming, we were able to apply RSA on rPCC model components using a similar approach to that of Blank and Davis (2016). Contrary to Blank and Davis's prediction error model simulations (which was not a formal PC model), the rPCC model failed to reliably exhibit a putatively PC-diagnostic RSA crossover interaction analogous to the interaction Blank & Davis (2016) observed in human pSTS. The discrepancy between the rPCC simulations and empirical RSA result opens several interesting theoretical avenues. While we cannot rule out predictive processing accounts of the empirical results, our result indicates the possibility that the interaction observed in Blank & Davis is not a signature of predictive coding in speech recognition.

6 Conclusion

We have introduced a promising predictive coding model for human SWR that accepts time-varying inputs. In future studies we aim to expand the lexicon, adapt the model to take real speech as input, and, importantly, conduct further tests of behavioral and neural adequacy (i.e., comparing the model to more empirical human behavioral and neural data). It would also be useful to compare it to the TRACE model (McClelland & Elman, 1986) as a benchmark. Despite TRACE's age and rather unusual architecture, it

is consistent with a remarkable number of phenomena in human SWR (see Magnuson & Crinnion, 2022, who review a list of approximately 30 key aspects of SWR and speech perception that TRACE accommodates). If we wish to place a PC model of human SWR on the same plane as TRACE, it should be able to simulate the same key behavioral phenomena. Finally, we note that our results do not prove that there is no measurable quantity that will distinguish PC from IA architectures. Taken together with Luthra et al. (2021), we *have* shown that previous suggestions for such a diagnostic criterion are insufficient, but this does not rule out other as-yet-unknown model quantities or phenomena that may be able to do discriminate between the two computational paradigms.

Appendix

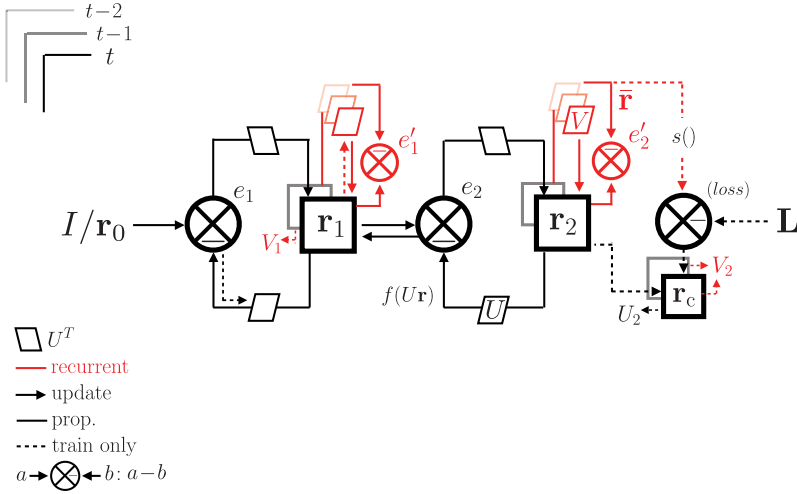


Figure A1: rPCC training and evaluation mechanics. Component sizes: $I(t)_{(21,)}; r_1_{(10,)}; r_2/r_c/L_{(12,)}; U_1_{(21,10)}; U_2_{(10,12)}; V_1_{(10,10)}; V_2_{(12,12)}; e_i/e'_{i((\text{nearest } r_i))}$. Terms are postupdate (“hatted” for $\hat{r}$, $\hat{U}$, $\hat{V}$) unless otherwise noted, and evaluated at time t , $t-1$, or $t-2$ according to the pattern in the upper left page corner. Activation function $f = id$. Red lines and text depict the recurrent components, connections, signal propagation, and updates, while black lines show the static components. *Error components*: The between-layer error component $e_i = \hat{r}_{i-1} - f(\hat{U}_i \hat{r}_i)$ represents the difference between a bottom-up signal and a top-down prediction; the within-layer error component $e'_i = \hat{r}_i - \bar{r}_i$ represents the difference between a hidden state after incorporating input at t and before incorporating that same input. In our two-layer model, $i = 1$ or $i = n = 2$, though the number of pre- n layers can be extended arbitrarily.

Acknowledgments

This research was supported by U.S. National Science Foundation awards BCS-PAC 1754284 (PI: JSM) and BCS-PAC 2043903 (PIs: JSM, KB), and by the Basque government through the BERC 2022-2025 program, and by the Spanish State Research Agency through Severo Ochoa excellence accreditation CEX2020-001010-S to BCBL, and award PID2023-149585NB-I00 (JSM, PI). All stimuli, data, and analysis scripts are available on GitHub. T.H. contributed to this project while at the University of Connecticut. We thank David Gow, Roeland Hancock, Emily Myers, and Fumiko Hoeft for comments that improved this work. B. Rogers personally thanks Yahweh, as well as Jean and Rich Josephson and ARCS Foundation Oregon for their generous Scholar Award, and OSU (both COE and CBEE) for graduate awards that assisted in conducting this research. Correspondence concerning this article should be addressed to Kevin Brown, Department of Pharmaceutical Sciences and School of Chemical, Biological, and Environmental Engineering, Oregon State University, Corvallis, OR, 97331, USA. E-mail: kevin.brown@oregonstate.edu.

References

- Allopenna, P. D., Magnuson, J. S., & Tannenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language*, 38(4), 419–439. 10.1006/jmla.1997.2558
- Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of a subsequent reference. *Cognition*, 73(3), 247–264. 10.1016/S0010-0277(99)00059-1
- Blank, H., & Davis, M. H. (2016). Prediction errors but not sharpened signals simulate multivoxel fMRI patterns during speech perception. *PLOS Biology*, 14(11), 1–33. 10.1371/journal.pbio.1002577
- Davis, M. H., & Sohoglu, E. (2020). Three functions of prediction error for Bayesian inference in speech perception. In M. Gazzaniga, R. Mangun, & D. Poeppel (Eds.), *The cognitive neurosciences*. MIT Press.
- DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, 8(8), 1117–1121. 10.1038/nn1504
- Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, 14, 179–211. 10.1207/s15516709cog1402_1
- Elman, J. L., & McClelland, J. L. (1988). Cognitive penetration of the mechanisms of perception: Compensation for coarticulation of lexically restored phonemes. *Journal of Memory and Language*, 27(2), 143–165. 10.1016/0749-596X(88)90071-X
- Forgy, E. W. (1965). Cluster analysis of multivariate data: Efficiency versus interpretability of classifications. *Biometrics*, 21(3), 768–780.
- Frauenfelder, U. H., & Peters, G. (1998). Simulating the time course of spoken word recognition: An analysis of lexical competition in TRACE. In J. Grainger &

- A. Jacobs (Eds.), *Localist connectionist approaches to human cognition* (pp. 101–146). Erlbaum.
- Friston, K. J. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1456), 815–836. 10.1098/rstb.2005.1622
- Friston, K. J., Sajid, N., Quiroga-Martinez, D. R., Parr, T., Price, C. J., & Holmes, E. (2021). Active listening. *Hearing Research*, 399, 107998. 10.1016/j.heares.2020.107998
- Gagnepain, P., Henson, R. N., & Davis, M. H. (2012). Temporal predictive codes for spoken words in auditory cortex. *Current Biology*, 22(7), 615–621. 10.1016/j.cub.2012.02.015
- Gaston, P., Rogers, B., Kadhim, K., Brown, K. S., & Magnuson, J. S. (N.d.). Amended prediction error and signal sharpening models in word recognition. In preparation.
- Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Transactions of the ASME—Journal of Basic Engineering*, 82(Series D), 35–45.
- Kriegeskorte, N. (2008). Representational similarity analysis: Connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2(November), 4. 10.3389/neuro.01.016.2008
- Kukona, A., Fang, S. Y., Aicher, K. A., Chen, H., & Magnuson, J. S. (2011). The time course of anticipatory constraint integration. *Cognition*, 119(1), 23–42. 10.1016/j.cognition.2010.12.002
- Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology*, 62, 621–647. 10.1146/annurev.psych.093008.131123
- Luce, R. D. (1959). *Individual choice behavior*. Wiley.
- Luthra, S., Li, M. Y. C., You, H., Brodbeck, C., & Magnuson, J. S. (2021). Does predictive processing imply predictive coding in models of spoken word recognition? *Psychonomic Bulletin and Review*. 10.3758/s13423-021-01924-x.
- Magnuson, J. S. (2025). TRACE-ing fixations in the Visual World Paradigm: Extending linking hypotheses and addressing individual differences by simulating trial-level behavior. *Brain Research*, 1856, 149563. 10.1016/j.brainres.2025.149563
- Magnuson, J. S., & Crinnion, A. M. (2022). Spoken word recognition. In J. Trueswell, A. Papafragou, & L. Gleitman (Eds.), *The Oxford handbook of the mental lexicon* (pp. 461–490). Oxford University Press.
- Magnuson, J. S., Crinnion, A. M., Luthra, S., Gaston, P., & Grubb, S. (2024). Contra assertions, feedback improves word recognition: How feedback and lateral inhibition sharpen signals over noise. *Cognition*, 242, 105661. 10.1016/j.cognition.2023.105661
- Magnuson, J. S., McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2003a). Lexical effects on compensation for coarticulation: The ghost of Christmas past. *Cognitive Science*, 27(2), 285–298. 10.1207/s15516709cog2702_6
- Magnuson, J. S., McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2003b). Lexical effects on compensation for coarticulation: A tale of two systems? *Cognitive Science*, 27(5), 801–805. 10.1207/s15516709cog2705_7
- Magnuson, J. S., Mirman, D., Luthra, S., Strauss, T., & Harris, H. D. (2018). Interaction in spoken word recognition models: Feedback helps. *Frontiers in Psychology*, 9(April), 369. 10.3389/fpsyg.2018.00369

- Magnuson, J. S., Mirman, D., Luthra, S., Strauss, T., & Harris, H. (2018). Interaction in spoken word recognition models: Feedback helps. *Frontiers in Psychology*, 9, 369. 10.3389/fpsyg.2018.00369
- Magnuson, J. S., Tanenhaus, M. K., Aslin, R. N., & Dahan, D. (2003). The time course of spoken word learning and recognition: Studies with artificial lexicons. *Journal of Experimental Psychology: General*, 132(2), 202–227. 10.1037/0096-3445.132.2.202
- Magnuson, J. S., You, H., Luthra, S., Li, M., Nam, H., Escabí, M., . . . Rueckl, J. G. (2020). EARSHOT: A minimal neural network model of incremental human speech recognition. *Cognitive Science*, 44(4), e12823. 10.1111/cogs.12823
- McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, 18(1), 1–86. 10.1016/0010-0285(86)90015-0
- McQueen, J. M. (2003). The ghost of Christmas future: Didn't Scrooge learn to be good? Commentary on Magnuson, McMurray, Tanenhaus, and Aslin (2003). *Cognitive Science*, 27(5), 795–799. 10.1207/s15516709cog2705_6
- McQueen, J. M., Jesse, A., & Norris, D. (2009). No lexical-prelexical feedback during speech perception or: Is it time to stop playing those Christmas tapes? *Journal of Memory and Language*, 61(1), 1–18. 10.1016/j.jml.2009.03.002
- McQueen, J. M., Norris, D., & Cutler, A. (2006). Are there really interactive processes in speech perception? *Trends in Cognitive Sciences*, 10(12), 533. 10.1016/j.tics.2006.10.004
- Norris, D., McQueen, J. M., & Cutler, A. (2000). Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences*, 23(3), 299–325. 10.1017/S0140525X00003241
- Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, 47(2), 204–238. 10.1016/S0010-0285(03)00006-9
- Norris, D., McQueen, J. M., & Cutler, A. (2016). Prediction, Bayesian inference and feedback in speech recognition. *Language, Cognition and Neuroscience*, 31(1), 4–18. 10.1080/23273798.2015.1081703
- Pitt, M. A., & McQueen, J. M. (1998). Is compensation for coarticulation mediated by the lexicon? *Journal of Memory and Language*, 39, 347–370. 10.1006/jmla.1998.2571
- Popal, H., Wang, Y., & Olson, I. R. (2020). A guide to representational similarity analysis for social neuroscience. *Social Cognitive and Affective Neuroscience*, 14(11), 1243–1253. 10.1093/scan/nsz099
- Rabovsky, M., Hansen, S. S., & McClelland, J. L. (2018). Modeling the N400 brain potential as change in a probabilistic representation of meaning. *Nature Human Behavior*, 2, 693–705. 10.1038/s41562-018-0406-4
- Rabovsky, M., & McClelland, J. L. (2020). Quasi-compositional mapping from form to meaning: A neural-network-based approach to capturing neural responses during language comprehension. *Philosophical Transactions of the Royal Society B*, 375, 20190313. 10.1098/rstb.2019.0313
- Rao, R. P. N. (1999). An optimal estimation approach to visual perception and learning. *Vision Research*, 39(11), 1963–1989. 10.1016/S0042-6989(98)00279-X
- Rao, R. P. N., & Ballard, D. H. (1997). Dynamic model of visual recognition predicts neural response properties in the visual cortex. *Neural Computation*, 9(4), 721–763. 10.1162/neco.1997.9.4.721

- Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1), 79–87. 10.1038/4580
- Rogers, B., Magnuson, J. S., & Brown, K. S. (N.d.) Object classification based on predictive coding. In preparation.
- Samuel, A. G., & Pitt, M. A. (2003). Lexical activation (and other factors) can mediate compensation for coarticulation. *Journal of Memory and Language*, 48(2), 416–434. 10.1016/S0749-596X(02)00514-4
- Sohoglu, E., & Davis, M. H. (2016). Perceptual learning of degraded speech by minimizing prediction error. *Proceedings of the National Academy of Sciences*, 113(12), E1747–E1756. 10.1073/pnas.1523266113
- Sohoglu, E., & Davis, M. H. (2020). Rapid computations of spectrotemporal prediction error support perception of degraded speech. *eLife*, 9, 1–25. 10.7554/eLife.58077
- Spratling, M. (2008). Reconciling predictive coding and biased competition models of cortical function. *Frontiers in Computational Neuroscience*, 2. 10.3389/neuro.10.004.2008.
- Taylor, W. (1953). Cloze procedure: A new tool for measuring readability. *Journalism Quarterly*, 30, 414–433. 10.1177/107769905303000401
- Vinh, N. X., Epps, J., & Bailey, J. (2009). Information theoretic measures for clusterings comparison. In *Proceedings of the 26th Annual International Conference on Machine Learning* (pp. 1073–1080).
- Yildiz, I. B., von Kriegstein, K., & Kiebel, S. J. (2013). From birdsong to human speech recognition: Bayesian inference on a hierarchy of nonlinear dynamical systems. *PLOS Computational Biology*, 9(9), e1003219. 10.1371/journal.pcbi.1003219

Received July 23, 2025; accepted May 14, 2026.