Phoneme restoration and empirical coverage of
Interactive Activation and Adaptive Resonance models
of human speech processing

Stephen Grossberg\textsuperscript{1} and Sohrob Kazerounian\textsuperscript{2}

\textsuperscript{1}Center for Adaptive Systems
Graduate Program in Cognitive and Neural Systems
Center for Computational Neuroscience and Neural Technology
Departments of Mathematics, Psychology, and Biomedical Engineering
Boston University
Boston, MA 02215

and

\textsuperscript{2}Nuance Communications, Inc.
1 Wayside Road
Burlington, MA 01803

November 23, 2015
Revised: March 9, 2016

All correspondence should be addressed to
Professor Stephen Grossberg
Center for Adaptive Systems, Room 205
Boston University
677 Beacon Street
Boston, MA 02215
Phone: 617-353-7858
Fax: 617-353-7755
Email:steve@bu.edu

Running Title: Interactive activation and adaptive resonance
Abstract

Magnuson (2015) makes claims for Interactive Activation (IA) models and against Adaptive Resonance Theory (ART) models of speech perception. Magnuson also presents simulations that claim to show that the TRACE model can simulate phonemic restoration, which was an explanatory target of the cARTWORD ART model. The theoretical analysis and review herein show that all these claims are incorrect. More generally, the TRACE and cARTWORD models illustrate two diametrically opposed types of neural models of speech and language. The TRACE model embodies core assumptions with no analog in known brain processes. The cARTWORD model defines a hierarchy of cortical processing regions whose networks embody cells in laminar cortical circuits, as part of the paradigm of Laminar Computing. cARTWORD further develops ART speech and language models that were introduced in the 1970s. It builds upon Item-Order-Rank working memories which activate learned list chunks that unitize sequences to represent phonemes, syllables, and words. Psychophysical and neurophysiological data support Item-Order-Rank mechanisms and contradict TRACE representations of time, temporal order, silence, and top-down processing that exhibit many anomalous properties, including hallucinations of non-occurring future phonemes. Computer simulations of the TRACE model are presented that demonstrate these failures.

Key Words: Interactive Activation; Adaptive Resonance Theory; TRACE; cARTWORD; phonemic restoration; speech perception; learning; long-term memory; working memory; Item-Order-Rank; competitive queuing; primacy gradient; bowed gradient; chunking; attention; competition; silence; time; temporal order; top-down processing; inhibition of return; Magical Number 7; Magical Number 4; prefrontal cortex; neural network; laminar cortical model; Laminar Computing
Several qualitatively different kinds of models are attempting to explain and predict data about speech and language processing. The JASA article of Magnuson (2015) espouses the TRACE model of McClelland and Elman (1986), which is a member of the family of Interactive Activation models that were introduced by McClelland and Rumelhart (1981). The cARTWORD model (Figure 1b; Grossberg and Kazerounian, 2011), also published in JASA, contributes to the family of Adaptive Resonance Theory, or ART, models of speech and language processing that was introduced by Grossberg (1978a, 1978b) and has been incrementally developed in a series of articles over the past 40 years (e.g., Ames and Grossberg, 2008; Boardman et al., 1999; Bradski, Carpenter, and Grossberg, 1994; Cohen and Grossberg, 1986; Cohen, Grossberg, and Wyse, 1995; Grossberg, 1984, 1986, 2003; Grossberg, Boardman, and Cohen, 1997; Grossberg, Govindarajan, Wyse, and Cohen, 2004; Grossberg and Myers, 2000; Grossberg and Pearson, 2008; Grossberg and Stone, 1986a, 1986b; Kazerounian and Grossberg, 2014).

The article by Magnuson (2015) is entirely devoted to a critique of cARTWORD and thus, by extension, the entire emerging ART theory of speech and language processing. Magnuson (2015) criticizes: (1) the inability of our model to represent repeated items in working memory; e.g., a list like ABACBD, (2) our claim that the representation in TRACE of temporal order information is not only biologically implausible but also contradicted by psychological and neurophysiological data, (3) our claim that the ability of top-down feedback in TRACE to activate target units, without bottom-up input support, is biologically incorrect and leads to serious computational problems, (4) the limited explanatory range of our model compared to that of TRACE, and (5) our explanation of why TRACE cannot simulate phonemic restoration, which was a key explanatory target of Grossberg and Kazerounian (2011), and of the TRACE simulations that Grossberg and Kazerounian (2011) carried out to demonstrate this failure.

Because all of Magnuson’s criticisms are scientifically incorrect, or at best misleading, we are here forced to rebut them. In so doing, our goal is also to provide useful information about how ART explains and predicts challenging psychological and neurobiological data about working
memory, speech perception, and language learning that are, in principle, outside the explanatory range of the TRACE model and its variants.

II. ART vs. TRACE: HOW ARE REPEATED ITEMS REPRESENTED IN WORKING MEMORY?

Magnuson (2015) writes in several places (pp. 1481, 1483, and 1990) that cARTWORD cannot represent sequences in working memory that have repeated elements. For example, he writes on p. 1481:

“Representing ordered sequences is a fundamental problem in neuroscience, and is particularly salient in the case of speech...models of speech processing must distinguish temporal orderings. Models must also distinguish repetitions of elements; the second /d/ in /dæd/ must be encoded as a second /d/ event, not just further evidence that /d/ has occurred. The same is true for word sequences, such as DOG EATS DOG...Only one model provides truly deep and broad coverage of phenomena in human speech perception and spoken word recognition while providing a basis for representing temporal order including repeated elements: the TRACE model.”
It is true that Grossberg and Kazerounian (2011) did not simulate acoustic sequences with repeated elements. The reason was simply that this was not an explanatory target of that article. However, the more general ART theory of speech and language perception does model how repeated elements can be stored in working memory (e.g., Braski, Carpenter, and Grossberg, 1994; Grossberg and Pearson, 2008; Silver et al., 2011), and quantitatively simulates neurophysiological data that support its Item-Order-Rank working memory model of how this is accomplished. In addition, the Item-and-Order working memory model (Figure 2) that was, for simplicity, simulated in Grossberg and Kazerounian (2011) can trivially be extended to an Item-Order-Rank working memory that is capable of storing sequences with repeated elements.

**Figure 2.** An Item-and-Order working memory for the short-term sequential storage of item sequences can activate a multiple-scale Masking Field list chunking network through a bottom-up adaptive filter. As in Figure 1b, the cognitive working memory uses two layers of cells with activities X and Y. The inputs are denoted by I. When embedded in a larger architecture like cARTWORD, these inputs are derived from item chunks, as in Figure 1b. Adaptive filter weights from the X activities to the Masking Field learn to selectively activate list chunks within the Masking Field. For simplicity, the Masking Field shows a single list chunk that receives one input (for the list “A”), two inputs (for the list “AC”), or three inputs (for the list “ABC”) from the cognitive working memory. The larger cell sizes and interaction strengths of the list chunks that categorize longer lists enable the Masking Field to choose the list chunk that currently receives the largest total input, and thus best predicts the sequence that is currently stored in the cognitive working memory. [Reprinted with permission from Kazerounian and Grossberg (2014).]
Magnuson (2015) is aware that other ART working memory models have indeed simulated the short-term storage of sequences with repeated elements. He criticizes this fact in two ways: (1) Those demonstrations do not apply to speech and language, and (2) Item-Order-Rank working memories exhibit the same kinds of problems in representing temporal order that TRACE faces. However, both of these assertions are incorrect.

In partial response to the first concern: The ART Item-and-Order model in Grossberg and Pearson (2008) quantitatively simulated psychophysical data about linguistic working memory in humans using the same working memory model that was used to quantitatively simulate neurophysiological data about motor working memory in monkeys. An Item-Order-Rank working memory was used by Silver et al. (2011) to simulate neurophysiological data about the role of spatial working memories in the learning and control of saccadic eye movement sequences by monkeys; see Figure 7 below.

Why should a similar kind of working memory circuit be used for linguistic, motor, and spatial working memories? To understand this property, and also why the ART and TRACE mechanisms of temporal order are fundamentally different, one needs to review how Item-and-Order working memories are designed, how they are naturally extended to Item-Order-Rank working memories, and why a similar recurrent shunting on-center off-surround network design, properly regulated by rehearsal and inhibition-of-return mechanisms, is used to represent all linguistic, motor, and spatial working memories. Indeed, if an Item-Order-Rank working memory was used in the Grossberg and Kazerounian (2011) simulations, instead of an Item-and-Order working memory, it would have yielded the identical results, since the inputs used in these simulations included no repeated phonemes. Why this is so is explained in Section III.L.

It should be noted in advance that none of the properties of ART working memories, and of the psychological and neurophysiological data that support them, can be explained by the TRACE mechanism of sequence representation. Indeed, these data contradict the key TRACE hypotheses. Why this is so can explained in IV after the following summary of key properties of ART working memories and some of the data that support them.

III. ART WORKING MEMORY AND LIST CHUNKING

A. The predicted link between working memory and list chunking

The Grossberg (1978a, 1978b) neural model of working memory (WM) posits that a temporal stream of inputs is stored through time as an evolving spatial pattern of content-addressable item
representations. These WM patterns are, in turn, unitized through learning into list chunk representations that can control context-sensitive behaviors. This WM model is called an Item-and-Order model because, in it, individual nodes, or cell populations, represent list items, and the temporal order in which the items are presented is stored by an activity gradient across the nodes.

The classical work of Miller (1956) on the Magical Number Seven showed that a key functional unit in speech and language is abstract, namely the “chunk”, that “the memory span is a fixed number of chunks [and] we can increase the number of bits of information that it contains simply by building larger and larger chunks, each chunk containing more information than before.” Chunks can thus be learned from multiple types of acoustic inputs that vary in size. Item-and-Order models like cARTWORD extend the classical work of Miller (1956) on chunks by defining the functional units that are proposed to exist at successive levels of the brain’s speech and language hierarchy. Instead of levels that process phonemes, letters, and words (e.g., McClelland and Rumelhart, 1981), Item-and-Order model levels represent distributed features, item chunks, and list chunks (Grossberg, 1978a, 1978b, 1984, 1986). An item chunk selectively responds to prescribed patterns of activity across the distributed feature detectors within a prescribed time interval (e.g., a phoneme). A list chunk selectively responds to prescribed sequences of item chunks that are stored in working memory. The properties of these functional units can explain data about word superiority effect, list length effect, and related speech phenomena that are incompatible with alternative processing levels; see below and Sections IV-VI.

B. Correct temporal order is stored temporarily in the brain by a primacy gradient

A primacy gradient stores items in WM in the correct temporal order. In a primacy gradient, the first item in the sequence activates the corresponding item chunk with the highest activity, the item chunk representing the second item has the second highest activity, and so on, until all items in the sequence are represented. For example, a sequence “1-2-3” of items is transformed into a primacy gradient of activity with cells encoding ‘1’ having the highest activity, cells encoding ‘2’ with the second highest activity, and cells encoding ‘3’ having the least activity. Item-and-Order working memories can, in a similar way, easily store sequences composed of the same items presented in different orderings.
C. Rehearsal and inhibition-of-return

How is a stored spatial pattern in WM used to recall a sequence of items performed through time? A rehearsal wave that is delivered uniformly, or non-specifically, from the basal ganglia to the entire WM enables read-out of stored activities (Figure 3).

**Figure 3.** A temporal sequence of inputs creates a spatial pattern of activity across item chunks in an Item-and-Order working memory (height of hatched rectangles is proportional to cell activity). Relative activity level codes for item and order. A rehearsal wave allows item activations to compete before the maximally active item elicits an output signal and self-inhibits via feedback inhibition to prevent its perseverative performance. The process then repeats itself. Solid arrows denote excitatory connections. Dashed arrows denote inhibitory connections. [Adapted from Grossberg, 1978a].

3). The node with the highest activity is read out fastest and self-inhibits its WM representation. Self-inhibition of the item that is currently being read out helps to explain the cognitive concept of *inhibition-of-return*, which prevents perseveration on the most recent item to be performed. This self-inhibition process is repeated until the entire sequence is reproduced in its correct order and there are no active nodes left in the WM. How different rehearsal strategies may depend on experimental conditions, such as during immediate free recall vs. immediate serial recall, is discussed in Grossberg and Pearson (2008).

D. Competitive queuing and primacy models

Since Grossberg (1978a, 1978b) introduced the Item-and-Order model, many modelers have used it and variations thereof (e.g., Boardman and Bullock, 1991; Bohland, Bullock, and
Guenther, 2010; Bradski, Carpenter, and Grossberg, 1994; Bullock and Rhodes, 2003; Grossberg and Pearson, 2008; Houghton, 1990; Page and Norris, 1998). For example, Page and Norris (1998) used a Primacy Model to explain and simulate cognitive data about immediate serial order working memory, notably experimental properties of word and list length, phonological similarity, and forward and backward recall effects. The Item-and-Order model is also known as the Competitive Queuing model (Houghton, 1990).

E. Psychological and neurophysiological data confirm Item-and-Order predictions

Both psychophysical and neurophysiological data have supported the Grossberg (1978a, 1978b) predictions that neural ensembles encode item order with relative activity levels, and are reset by self-inhibition. For example, Farrell and Lewandowsky (2004) did psychophysical experiments in humans that studied the latency of responses following serial performance errors. They concluded that (p. 115):

“Several competing theories of short-term memory can explain serial recall performance at a quantitative level. However, most theories to date have not been applied to the accompanying pattern of response latencies...Data from three experiments...rule out three of the four representational mechanisms. The data support the notion that serial order is represented by a primacy gradient that is accompanied by suppression of recalled items [italics ours].”

Electrophysiological experiments have directly demonstrated these predicted properties. For example, macaque monkeys stored primacy gradients in their dorsolateral prefrontal cortex to control their performance of arm movement sequences that copy geometrical shapes (e.g., Averbeck et al., 2002). The predicted properties of a primacy gradient and a self-inhibitory form of inhibition-of-return are evident in these data (Figure 4a), which were simulated (Figure 4b) by a motor Item-and-Order working memory in the laminar cortical LIST PARSE model of Grossberg and Pearson (2008) that is a precursor of cARTWORD. An Item-Order-Rank spatial working memory in the lisTELOS model (Silver et al., 2011) was used to simulate neurophysiological data (Histed and Miller, 2006) about how microstimulation changes a stored primacy gradient, and thus the order of sequential saccadic eye movement performance. These examples illustrate that the Item-and-Order model predicted the kind of working memory representation that occurs in mammalian brains more than 20 years before it was supported by neurophysiological experiments.

F. Bowed gradients during free recall

Item-and-Order working memories have been used to explain and predict many types of data about temporal order and how it is unitized through learning by list chunks. A key theme in this
development surrounds the question: What is the longest list that the brain can store in working memory in the correct temporal order? Why can only relatively short lists be stored with the correct temporal order \textit{in vivo}? In an Item-and-Order working memory, this question translates into: What is the longest primacy gradient that the working memory can store? And why is it so short? Indeed, in free recall tasks, if too long a list

\begin{figure}
\centering
\includegraphics[width=\textwidth]{figure4.png}
\caption{Neurophysiological data and simulations of monkey sequential copying data. (a) Each plot shows the relative strength of representation to control drawing of each segment for each time bin (at 25ms) of the task. The number of movement segments is due to the starting positions of each movement sequence on the corresponding geometrical figure. Time 0 indicates the onset of the template. Lengths of segments were normalized to permit averaging across trials. Plots show parallel representation of segments before initiation of copying. Further, rank order of}
strength of representation before copying corresponds to the serial position of the segment in the series. The rank order evolves during the drawing to maintain the serial position code. At least four phases of the Averbeck et al. (2002; Figure 9a) curves should be noted: (1) presence of a primacy gradient; that is, greater relative activation corresponds to earlier eventual execution in the sequence during the period prior to the initiation of the movement sequence (period -500 to 400ms); (2) contrast enhancement of the primacy gradient to favor the item to be performed (greater proportional representation of the first item) prior to first item performance (period ~100 to 400ms); (3) inhibition of the chosen item’s activity just prior to its performance and preferential relative enhancement of the representation of the next item to be preformed such that it becomes the most active item prior to its execution (period ~400ms to near sequence completion); and (4) possible re-establishment of the gradient just prior to task completion.

[Reprinted with permission from Averbeck et al., 2002.] (b) Simulations of item activity across the motor plan field of the LIST PARSE model for 3, 4, and 5 item sequences vs. simulation time. In both (a) and (b), line colors correspond to representations of segments as follows: yellow, segment 1; green, segment 2; red, segment 3; cyan, segment 4; magenta, segment 5.

[Reprinted with permission from Grossberg and Pearson, 2008.]

is presented, a bowed serial position curve is often observed, such that items at the beginning and the end of the list are performed earliest, and with the highest probability of recall (Figure 5).

Figure 5. The simulation to the left shows a primacy gradient of activation that is stored in working memory in response to presenting a list of seven items. The simulation to the right shows how this primacy gradient becomes a bowed gradient when more items are presented. Note, in addition, that the activities of the stored items in response to the longer list are smaller due to the self-normalizing network competition, which realizes the Normalization Rule.

[Reprinted with permission from Grossberg and Pearson (2008).]

Grossberg (1978a, 1978b) noted that these free recall properties have a natural explanation if the working memory gradient that stores the list items is also bowed, with the first and last items having the largest activities, and items in the middle having less activity. If the item with the
largest activity is read out first, whether at the list beginning or end, and then self-inhibits its item representation to prevent preservation, then the next largest item will be read out, and so on in the order of item relative activity. The greater probability of items being recalled at the beginning and end of the list also has a simple explanation, since items that are stored with larger activities have greater resilience against perturbation by cellular noise. Transpositions of order recall are also easily explained, because they belong to items with similar stored activities.

G. Magical numbers four and seven: Immediate and transient memory spans

What is the longest primacy gradient that can be stored? The classical Magical Number Seven, or immediate memory span, of 7 ± 2 items that is found during free recall (Miller, 1956) estimates the upper bound. Grossberg (1978a) distinguished between the immediate memory span and the then new concept of transient memory span. The transient memory span was predicted to be the result of recall from short-term working memory without the benefit of top-down read-out of learned expectations from list chunks. That is, the transient memory span is the longest list for which a primacy gradient may be stored in short-term memory solely as the result of bottom-up inputs. In contrast, the immediate memory span was predicted to arise from the combined effect of bottom-up inputs and top-down long-term memory read-out. Grossberg (1978a) proved that the read-out of top-down long-term memories can only increase the maximal primacy gradient that can be stored, and thus that the immediate memory span is longer than the transient memory span. Given an estimated immediate memory span of approximately seven items, it was estimated that the transient memory span should be approximately four items. Cowan (2001) has beautifully summarized data showing that, when the influences of long-term memory and grouping effects are minimized, there is indeed a working memory capacity limit of 4 ± 1 items. There is thus also a Magical Number Four, as predicted. As will be discussed in Section V, neither Magical Number can be explained by the TRACE model because of how it represents temporal order information.

H. LTM Invariance Principle: Designing working memory to learn stable list chunks

Why is the transient memory span so short? To explain this, Grossberg (1978a, 1978b) noted that it is pointless to store sequences temporarily in short-term working memory if they cannot lead to learning of unitized list chunks. And without such learning, it would be impossible to learn language, motor skills, or spatially organized movements or navigation routes. Grossberg (1978a, 1978b) therefore predicted that all working memories for the short-term storage of items
are designed to enable learning and stable memory of list chunks, and showed that two simple postulates imply these properties: the LTM Invariance Principle and the Normalization rule. Item-and-Order working memories were derived from these postulates. Grossberg (1978a, 1978b) then proceeded to demonstrate mathematically how these postulates can be realized, and thereby derived laws for Item-and-Order working memories, and how they generate primacy and bowed gradients to help explain free recall data. Since this early derivation, the understanding of how these working memories are realized in vivo has been incrementally refined (e.g., Bradski, Carpenter, and Grossberg, 1992, 1994), leading most recently to laminar cortical models of how prefrontal circuits realize Item-Order-Rank working memories (Grossberg and Pearson, 2008; Silver et al., 2011).

The LTM Invariance Principle implies that novel sequences of items may be stored and chunked through learning in a way that does not destabilize memories of previously learned chunk subsequences. Without such a property, longer chunks (e.g., for MYSELF) could not be stored in short-term working memory without risking the catastrophic forgetting of previously learned memories of shorter chunks (e.g., for MY, SELF, and ELF). Language, motor, and spatial sequential skills would then be impossible. The LTM Invariance Principle insists that, if bottom-up inputs activate a familiar subset chunk, such as the word MY, then the arrival of the remaining portion SELF of the novel word MYSELF during the next time interval will not erode the previously learned weights that activate the list chunk of MY. This principle is achieved mathematically by preserving the relative activities, or ratios, between previously stored working memory activities as new items are presented through time. Newly arriving inputs may, however, alter the total activity of each active cell across the working memory.

How does preserving activity ratios help to stabilize previously learned categories? These activities send signals to the next processing stage, where the category cells are activated. The signals are multiplied by adaptive weights, or LTM traces, before the net signals activate their target categories (Figure 2). The total input to a category thus multiplies a pattern, or vector, of activities times a pattern, or vector, of LTM traces. By preserving relative activities, the relative sizes of these total inputs to the category cells do not change through time, and thus nor do the corresponding LTM patterns that track these activities when learning occurs at their category cells.
Consider for example, what happens as bottom-up acoustic inputs arrive in time, activating their corresponding chunked (word) representations. As these inputs arrive, a chunk such as 'MY' may become active once it receives all or most of its expected bottom-up input. If the acoustic inputs are then followed immediately by silence, the chunked representation of 'MY' could stably learn from the stored STM pattern of activity that first supported it. On the other hand, as is often the case, the acoustic inputs might not simply be followed by silence, but rather by further acoustic information (e.g., the inputs corresponding to the super-set word or chunk 'MYSELF'). In this case, if the newly arriving inputs could drastically alter the pattern of activation reverberating in STM if the LTM Invariance Principle did not hold. As a result, the chunked representation for 'MY' would begin to degrade as the weights to 'MY' change in response to the now altered STM pattern in working memory. If however, the newly arriving inputs (corresponding to 'SELF') leave intact the relative pattern of activity in STM of the already occurring acoustic inputs (corresponding to 'MY'), a new chunk for the full superset word (in this case, 'MYSELF') could be learned without destabilizing the already learned LTM pattern for its subset components (e.g., 'MY').

The Normalization Rule insists that, when a working memory is activated, its total activity is approximately constant through time, and tends to be independent of the number of items stored in working memory. Thus, if more items are stored in working memory, then each item tends to be stored with less activity (see Figures 4 and 5). This normalization property implies the familiar limited capacity of working memory by redistributing, rather than simply adding, activity when new items are stored.

**I. Bowed gradients for long lists follow from self-stabilizing memory**

Grossberg (1978a, 1978b) mathematically proved that, if both the LTM Invariance Principle and the Normalization Rule hold in a working memory, then there is a transient memory span; that is, lists no longer than the transient memory span can be stored as a primacy gradient and thus recalled in their correct temporal order. If a list is longer than the transient memory span, the primacy gradient that is initially stored will evolve into a bowed gradient as more items are stored.

In other words, the ability of a working memory to enable learning and stable memory of stored sequences implies an upper bound on the length of lists that can be temporarily stored in the correct temporal order. The bowed serial position curves of free recall data could then be
understood as the price paid for being able to rapidly learn, and stably remember, language and sequential spatial and motor skills.

These results hold when the same amount of attention is paid to each item as it is stored. If attention is not uniform across items, then multi-modal bowing can occur, as during von Restorff (1933) effects, also called isolation effects (Hunt and Lamb, 2001), which occur when an item in a list “stands out like a sore thumb” and is thus more likely to be remembered than other list items.

**J. Universal design for linguistic, spatial, and motor working memories**

If all working memories obey these postulates, then all linguistic, motor, and spatial working memories should have a similar design. Psychological and neurobiological data have supported this prediction, as reviewed in Grossberg and Pearson (2008), Silver et al. (2011), and Grossberg (2013). Such data exhibit similar data patterns across modalities, including bowing effects on performance order and error probabilities. In particular, the LIST PARSE model of Grossberg and Pearson (2008) used a prefrontal *linguistic* working memory to quantitatively simulate psychophysical data about immediate serial recall, and immediate, delayed, and continuous distractor free recall; and a similarly designed prefrontal *motor* working memory to quantitatively simulate neurophysiological data about sequential recall of stored motor sequences (Figure 4). The listTELOS model of Silver et al. (2011) used a prefrontal *spatial* working memory to quantitatively simulate neurophysiological data about the learning and planned performance of saccadic eye movement sequences.

Such results provide accumulating evidence for the prediction that all working memories have a similar design because they all need to obey the LTM Invariance Principle. List chunks in all these modalities can then be learned and stably remembered, and cross-modality interactions of such working memories can occur, since they all obey variations of the same circuit design. Because of this shared design, it becomes easier to understand how language in young children can begin to develop in a way that parallels the motor behaviors of adult teachers during mutual play (Bruner, 1975), or how sign language by hearing adults can coordinate signing with speaking (Neville et al., 2002).

It remains to answer the nagging question: What is this shared design?

**K. Recurrent shunting on-center off-surround networks embody working memories**
Are postulates such as the LTM Invariance Principle and the Normalization Rule too sophisticated to be discovered by evolution? This type of concern arises whenever one confronts the variety of intelligent brain competences, whether it is about the origins of working memory, numerical representation, or handwriting. This concern was allayed by the demonstration that both the LTM Invariance Principle and the Normalization Rule occur within a ubiquitous neural design; namely, a recurrent on-center off-surround network of cells that obey the membrane equations of neurophysiology, otherwise called shunting dynamics. How such recurrent shunting networks process ratios (LTM Invariance Principle) and conserve total activity (Normalization Rule) was mathematically proved in Grossberg (1973) and reviewed in Grossberg (1978a, 1980). Bradski, Carpenter, and Grossberg (1992, 1994) went on to prove theorems about how Item-and-Order recurrent shunting working memory networks generate primacy and bowed gradients, among other properties, as a function of network parameters.

In brief, the excitatory feedback due to the recurrent on-center interactions in such a network helps to store an evolving spatial pattern of activities in response to a sequence of inputs. The recurrent shunting off-surround, in concert with the on-center, helps to preserve the relative activities that are stored. A volitional rehearsal wave from the basal ganglia enables the highest stored activity to be read out first, and self-inhibitory feedback prevents perseverative performance of this most highly activated cell population, thereby enabling less active populations to be performed (Figure 3), while the network as a whole gradually renormalizes its activity through time.

The effects of recurrent inhibition are evident in the data and simulation within Figure 4: After the next-to-last item is performed, the population storing the last item is disinhibited and reaches the highest activity through time of any population that stored the list.

L. Storing lists with multiple item repetitions: Item-Order-Rank coding

In its simplest form, an Item-and-Order working memory does not represent the same item in multiple positions, or ranks, of a list. However, humans can easily do this, and there are many examples in cognitive data of sensitivity to list position (e.g., Henson, 1998), including spoonerisms, wherein phonemes or syllables in similar positions in different words are selectively interchanged; e.g., “hissed my mystery lesson”. It is also known that the activity of some neurons in prefrontal cortex for a given list item is sensitive to the rank of that item within the sequence (e.g., Averbeck et al., 2003a; Barone and Jacobs, 1989; Funahashi et al., 1997;
Inoue and Mikami, 2006; Kermadi and Joseph, 1995; Ninokura et al. 2004). Error data in human serial recall experiments also indicate that rank information is available, which some models of serial recall have incorporated (see Grossberg and Pearson (2008) for a review).

Despite some positive results from rank-based models, Farrell and Lewandowsky (2004) have, as noted above, shown that latency data from error trials can be best explained by models that use a primacy gradient and self-inhibition (i.e., Item-and-Order models), but not by those that use rank alone. Some Item-and-Order models incorporated rank information (e.g., Bohland et al., 2010; Bradski et al., 1994). Indeed, Bradski et al. (1994) proposed the first Item-Order-Rank working memory model that can incorporate rank-order coding into an Item-and-Order working memory to represent item repeats at arbitrary list positions; e.g., ABACBD.

The LIST PARSE model (Grossberg and Pearson (2008, Figure 18) deepened understanding of where such rank order coding may arise in the brain, and how it gets represented in working memory. This model predicted how an Item-Order-Rank working memory can be created in prefrontal cortex by deriving its rank selectivity from the analog spatial representations of numbers in the parietal cortex via parietal-prefrontal projections. This prediction built upon the Spatial Number Network, or SpaN, model of Grossberg and Repin (2003) which simulated how the known analog map of ordered numerical representations in inferior parietal cortex may control the ability of animals and humans to estimate and compare sufficiently small numerical quantities. The predicted properties of SpaN model parietal neurons were supported by neurophysiological data of Nieder and Miller (2004), who also studied the prefrontal projections of these parietal numerical representations.

In such an Item-Order-Rank working memory, a spatial gradient of activity still represents temporal order, with the most active cell population being performed first. To enable the storage of the same item at multiple list positions in this gradient, the parietal-prefrontal projection of the analog spatial map of parietal numerical representations embeds numerical hypercolumns into the prefrontal working memory, so that each item is stored in a different position in its hypercolumn if it is repeated in the list more than once (Figure 6). A single numerical hypercolumn that represents a particular list item can store that item in multiple list positions, just as a positional hypercolumn in the visual cortical map of the primary visual cortex can selectively respond to multiple orientations at that position (Hubel and Wiesel, 1962, 1963). For example, to store and perform in its correct order the short list ABAC, item A would be stored in
two different positions within its hypercolumn, whereas items B and C would be stored only in one position in their respective hypercolumns. A primacy gradient of activity would still represent the temporal order of a short stored list, whether or not it had repeated items. Davis (2010) has proposed a related concept to model letter repetitions during visual word identification.

The recurrent on-center off-surround network that stores items in an Item-Order-Rank working memory can still have the same simple anatomy as it does for an Item-and-Order working memory that does not store repeats: self-excitatory feedback from each cell population

![Diagram](image)

**Figure 6.** How inputs from the analog number field that is found in the parietal cortex can generate rank-sensitive inputs (see numbers 1, 2, 3, and 4) to a prefrontal Item-and-Order working memory that convert it into an Item-Order-Rank working memory which can store the same item at multiple positions in a list. Each circle in the sensory working memory represents a different item. The numerical hypercolumn for each item has, in this example, four cell populations that can be activated by the parietal numerical map after 1, 2, 3, or 4 items has been presented. The maximum number 4 was chosen for ease of exposition. In response to a sequence of sensory inputs, an integrator cell population increases its activity proportionally and broadcasts this activity to the entire parietal numerical map. The map responds by shifting its locus of maximal activity to the right as larger numbers of inputs occur (see Grossberg and Repin (2003) for an explanation of how this is proposed to happen). Each parietal locus projects to a corresponding position in multiple prefrontal numerical hypercolumns. A prefrontal cell can fire only if it receives an item input and a numerical input. Thus, in response to a list ABA, the item representation for A will be activated in hypercolumn slots 1 and 3, and the item representation
for B will be activated in hypercolumn slot 2. A primacy gradient will develop over these three active item representations. [Reprinted with permission from Grossberg and Pearson (2008).]

to itself, and a broad off-surround that equally inhibits all other populations in the working memory.

**M. Simulating Item-Order-Rank working memory cells in prefrontal cortex**

The lisTELOS model of Silver et al. (2011) implemented the Grossberg and Pearson (2008) proposal by simulating an Item-Order-Rank model of spatial working memory in prefrontal cortex, and its interactions with other brain regions, to control planning, working memory storage, and execution of saccadic eye movement sequences. The model predicts and simulates how the supplementary eye fields (SEF) may select saccades from sequences that are stored in this prefrontal working memory. It also predicts and simulates how SEF may interact with downstream regions such as the frontal eye fields during memory-guided sequential saccade tasks, and how the basal ganglia may control the flow of information through time. Model simulations reproduce behavioral, anatomical and electrophysiological data under multiple experimental paradigms, including visually- and memory-guided single and sequential saccade tasks, and behavioral data from SEF microstimulation paradigms. In particular, lisTELOS simulates neurophysiological properties of rank-sensitive working memory cells in monkey SEF, thereby clarifying how Item-Order-Rank working memories store sequences of repeated target positions in brain spatial working memories (Figure 7).
Figure 7. Left two columns: SEF rank-direction cell that responds phasically before rightward saccades that occur at the third ordinal position in a sequence of saccades. Each row, from left to right, shows the cell’s responses when the first, second, and third saccades occur. Symbols U, L, and R denote upward, left, and right saccades. The cell responds vigorously only when the third saccade in the sequence moves to the right, as summarized in the left column and last two rows, and the right column and top row. [Data adapted with permission from Isoda and Tanji (2002).] See figure caption at the end of the article for mathematical symbols and citation.

Given that all working memories have a similar network design in order to realize the LTM Invariance Principle and Normalization Rule, the Item-Order-Rank working memory of the lisTELOS model is a prototype for linguistic and motor working memories, no less than for spatial working memories. Indeed, the basal ganglia play a gating function in cARTWORD (Figure 1b) that is similar to its role in lisTELOS.

The above summary thus shows that ART models of temporal order information can, and have, simulated challenging data about the storage of lists with repeated items in linguistic, motor, and spatial working memories. Moreover, these Item-Order-Rank working memories embody design principles about how the brain can stably learn list chunks, such as syllables and words, because of the way that these working memories are designed. None of these principles, mechanisms, or data can, in principle, be explained by TRACE.

IV. REBUTTING MAGNUSON’S CLAIMS AGAINST cARTWORD AND FOR TRACE

With this background of concepts, data, and their explanations, we can now respond to various of Magnuson’s criticisms of cARTWORD and, by extension, the ART family of speech and language models to which it contributes, as well as his claims for the TRACE model. Section V will present additional ART design principles and mechanisms that have been used to explain other data which cannot, in principle, be explained by TRACE. This review will, in particular, explain how ART can represent and store sequences with repeated words, thereby contradicting another criticism of Magnuson.

A. Claim that cARTWORD cannot, in principle, store repeated items in working memory

Magnuson asserts in multiple places that:

“First and most crucially, cARTWORD can represent sequences, but cannot [italics his] represent sequences that contain repeated elements…This rules out cARTWORD as a plausible model of sequence encoding for word recognition” (p. 1483).

The cARTWORD simulations of Grossberg and Kazerounian (2011) did not include repeated elements because that was not the explanatory goal of this article. However, because
cARTWORD uses an Item-and-Order working memory, it can easily be extended to include rank-sensitivity in an Item-Order-Rank working memory to simulate speech and language data where rank sensitivity is needed, without undermining any results that are derived where it is not, as was explained in Section III.L.

If Magnuson’s claim about the rank-sensitivity of cARTWORD is to be taken literally (e.g., “cARTWORD cannot represent sequences that contain repeated elements”), then it must be interpreted as a claim that, in principle, such an extension of cARTWORD is impossible. That claim is false.

**B. Claim that IOR working memory needs time-specific slices and slice-specific inhibition**

Magnuson does discuss the Silver et al. (2011) use of an Item-Order-Rank working memory on p. 1483:

“...it is not clear how to extend it sufficiently for speech, where each IOR [Item-Order-Rank] node might need to be capable of coding dozens of time steps (again reflecting perhaps the duration of echoic memory). Furthermore...time-specific inhibitory connectivity between IOR nodes would likely be required, as in TRACE. Indeed, extending the IOR approach in these ways would result in a mechanism for encoding temporal sequences not terribly different from that employed by TRACE, since each ‘pie slice’ in an IOR phoneme node because a time-specific representation of that phoneme—exactly the reduplication problem cARTWORD claims to avoid. Similar problems necessarily arise at supra-phonemic levels of encoding (cARTWORD’s list chunks and lexical nodes); repeated words will be as problematic as repeated phonemes, and the IOR framework must again be extended to include many time-specific representations for words.”

These claims are also mistaken. To understand Magnuson’s misunderstanding, one needs to realize that TRACE creates a new “slice” that reduplicates each phonetic item at every time. If 1000, or 10,000, time steps have occurred since a phonetic item occurred just once, then there are 1000, or 10,000, duplications of this phonetic item in TRACE. This representation is physically implausible and would cause a major combinatorial explosion when dealing with natural speech. Such a representation also cannot support learning of speech in real time, and has no neurobiological evidence to support it. Indeed, because of this reduplication property, TRACE has no representation of real time, as explained below and in Section VII.D.

In contrast, there is now abundant psychological and neurobiological evidence of Item-Order-Rank coding, some of it summarized in Section III. There is no replication of each phonemic item through time in an Item-Order-Rank working memory. Its unique content-addressable item chunks gets activated just once. If a phonemic item occurs more than once in a sequence, then its unique rank-sensitive representations get activated just once, as
neurophysiological data from prefrontal cortex has shown; e.g., Figure 7. There is no need for “time-specific inhibitory connectivity between IOR nodes”. Instead, as noted in Section III.L, a uniform off-surround from every cell to all other cells is sufficient, whether or not there is rank-sensitive coding. These off-surrounds respond in real time to whatever combination of item chunks is activated by a particular input sequence.

Magnuson’s discussion seems to conflate the passage of real time with “time steps”, as in his phrase “coding dozens of time steps”. This is indeed a property, and a problem, of the TRACE model because there is no evolution of real time in TRACE, since it formally creates new slices to represent each new sequential input. In contrast, an Item-Order-Rank working memory represents “temporal order”, not “time steps”. When one considers temporal order information, which is studied ubiquitously in the cognitive psychology of language, motor control, and space, one needs to explain the Magical Numbers Four and Seven, which the Item- and-Order framework explains in a principled way, and indeed predicted the Magical Number 4 and its explanation (Section III.G). These Magical Numbers, and the bowed gradients that occur when they are exceeded, have no natural explanation within TRACE’s myriad of artificially created “slices”.

C. TRACE is incompatible with psychophysical and neurobiological data

Magnuson notes (p. 1482) that “McClelland et al. (2014) remind us that TRACE is not meant to provide a neural-level solution. ‘The structure of the TRACE model should not be viewed as a literal claim about the neural mechanism.’” However, TRACE’s way of representing sequences is directly contradicted by psychophysical and neurobiological data about sequence representation, some of it summarized above. These data also directly support an Item-Order-Rank model of working memory, and also support the prediction of the LTM Invariance Principle that there exists an intimate linkage between mechanisms for storing sequences in short-term working memory and for learning list chunks of these sequences. A nice example of this linkage between STM and LTM during the learning of novel arm movement sequences is provided by human psychophysical data of Agam et al. (2007).

If the foundational hypotheses of sequence representation in TRACE are wrong, then there is no logical reason to believe any conclusion that is derived from them. Any resemblance between data and a TRACE simulation must therefore be interpreted as a simulation of weak model-independent properties of the data, ones that do not constrain, or explain, how the brain
actually works. Indeed, Magnuson et al. (2012) himself has written about the importance of such a model failure:

“[M]odel success or failure can be linked to one of four levels of decreasing importance: theory, parameters, or linking hypotheses. As we have just discussed, a “failure” or “success” due to improper linking hypotheses is not informative in the same way that an experimental failure due to improper operational definitions is not informative. A failure at the level of theoretical assumptions is of greatest interest and holds the greatest possibility for progress (i.e., theory falsification).”

Magnuson (p. 1482) goes on to write that “while TRACE can be fairly criticized for its reduplication mechanism, it is not wildly implausible”. In reality, its item replication mechanism is not just implausible; it has been directly disconfirmed by neurophysiological evidence. To support the claim that TRACE is “not wildly implausible”, Magnuson asks the reader to:

“consider a model of echoic memory based on a frequency-by-time matrix (perhaps 1 to 4 seconds in duration, the approximate duration of echoic memory for speech; Connine et al., 1991; Watkins and Watkins, 1980), with the simplifying assumption that time is discretized into steps. Now as auditory input is encountered at time 0, the time = 0 frequency vector would encode the input on position 0. At time = 1, the time = 0 vector would shift to position 1, and the new input would be encoded at slot 0 (and at some point, the system would have to 'wrap', recycling position 0, etc.). It is a small step to imagine a similar memory where frequency vectors would be replaced or augmented by phonemic vectors, or some other phonetic or phonemic recoding. As each phoneme is processed, the matrix corresponding to the previous phonological state could be shifted on the memory matrix, replaced with the current one aligned at slot 0. “

This is the primary assumption of the once-popular Atkinson and Shiffrin (1971) shift-register model of working memory, which is not cited. Although this model was a useful contribution 40 years ago, it has long ago been discarded because it is neurally impossible and incapable of learning.

D. The explanatory power of ART vs. TRACE

Magnuson (2015) claims in several places that the explanatory range of cARTWORD is far inferior to that of TRACE; e.g., his claims that cARTWORD “has been applied to only one phenomenon (phoneme restoration)” (p. 1481) and that “no other model comes close to the depth and breadth of TRACE’s coverage” (p. 1483). These claims are based upon the fact that the Grossberg and Kazerounian (2011) article focused on providing the first real-time neural simulation of phonemic restoration, in which future context can disambiguate noise-occluded previously occurring phonemes and generate a temporally evolving representation of the restored sequence that is consciously heard, including the order, timing, and amplitude with which the restored phoneme is heard. Magnuson’s criticism is, however, misleading in two major ways:
(1) As we have noted already, cARTWORD is just one contribution to 40 years of incremental development of the ART theory of speech and language learning, perception, and recognition. The models PHONET, ARTPHONE, ARTWORD, cARTWORD, ARTSPEECH, and NormNet, as well as SPINET and ARTSTREAM, all embody the same core design principles. Each incremental refinement of these models includes the same core mechanisms of working memory storage and unitization of sequences in working memory to selectively activate list chunks that can represent phonemes, syllables, and words. This ART theory has explained and predicted many psychological data that are outside the explanatory range of TRACE. Many of these data describe how consciously heard speech sounds depend on past or future linguistic contexts that can span 100-150 milliseconds, and how the brain attempts to form a rate- and speaker-independent representation of variable-rate and variable-speaker speech from multiple auditory streams (e.g., Ames and Grossberg, 2008; Boardman et al., 1999; Cohen and Grossberg, 1986; Cohen, Grossberg, and Wyse, 1995; Grossberg, Boardman, and Cohen, 1997; Grossberg, 1978a, 1978b, 1984, 1986, 2003; Grossberg, Govindarajan, Wyse, and Cohen, 2004; Grossberg and Myers, 2000; Grossberg and Stone, 1986a, 1986b). Because TRACE has no representation of the passage of real time, it can explain none of these data.

(2) cARTWORD is a neural model of speech that is defined by a hierarchy of cortical processing regions whose networks embody cells in laminar cortical circuits (Figure 1b). Variations of these same circuits have also been used in the 3D LAMINART model to explain data about 3D vision and figure-ground separation (Cao and Grossberg, 2005, 2012; Fang and Grossberg, 2009; Grossberg and Swaminathan, 2004; Grossberg and Versace, 2008; Grossberg and Yazdanbaksh, 2005; Grossberg et al., 2008; Raizada and Grossberg, 2003) and in the LIST PARSE model to simulate data about immediate serial recall; immediate, delayed, and continuous distractor free recall; and sequential planned arm movement control (Grossberg and Pearson, 2008). Thus, cARTWORD is part of a larger theory of how the cerebral cortex works, which has already explained how variations on the same canonical laminar cortical circuits can support several kinds of biological intelligence.

cARTWORD hereby contributes to the rapidly emerging paradigm of Laminar Computing. Laminar Computing describes how the cerebral cortex is organized into layered circuits whose specializations can support all forms of higher-order biological intelligence. Indeed, the laminar circuits of cerebral cortex seem to realize a revolutionary computational
synthesis of the best properties of feedforward and feedback processing, digital and analog processing, and data-driven bottom-up processing and hypothesis-driven top-down processing (Grossberg, 2007, 2013). The fact that variations of the same canonical laminar cortical circuits, supported by data about identified neurons, have been used to simulate challenging data about vision, speech, and cognition provides converging evidence that the models that embody these circuits are tapping real brain designs. TRACE cannot make any such claim.

The ART neural models of speech and language are also part of the more comprehensive ART cognitive and neural theory of how the brain autonomously learns to attend, recognize, and predict objects and events, and sequences of them, in a changing world. ART currently has the broadest explanatory and predictive range, and all of its main predictions have been supported by psychological and neurobiological data; see Grossberg (2013) for a review. This explanatory and predictive range emerges from ART analyses of mechanistic links between processes of consciousness, learning, expectation, attention, resonance, and synchrony (the CLEARS processes) during both unsupervised and supervised learning. These general design principles and mechanisms are specialized in ART architectures for speech and language.

TRACE, in contrast, cannot explain any neural data because the main representations of TRACE have been directly contradicted by neurophysiological data; e.g., Figures 4 and 7. The mechanisms of TRACE capture neither the design heuristics nor the mechanistic properties of brain representations. Thus, whereas ART is a principled cognitive and neural theory that has rapidly expanded its explanatory and predictive range over the years, TRACE offers a computational metaphor.

V. LEARNING CHUNKS OF VARIABLE LENGTHS AND SEQUENCES OF REPEATED WORDS

A. Masking field working memory chunks variable-length lists

This section describes ART properties that contradict another of Magnuson’s claims, one that must be faced by all models of language; namely, that “repeated words will be as problematic as repeated phonemes” (p. 1493). The section also reviews ART explanations that contradict another strong claim of Magnuson, namely that “only one model provides truly deep and broad coverage of phenomena in human speech perception and spoken word recognition while providing a basis for representing temporal order including repeated elements: the TRACE model.” This claim is countered, first, by summarizing some of the other data for which ART
has proposed principled explanations but which TRACE cannot explain; second, by
demonstrating that claims about data that TRACE can explain are inaccurate; and third by the
fact that, in subsequent modeling efforts, Magnuson himself has attempted to move beyond the
representations of temporal order as used in TRACE and IA models. For example, the ART
explanation of how sequences of repeated words, not just repeated phonemes, are represented in
the brain also helps to explain the Magical Numbers Four and Seven that directly contradict the
TRACE representation of temporal order.

A neural explanation of the Magical Numbers Four and Seven was first given using an
Item-and-Order working memory that is called a Masking Field (Figure 2; Grossberg, 1978a,
1984, 1986). A Masking Field is a specialized type of Item-and-Order working memory. As with
all Item-and-Order working memories, it is defined by a recurrent on-center off-surround
network whose cells obey the membrane equations of neurophysiology. In a Masking Field,
however, the “items” are list chunks that are selectively activated, via a bottom-up adaptive filter,
by prescribed sequences of items that are stored in an Item-and-Order working memory at an
earlier processing level (Figures 1 and 2). In other words, Masking Field cells represent list
chunks because each of them is activated by a particular temporal sequence, or list, of items that
is stored within the Item-and-Order working memory at the previous processing level. Thus, both
levels of the item and list processing hierarchy are composed of working memories that obey
similar laws.

In order for Masking Field list chunks to represent lists (e.g., syllables or words) of multiple
lengths, its cells interact within and between multiple spatial sizes, or scales, with the cells of
larger sizes capable of selectively representing item sequences of greater length, and of inhibiting
smaller Masking Field cells that represent item sequences of lesser length. As items are stored in
working memory, an adaptive filter activates the learned Masking Field list chunks that represent
the most predictive item groupings at any time, while its recurrent inhibitory interactions
suppress less predictive list chunks. Kazerounian and Grossberg (2014) have simulated how
variable-length list chunks of a Masking Field can be learned as a list of items is stored in
working memory in real time.

An item is more properly called an item chunk, which, just like any chunk, is a compressed
representation of a spatial pattern of activity within a prescribed time interval. In the case of an
item chunk, the spatial pattern of activity exists across acoustical feature detectors that process
sounds through time, and which are compressed by an adaptive filter to activate item chunks. The prescribed time interval is short, and is commensurate with the duration of the shortest perceivable acoustic inputs, of the order of 10 – 100 msec. Some phonemes may be coded as individual items, but others, in which two or more spatial patterns are needed to identify them, may be coded in working memory as a short sequence of item chunks, and are fully unitized as a list chunk. Thus the model in Figure 2 first compresses spatial patterns of feature detectors into item chunks, and then sequences of the item chunks that are stored in working memory are compressed into list chunks.

B. Temporal chunking problem: Learning words of variable length

Masking Fields were introduced to solve the temporal chunking problem (Cohen and Grossberg 1986, 1987; Grossberg 1978a, 1986), which asks how an internal representation of an unfamiliar list of familiar speech units—for example, a novel word composed of familiar phonemes or syllables—can be learned under the type of unsupervised learning conditions that are the norm during daily experiences with language. Before a novel word, or list, can fully activate the adaptive filter, all of its individual items must first be presented. By the time the entire list is fully presented, all of its familiar sublists will have also been presented. What mechanisms prevent the familiarity of smaller sublists (e.g., MY, ELF, and SELF), which have already learned to activate their own list chunks, from forcing the novel longer list (e.g., MYSELF) to always be processed as a sequence of these smaller familiar chunks, rather than eventually as a newly learned unitized whole? How does a not-yet-established word representation overcome the salience of already well-established phoneme, syllable, or word representations to enable learning of the novel word to occur?

C. Self-similar competition solves the temporal chunking problem

A Masking Field accomplishes this using cells with multiple cell and receptive field sizes, or scales (Figure 2), that are related to each other by a property of self-similarity; that is, each scale's properties, including its cell body sizes and their excitatory and inhibitory connection lengths and interaction strengths, are a multiple of the corresponding properties in another scale. Such a self-similarity property can develop as a result of simple activity-dependent growth laws (Cohen and Grossberg 1986, 1987) in the following way:

It is assumed that item chunk cells are endogenously active during a critical period of development. As a result, Masking Field cells that receive inputs from a larger number of item
chunk cells receive a larger total input activity through time. Activity-dependent cell growth causes the Masking Field cell bodies and connections to grow approximately proportionally. This property is called self-similar growth. Cell growth terminates when the cell bodies become large enough to dilute their activities sufficiently in response to their inputs no longer exceed a growth-triggering threshold. Cells that receive more inputs grow larger as a result, so that the effects of individual inputs are smaller on larger cells. In effect, self-similar growth normalizes the total effect of all the inputs that converge on a Masking Field cell. Consequently, such a cell only fires vigorously if it receives active inputs from all of its item chunk cells.

Due to self-similar growth, larger list chunks selectively represent longer lists because they need more inputs, and thus more evidence, to fire. Once they fire, their stronger inhibitory interaction strengths than those of smaller list chunks can inhibit the smaller list chunks more than conversely (“asymmetric competition”). The intuitive idea is that, other things being equal, the longest lists are better predictors of subsequent events than are shorter sublists, because a longer list embodies a more unique temporal context. The stronger inhibition from list chunks of longer, but unfamiliar, lists (e.g., MYSELF) enables them to inhibit the chunks that represent shorter, but familiar, sublists (e.g., MY), more than conversely, thereby providing a solution of the temporal chunking problem.

**D. Magical Number Seven and word superiority**

The word length effect in word superiority studies and the Magical Number Seven both follow from the self-similarity property. This word length effect was discovered by Samuel, van Santen, and Johnston (1982, 1983) who showed that a letter is progressively better recognized when it is embedded in longer words of lengths from 1 to 4. The word length effect is relevant to self-similarity because larger list chunks are more potent and predictive than smaller list chunks in a Masking Field. However, self-similarity implies that the list chunk of a familiar multi-letter word can inhibit the list chunk of a familiar letter, which seems to contradict the property that the word can facilitate perception of its constituent letters, which is the main result of word superiority studies.

This problem is resolved in ART systems with item chunk and list chunk processing levels. The cARTWORD model is such an ART system, as was its predecessor, the ARTWORD model (Grossberg and Myers, 2000). Their list chunk levels are represented by a Masking Field (Figures 1 and 2). In particular, although chunks that represent lists of multiple lengths compete...
within the Masking Field that categorizes list chunks, the top-down expectations from the list chunk level to the item chunk level are excitatory. By self-similarity, list chunks that represent longer words generate larger recurrent inhibitory signals and larger top-down excitatory priming signals to the item chunk level.

The Magical Number Seven also arises due to self-similarity and asymmetric inhibition among list chunks that represent multiple list lengths. This is easy to see by considering how all the cells of a given scale in a Masking Field interact among themselves via a recurrent on-center off-surround network. Call such a network a single-scale network. Each single-scale network is self-similar to all the other single-scale networks that comprise the Masking Field, but each single-scale network chunks item lists of a different length. The largest active chunks tend to win the asymmetric competition in a Masking Field in response to an input sequence. How big these chunks are will depend on prior experience. By the self-similarity of Masking Field scales, the same number of winning chunks will tend to be active, no matter how big their chunks may be. This explains the Magical Number Seven because, as noted by Miller (1956), “the memory span is a fixed number of chunks”. This number turns out to be seven plus or minus two because of parameter choices that define these working memory networks (see Bradski, Carpenter, and Grossberg, 1994).

These explanations of the Magical Number Seven and the word length effect provide further support for the ART prediction that item chunk and list chunk levels process speech and language (Grossberg, 1978a, 1984), rather than the phoneme, letter, and word levels that were used in the Interactive Activation Model (McClelland and Rumelhart, 1981).

E. Conscious speech is a resonant wave: The units of speech and language

Both ARTWORD and cARTWORD simulated parametric psychophysical data about speech that illustrate the revolutionary ART predictions that “conscious speech is a resonant wave” and that “perceived silence is a temporal break in the rate that the resonance evolves”. In particular, the “resonant wave” that embodies the properties of phonemic restoration illustrates how the conscious percepts of speech sounds can proceed from past to future, even while sounds that are heard in the past are determined by future contextual information.

ARTWORD further illustrated how listeners integrate temporally distributed phonemic information into coherent representations of syllables and words. During fluent speech perception, variations in the durations of speech sounds and silent pauses can produce different
perceived word groupings. For example, increasing the silence interval between the words “gray” and “chip” in theutterance “gray chip” may result in the percept “great chip”, whereas increasing the duration of fricative noise in “chip” may alter the percept to “great ship” (Repp et al., 1978). In the “gray chip” to “great chip” example, why should increasing the silence interval between two words, which one might think should make them more distinct, increase the probability that the fricative noise from the second word would leap backwards-in-time over 100 milliseconds of silence to join the percept of the first word? In the “gray chip” to “great ship” example, why should increasing the duration of fricative noise in the second word cause that noise to jump backwards-in-time over an interval of 30 - 80 milliseconds of silence to join the first word? The ARTWORD neural model explains how these percepts arise naturally from Adaptive Resonance Theory concepts, and quantitatively simulates these context-sensitive speech data, data of a kind that TRACE cannot explain because of how it represents time and silence.

**Figure 8.** Hierarchy of speech processing levels. Interactions among three speech processing levels are capable of learning, working memory storage, and performance of word sequences. Each level consists of an Item-Order-Rank working memory. The second and third levels are, in addition, multiple-scale Masking Fields that enable selection and storage of sequences of variable length. All the levels are connected by Adaptive Resonance Theory bottom-up adaptive filters and top-down learned expectations and their attentional focusing capabilities. The first level stores sequences of item chunks. Its inputs to the second level enable that level to store list chunks of item chunks. The inputs of the second level to the third level enable it to store list chunks of list chunks, in particular sequences of words.
F. Item-Order-Rank Masking Field hierarchy chunks lists of repeated words

With this background in mind, it is easy to explain how ART represents lists of repeated words. This can be accomplished by a three-level network (Figure 8): Each processing level in this network is an Item-Order-Rank (IOR) working memory that can store sequences with repeated items in short-term memory. The second and third IOR working memories are, in addition, multiple-scale Masking Fields (MF) that can chunk input sequences of variable length, and choose the sequence, or sequences, for storage that receive the most evidence from its inputs. Each level receives its bottom-up inputs from an adaptive filter and reads-out top-down expectations that focus attention on the feature patterns in their learned prototypes at the previous level. The first level stores sequences of item chunks. The second level stores sequences of list chunks. The individual list chunks of the third level thus represent sequences of list chunks at the second level, including sequences with repeated words, as in the “DOG EATS DOG” example in Magnuson (2015, p. 1481).

VI. HOW DOES TOP-DOWN ATTENTIVE FEEDBACK WORK?

A. TRACE top-down feedback is incompatible with the data

Magnuson (2015) comments that: “Grossberg and Kazerounian (2011) also take issue with TRACE’s lack of absolute constraints on top-down feedback (specifically, they argue that top-down feedback must not be allowed in the absence of any bottom-up support). They cite a passage from McClelland and Elman (1986) (p.75) where those authors speculated about how feedback might be used in a learning variant of TRACE. Grossberg and Kazerounian (2011) argue that the mechanism outlined there would lead to unstable learning…” (pp. 1482-1483).

Grossberg and Kazerounian (2011) did not write that “‘’top-down feedback must not be allowed in the absence of any bottom-up support”. Top-down feedback is ubiquitous in the brain. It is, however, modulatory feedback, not driving feedback, under most conditions, unlike the feedback in TRACE. Driving top-down feedback is feedback that can activate cells without bottom-up support. Driving feedback is incompatible with a large number of psychological and neurobiological experiments about how top-down attention works. These experiments confirm that attention is controlled by a top-down, modulatory on-center, off-surround network. A modulatory on-center cannot drive its target cells without bottom-up support. It can only sensitize or prime its target cells to respond more vigorously to matching bottom-up inputs that may or may not later occur. Because the competitive off-surround can suppress unattended feature combinations, this kind of attentional network is sometimes called “biased competition” (Desimone, 1998).
These mechanisms and data about how attentional matching works were predicted by ART, and are said to obey the ART Matching Rule. Recent models agree about how to mathematically instantiate the ART Matching Rule. Moreover, as Grossberg and Kazerounian (2011) have reviewed, the ART Matching Rule has been mathematically proved to enable fast category learning to occur without catastrophic forgetting, and its violation can cause catastrophic forgetting, as proved by Carpenter and Grossberg (1987). Many subsequent psychophysical and neurophysiological experiments have supported the ART prediction about how this form of top-down feedback can modulate plasticity, for example, during visual perceptual learning (Ahissar and Hochstein, 1993). auditory learning (Gao and Suga, 1998) and somatosensory learning (Krupa, Ghazanfar, and Nicolelis, 1999; Parker and Dostrovsky, 1999). See reviews by Grossberg (2013) and Kaas (1999).

Problems with using driving top-down feedback were noted in Magnuson et al. (2012):

“An intriguing aspect of ART’s processing assumptions is that its version of top-down feedback cannot cause hallucinatory representations. A ‘2/3 rule’ means that weak inputs (e.g., phonetic features corrupted by noise) can be strengthened once recognized by higher levels of processing, but completely absent inputs cannot be created from nothing. As we discuss below, a common criticism of feedback in TRACE is that it could make the system hallucinate (Norris et al., 2000). Although, in practice, misperception in TRACE seems generally similar to misperception in humans (Mirman, McClelland, & Holt, 2005) and the default TRACE parameters also give it strong, bottom-up priority, future modeling efforts might benefit from nonsymmetrical feedback rules such as those implemented in ART.”

The ART Matching Rule also enables top-down feedback to generate suprathreshold responses when it is supplemented by basal ganglia volitional inputs; e.g., Figure 1b. This allows top-down feedback to activate visual imagery and internal thought by being converted from a modulatory to a driving mode. However, if this basal ganglia input becomes tonically hyperactive, it can create visual or auditory hallucinations, as can occur during schizophrenia (Grossberg, 2000). Such considerations are, however, entirely outside the explanatory range of TRACE.

**B. The ART Matching Rule implies phonemic restoration**

The previous discussion illustrates that a major prediction of ART concerns how fast learning can occur without causing catastrophic forgetting; that is, ART proposes how to solve the *stability-plasticity dilemma*. ART predicted in the 1970s that this is accomplished by learned top-down expectations that are matched against bottom-up information. The match focuses attention upon expected combinations of critical features. As noted above, ART predicted that this ART Matching Rule is realized by a top-down, modulatory on-center, off-surround network. In brief,
the ART Matching Rule shows how attentional matching enables fast learning with self-stabilizing memory.

In the case of speech perception, the ART Matching Rule implies the main properties of phonemic restoration: A top-down expectation can select a bottom-up signal that is consistent with it, such as those noise components that match the learned features in the expectation. However, because it is modulatory, such a top-down expectation cannot create something out of nothing, so silence does not lead to restoration. Because it takes awhile for a resonance to form, future context can influence the expectation that controls restoration. Thus, phonemic restoration in response to future context is a consequence of the brain’s mechanism for learning language quickly without experiencing catastrophic forgetting. Said in another way, phonemic restoration supplies additional experimental evidence for the ART Matching Rule operating in real time.

These observations about phonemic restoration were made long before cARTWORD was developed; e.g., in Grossberg (1986). It took almost 30 years of theory development to finally be able to simulate how restoration could be generated in real time in the correct temporal order within a laminar cortical model of speech and word recognition, such as cARTWORD.

The TRACE model uses driving top-down feedback that embodies none of the properties of the ART Matching Rule. Hence it cannot provide the kind of elegant explanation of phonemic restoration that ART and cARTWORD have provided as a manifestation of the brain’s ability to learn language quickly and stably.

This leaves one remaining question: Can TRACE simulate phonemic restoration at all?

VII. TRACE CANNOT EXPLAIN PHONEMIC RESTORATION DUE TO HOW IT PROCESSES SPEECH

This section shows that, contrary to claims in Magnuson (2015), TRACE cannot simulate phonemic restoration. Our demonstrations of this failure also highlight fundamental problems with the representations in TRACE of time, temporal order, silence, and top-down processing.

A. Background leading to Magnuson (2015)

A reviewer of the Grossberg and Kazerounian (2011) manuscript presented simulations that purported to emulate phonemic restoration using the TRACE model, and used them to claim that TRACE can explain this class of phenomena. Grossberg and Kazerounian (2011) responded to those simulations by showing that, when they were carefully analyzed, they did not simulate phonemic restoration, and indeed illustrated fundamental problems of the TRACE model.
Magnuson later submitted a manuscript to JASA purporting to show, once again, how TRACE (by way of jTRACE) could account for phonemic restoration. After our referee report of this Magnuson manuscript once again pointed out serious problems with his simulations, Magnuson withdrew that manuscript. Magnuson soon after that resubmitted a variation of this failed manuscript to JASA. This manuscript was published as Magnuson (2015) without our knowledge, even though it purported yet again to simulate phonemic restoration and focused yet again on purported shortcomings of the cARTWORD model in Grossberg and Kazerounian (2011).

The criteria laid out in Grossberg and Kazerounian (2011) argued that any model of phonemic restoration should show three things, in accordance with psychological data. First, when phonemes are replaced with silence, the model should not give rise to a percept of the removed phoneme. Second, when a phoneme is replaced with broadband noise, the model should give rise to a percept of the phoneme that was removed. And third, a model of phonemic restoration should be able to explain “backwards effects in time”. For example, Warren and Sherman (1974) found that when replacing the phonemes /v/ and /b/ in “delivery” and “deliberation”, which are contextually neutral up to the removed portion, restoration of the removed phoneme could only occur when the disambiguating portions, “ery” and “eration”, were finally presented. As such, restoration of the removed phonemes relies on future information in order to determine what is to be restored.

The initial referee report claiming to show restoration in TRACE, as well as Magnuson’s subsequent paper, focused on showing phonemic restoration using the word “luxury” (represented in TRACE as ‘l^kS^ri’), and how restoration can occur when word initial (/l/), medial (/S/), and final (/i/) portions, are replaced. In order to show restoration, Magnuson first modified noise representations, such that input noise no longer ramped on and off (as is the case with normal phonemes in TRACE), and modified representations of silence, such that they do not contain the overlapping (co-articulated) portions of adjacent phonemes. However, as pointed out in our referee report of the manuscript that Magnuson withdrew from JASA, simulations resulting from these modifications alone (as shown in Figure 3 of Magnuson (2015)) do not show phonemic restoration for a number of reasons.

Most critically, as can be seen in Figure 4 of Magnuson (2015), or in our recreation of it in Figure 9, when the initial /l/ and medial /S/ phonemes are replaced by noise, the most active
phoneme representations are not /l/ and /S/, but rather /k/ and /g/. Furthermore, in the case of replacement of the initial /l/, Magnuson claims that only the activity of the corresponding phoneme activity prior to cycle 25 should be considered. In this time window, it is clear that the /l/ phoneme becomes active between time cycles 5-15. What Magnuson does not show, however, is that in this simulation, the lexical node for “luxury” first becomes active at time cycle 20, which is after the phoneme activation levels for /l/ have already dropped below zero. This means that the peak in the /l/ activation that Magnuson claimed as evidence for restoration could not have been due to top-down feedback from the lexical representation of “luxury”. Indeed, the noise has only just turned off by time cycle 12, so there is no plausible way for restoration resulting from lexical feedback to have occurred in the 5-15 cycle time window. This problem, and more generally the issue of how it arises from the TRACE representation of time, is discussed in detail in Section VII.C.

In order to try to overcome the problem that the initial modifications to noise/silence representations in TRACE resulted in the wrong phonemes becoming most active, Magnuson then collapsed the input feature specifications for the phonemes, so that every phoneme receives an equivalent amount of bottom-up input. This was done to prevent noise from preferentially activating certain phonemes over others. As before, however, this change was insufficient for showing phonemic restoration when either the initial or medial phonemes have been replaced by noise. This can be seen in Figure 8 of Magnuson (2015). When /l/ and /S/ are replaced with noise, their phoneme representations are again not the most active ones over the course of the simulation, suggesting that alternative phonemes would in fact be perceived.

Magnuson argued that this is due to biasing of phonotactic probabilities for certain phonemes, by way of early top-down feedback from the full lexicon. Because top-down feedback from the lexical layer begins as soon as any lexical node becomes active, any lexical entry that even partially codes for some of the incoming acoustic input can begin to feed back and excite phoneme representations. As such, with any non-trivial lexicon, the distribution of phonemic activations at various locations in a word will be skewed, resulting in top-down feedback from active lexical nodes that preferentially excites phonemes occurring more often at a particular location, than the phonemes that occur less frequently in that position. Magnuson argues that the skewed excitatory top-down feedback, to certain phonemes over others, means that “…a pure test of lexical restoration in TRACE is virtually impossible in a lexicon with even
In order to eliminate such bias from the lexicon, Magnuson next removed all lexical entries—that is, all word representations—except for the single word “luxury”.

Although Magnuson claims that the model, when incorporating all these changes, with only the single word lexicon, is then able to show phonemic restoration, his simulations do not reflect the data. In the simulation results in Figure 12 of Magnuson (2015), not only does noise replacement of the initial phoneme fail to show restoration (i.e., positive activation of the /l/ phoneme), but also, in cases where the initial /l/ and final /i/ have been replaced by silence, the activities of the corresponding phoneme representations become activated, and thus would be perceived despite having been replaced by silence. Moreover, in the cases where a percept is formed when the word medial and final /S/ and /i/ are replaced by noise, the lexical representation for “luxury” has already become active, meaning that none of the instances of restoration show any backwards effects in time.

In order to demonstrate these results in enough detail for these problems to be fully understood, we did our own simulations that incorporate the changes made by Magnuson to the TRACE model and inputs. In particular, in order to verify that our additional simulations accurately reflect the modifications made by Magnuson to the TRACE model and inputs, two of his figures were replicated after incorporating these changes. With this verification in hand, two new simulations were done to illustrate the seriousness of TRACE’s failures. After reviewing these results, a discussion is provided below of how these problems follow from the representations of time, temporal order, silence, and top-down processing in TRACE.

![Diagram](image)

**Figure 9.** Recreation of Figure 4, Magnuson (2015). This figure summarizes the phoneme activation levels for correctly time-aligned phonemes of the word “luxury”, represented as the
sequence 'l^kS^ri' in TRACE, when replaced by true silence (0.0 noise, top row), or by full noise (1.0 noise, bottom row). Each curve in the simulation is labeled with its phonemic descriptor. In this simulation, noise and silence have been modified according to Magnuson (2015). The plots in the first column show replacement of the word initial /l/, second column show replacement of word medial /S/, and the final column show replacements of word final /i/. As can be seen in the top row, despite receiving no input, the /l/, /S/, and /i/ grow steadily through time, with /l/ and /i/ becoming activated during the course of the simulation, thereby causing a kind of auditory hallucination. Bottom row: the most active phoneme representations when /l/ and /S/ have been replaced by noise, are /k/ and /g/ rather than the expected /l/ and /S/.

B. Recreating the Magnuson (2015) simulations

The first simulation, in Figure 9, replicates Figure 4 of Magnuson (2015). It shows the time-aligned phoneme node activations, in response to the input word ‘luxury’ (‘l^kS^ri’), with the word initial, medial or final phonemes, /l/, /S/, or /i/, replaced with either silence (0.0 noise), or with true noise (1.0 noise). These simulations include the noise and silence input representations introduced by Magnuson (2015), which, for the case of silence, remove overlapping portions of co-articulated phonemes, and for noise, use values that do not ramp on and off. Although the changes made for these simulations do not on their own suffice to show restoration, they are recreated here to validate our additional simulations. Indeed, in the case of silence in row 1, they show activations of phonemes that receive no bottom-up inputs, and in the case of noise in row 2, they show maximal activations of phonemes other than those that were replaced by noise.

Figure 10 replicates Figure 12 of Magnuson (2015). This simulation is the culmination of all changes made to the TRACE model and inputs, and is claimed by Magnuson to show the model correctly restoring phonemes when a removed phoneme is replaced by noise, but not by silence. The changes include the initial modifications to noise/silence input representations (as in Figure 9), modification to the phoneme/feature specifications such that all phonemes receive equivalent bottom-up support, and removal of all words from the lexicon except 'luxury' (‘l^kS^ri’).
Figure 10. Recreation of Figure 12, Magnuson (2015). This figure shows the phoneme activation levels when incorporating all the changes made by Magnuson in order to show restoration. This includes the modified noise/silence representations, the collapsed phoneme feature specification, and the use of a one-word lexicon. The properties of this simulation do not match properties of phonemic restoration in multiple cases. See the text for details.

In the case that silence replaces the word initial /l/ and word final /i/ (first and third columns of the top row), the corresponding phonemes become positively activated, and all of them grow during the same time windows as when the phoneme is either intact, or replaced by noise. Moreover, as can be seen in the first column, /l/ grows less when its activation is supported by noise (second row) than when it receives no input (first row). Thus, in order to advance a claim of phonemic restoration, if /l/ is assumed to be heard when it is supported by noise, then it must also be heard when it is replaced by silence, thereby contradicting the facts of phonemic restoration. In addition, /l/ grows less when it receives a noise input (second row, first column) than /i/ grows when it receives no inputs (first row, third column). Thus, if /l/ is heard when it is supported by noise, then both /l/ and /i/ are heard when they are replaced by silence, thereby contradicting phonemic restoration properties even more seriously. Another problem is that, if /l/ is assumed not to be heard when it is replaced by silence, then neither /l/ nor /S/ can be heard when they is supported by noise, since they both attain no more than the maximal value of .1, again contradicting basic data about phonemic restoration.

C. Silence, top-down processing, and hallucinations in TRACE

The above simulation raises the question of how, during simulations of phonemic restoration with the TRACE model, does a phoneme, when replaced by silence, gives rise to a percept of the removed phoneme, in contradiction of the data. In TRACE, this hinges on whether or not
excitatory top-down feedback from the lexical layer is sufficient on its own to activate, or drive, phoneme nodes that have not received any bottom-up input. If a phoneme representation becomes active due solely to such top-down signals, despite the removal of acoustic input to that phoneme, then TRACE would incorrectly predict that the phoneme was nonetheless perceived.

Given that this does happen in Figure 10, how does Magnuson argue that these simulations accurately reflect restoration phenomena? Magnuson makes two claims to defend this position:

First, Magnuson claims that we should ignore late-cycle activations of the phoneme representations: “The early time window is the critical region for restoration; what matters is whether there is a basis for differential behavior as the word is being experienced, rather than many time steps after the noise (e.g., approximately 30 slices after replacement onset, where the 0.0 noise case catches up to the noise replacements).”

Second, in footnote 5, Magnuson claims that late advantages for silence (0.0 noise) over true noise (1.0 noise)-replaced phonemes, can be “wiped out” by reducing the amount of phonemic lateral inhibition, which was presumed to decrease activations for noise-replaced phonemes.

The issues of time and late cycle activations are discussed in more detail in Section VII.C. It is useful, however, to immediately note that, in the final simulations purporting to show restoration, Magnuson’s two claims do not alleviate the problems that are exhibited in the TRACE simulations. First, note that in the case of replacement of the word final /i/ by silence, the phoneme representation for /i/ becomes active between time cycles 30-60, roughly the same time window that the phoneme is active when replaced by noise. Because the phoneme is active during the same window for noise and silence replacement, Magnuson’s claim provides no relief in this case, and thus the supra-threshold activation of /i/ in response to silence is contradicted by phonemic restoration data. Moreover, if one attempts to claim that the maximum activity of /i/ is not sufficient to generate a percept, then other cases where a percept should occur will also not generate a percept, as noted in the previous section. Thus, both the times of response and their amplitudes force the conclusion that phonemic restoration data are not properly simulated.

In the case of replacement of the word initial /l/ by silence, the phoneme representation becomes active at approximately time cycle 50, and only continues to become more active as the simulation proceeds. While Magnuson's choice of time interval seems to apply in this case, the
time cycle threshold he uses appears to be selected arbitrarily, and, more importantly, does not comport with the how time, and time-aligned phonemes, are represented in TRACE, as will be discussed further in the next section.

Figure 11. Simulation of TRACE showing phoneme activations when word initial /l/, medial /S/, and final /i/ have been replaced by extended temporal silence. In previous simulations, when /l/ is replaced by silence, the first 12 time cycles of acoustic input were removed. In this case, the first 13 time cycles are replaced by silence, thereby removing an additional time cycle during which the acoustic inputs for /l/ had previously been active. Similarly, the silence window for /S/ and /i/ have been extended, such that 14 cycles of acoustic input are replaced by silence in the case of /S/, while 13 cycles are removed in the case of /i/.

As for the second claim, when the TRACE model uses true silence representations, it is instructive to consider how the model actually prevents phoneme nodes from becoming active solely in response to top-down feedback. Magnuson is correct in noting that it has something to do with lateral inhibition within the phoneme layer. It is problematic, however, to suggest that TRACE can prevent activation of silence-replaced phonemes simply by decreasing the amount of phonemic lateral inhibition. In fact, because lateral inhibition is the only thing preventing top-down feedback from activating phoneme representations that did not receive any bottom-up input, a number of problems arise which are independent of any parameter choice for the level of inhibition in that layer.

Consider, for example, what happens when the word initial /l/ is replaced with silence. During normal presentation of the word “luxury” (‘lʌksəri’), the inputs corresponding to /l/ would be presented for 12 time cycles, ramping up to a peak at cycle 6, and ramping off by cycle 12. The inputs for the subsequent phoneme, /ʌ/, begin ramping on at time cycle 6, to a peak at time cycle 12, and ramping off by cycle 18. Subsequent phonemes would be presented in similar fashion, with co-articulated portions overlapping with one another. When the initial /l/ phoneme has been spliced out, the full 12 cycles of the corresponding /l/ input are removed, as are the first 6 cycles of the ramping on of /ʌ/. The primary source of inhibition on the /l/ phoneme centered
at the peak of its input (time cycle 6), is the strong activation of the following phoneme, /\^/, centered on the peak of its input at time cycle 12.

An obvious consequence of this fact is that, if the temporal extent of the silence is increased, the phoneme representation for /\^/ has less time to receive bottom-up input, thereby resulting in a lower activation level of the /\^/ phoneme node. This in turn will result in decreased inhibition of the time-aligned /l/ phoneme, allowing it to become more active than if the silence was shorter. At a psychophysical level, this would imply that replacing a larger portion of the acoustic signal for the word “luxury” by silence would increase, rather than decrease, the percept of the removed portion, as one might expect from the data. Importantly, because this is a structural problem with the way TRACE deals with driving top-down feedback, there is no clear parametric fix for this. For example, decreasing top-down feedback will not help, since it will remain the case that a larger silence window that replaces a portion of the word, will lead to stronger activations of the replaced phoneme(s), than smaller silence windows. Indeed, the simulations in Figure 11 show that the maximum activities in response to the slightly extended silence are, in the cases of /S/ and /i/, at least three times larger than in the corresponding cases in Figure 10. These activities must, moreover, be considered supra-threshold, and thus generate percepts, if the /S/ trace in response to a noise input in Figure 10 (second row, second column) is considered to be supra-threshold.

In particular, the simulations of TRACE in Figure 11 incorporate all the changes made by Magnuson in the simulations that he claimed to show restoration. Although not shown, additional simulations have shown that these results also hold when the model has a full lexicon, as well as an uncollapsed feature set. In the first simulation, instead of replacing only the first 12 cycles of input with silence, the amount of silence was extended by an additional cycle, so that the first 13 cycles of acoustic input are now replaced with true silence. When replacing the medial /S/ with silence, the silence was extended in both directions, so that there is an additional time cycle of silence before and after the normal window (i.e., 14 cycles of total silence, rather than 12). For word final /i/, the silence was again extended by 1 time step. In all cases, not only do all the phoneme nodes become strongly active despite the fact that they have been replaced by silence, but they become more active in the time window prescribed by Magnuson.

Even more problematic, perhaps, is what happens when larger chunks of the word are replaced with nothing but silence. In the case where the lexicon represents only one word, for
example, suppose that only inputs corresponding to the first two phonemes, /l/ and /^/, are presented. Because there are no lexical competitors, top-down feedback will activate all the remaining phonemes that comprise the word “luxury”, aside from the phoneme representation for /k/. Interestingly, the positive activation of the remaining phonemes occurs simultaneously. In Figure 12, this can be seen as the temporally overlapping traces for /S/, /^/, /r/ and /i/), suggesting further problems with how TRACE deals with time. In any case, this result shows that TRACE hallucinates four phonemes that were never present in the acoustic signal.

In the full lexicon case, presenting only /l/ and /^/ causes the model to hallucinate /k/, since the most active word is “luck”. If /S/ is also presented, so that the total presentation is 'l^kS', then top-down excitation will activate the remaining /r/, and /i/, causing the TRACE model to hallucinate the final portions of “luxury”. In all, in addition to not simulating critical properties of phonemic restoration, the TRACE model also hallucinates sequences of phonemes for which no inputs are presented, and can do so more vigorously if more silence is present in a given word.

D. Time and temporal order in TRACE

Another foundational problem of the TRACE model concerns how it represents time, which is a central concern for any model of speech and language. One of the defining characteristics of speech perception is the fact that speech is inherently temporal. Speech perception is also context-dependent, with both past and future contexts determining conscious speech percepts, as phonemic restoration illustrates. As such, a biologically plausible model of speech perception must be able to describe, not only how the
Figure 12. Simulation of TRACE showing presentation of the acoustic input containing only the first two phonemes of “luxury”, /l/ and /^/. The phoneme activations shown, are for the phoneme /l/ aligned at position 2, /^/ aligned at position 4, /k/ at position 6, /S/ at position 8, /^/ at position 10, /r/ at position 12, and /i/ at position 14. Despite most of the acoustic input corresponding to the word being absent from the input signal, top-down feedback activates the phonemes /S/, /^/, /r/ and /i/, such that they simultaneously activate at approximately time cycle 30, reaching a final activity level of .1 by time cycle 100. This shows the TRACE model hallucinating four phonemes in the absence of any acoustic input.

脑代表长期记忆，或LTM，学习的模式的迹，但也是，如何这些模式是暂时存储在工作记忆中响应底部-向上 acoustic inputs arriving in real time, even before sequence learning occurs. After learning occurs, such a theory needs to show how learned sequence chunks interact with bottom-up acoustic inputs as they arrive in real time and activate working memory. A theory of speech and language perception must thus explain both what we hear, and when we hear it, let alone how we learn this information through past experiences.

TRACE sidesteps the issue of how temporal order, and in particular temporal sequences, are represented in the brain by using a “two-fold” representation of time. Acoustic inputs are presented sequentially in what is referred to as “real-time”, whereas phoneme and lexical representations use a method of temporal alignment that requires multiple reduplications of each
representation, over many points in time. Each phoneme representation is copied, so that there is
one centered at every 3 “real-time” slices, with each phoneme spanning 6 “real-time” slices. The
“first” phoneme representation for a given phoneme would, for example, process inputs spanning
from “real-time” slices 0 to 6, while the “second” representation for that phoneme would process
temporal inputs from slices 3 to 9, the “third” from 6 to 12, and so on. These reduplicated
phoneme representations are said to occur at various alignment positions. When the acoustic
input for the word “luxury” is presented, the /l/ features ramp on from time slice 0, to a
maximum level at time slice 6, ramping off by time slice 12. As such, they most activate the /l/
representation at alignment position 2. The features corresponding to /^/ begin ramping on at
time slice 6, to a maximum at slice 12, and ramp off by slice 18, these features maximally
activate the /^/ representation at alignment position 4. Note here that the /^/ representations at
time alignment positions 3 and 5 will also receive some bottom-up input (since the representation
for /^/ at alignment position 3 processes acoustic inputs occurring between slices 6-12, and /^/ at
alignment position 5 processes those occurring between time slices 12-18), neither receives as
much bottom-up input as the representation at alignment position 2, which is fully centered over
the portion of its corresponding acoustic inputs when maximally activated. Lexical
representations are similarly reduplicated, so that all positions in “real-time” can be covered by
one of the copies of the lexical entry.

As discussed earlier, this method of representing time is at odds with both biological and
psychophysical data. Some of these shortcomings were mentioned in the original TRACE paper.
As McClelland and Elman (1986) note:

“One fundamental deficiency of TRACE is that fact that it requires massive duplication of
units and connections, copying over and over again the connection patterns that determine
which features activate which phonemes and which phonemes activate which words. As we
already noted, learning in activation models (e.g., Ackley, Hinton, & Sejnowski, 1985;
Grossberg, 1976; Rumelhart & Zipser, 1985) usually involves the retuning of connections
between units depending on their simultaneous activation. Given TRACE’s architecture, such
learning would not generalize from one part of the Trace to another and so would not be
accessible for inputs arising at different locations in the Trace. A second problem is that the
model, as is, is insensitive to variation in global parameters, such as speaking rate, speaker
characteristics and accent, and ambient acoustic characteristics.”

In addition to their inherent implausibility, such duplication is contradicted by psychophysical
data concerning how phonemes and words are represented (e.g., Bowers, Kazanina, and
Andermane, in press; Toscano, Anderson, and McMurray, 2013). More relevant to the issue of
phonemic restoration is the failure of such a representation to explain what is perceived, and when it is perceived. Each of the phoneme and word representations, at every alignment position, has an activation value for all of the real-time cycles of any given simulation. This leaves open various possible methods for determining which phonemes are perceived at a given point in time.

In the first possible method, all of the phonemes centered at a particular alignment position (or potentially adjacent positions) are interpreted as competing hypotheses about what is perceived at that position in time. The activity of those representations, as simulation real-time cycles unfold, then represent changes in those hypotheses in response to bottom-up and top-down interactions.

The second possible method assumes that, outside of the need to process acoustic inputs at various points in time, the alignment positions do not matter as such. To determine what is perceived at any given moment in real-time, simply look at all the phoneme representations (across all alignment positions) at a given point in real-time cycles, and read out the most active representation.

A third possibility is to use some hybrid procedure, whereby phonemes at a given alignment position determine what is perceived at the “real-time” cycles that position is centered on, but with an additional coordination mechanism that prescribes certain time-windows within which that representation is deemed to be relevant. No such method appears to have ever been elicited in any detail, and even if it had, there would be no clear mechanism for explaining how this coordination would occur, either in the model or in the brain. The very fact that the linking hypothesis between model representation and perception is so indeterminate is itself a problem, especially when it is contrasted with the unambiguous brain-to-behavioral linking hypotheses that naturally arise in neural models such as cARTWORD (Grossberg & Kazerounian, 2011).

Indeed, Dahan, Magnuson, and Tanenhaus (2001) note why such a coordination procedure is problematic (albeit in the context of word-activations):

“Determining activations from TRACE is not a trivial process. Word units in TRACE function as templates. For a word unit to become highly active, it must be well aligned with phonemic (and featural) inputs. TRACE avoids the alignment problem by aligning a copy of each word unit every three input slices. Given input, TRACE reports the activity of copies of each word unit aligned at different slices. The experimenter must decide how to decode the word-unit patterns of activation [italics ours]. The method we used was to determine which copy of a word unit reached the highest activation and then use the activation of that unit over all input cycles as the activation of that word...This procedure is problematic because it cannot be implemented in an incremental fashion; it requires an omniscient observer [italics
ours] to compare peak activations after processing is finished. Incremental methods are possible. Each lexical item could have an associated decision node that would either summate the responses of all copies of the word template at all slices or report the activation of the most active word template at each slice. For the purposes of the current article, we use the simple method we have described and leave this issue open for future research.”

It should be noted that an implementation of such a decision node would itself involve an omniscient observer that would need to be able to selectively determine where and when in the corpus of slices only certain activities should be counted. This would have to be repeated for all lexical units. Then all the decision nodes, or omniscient observers, would need to be globally coordinated, perhaps by an even more omniscient observer.

In either of the first two cases that were summarized above, the problem with time remains. Imagine, for example, a phoneme at alignment position 10 with an activity value at real-time cycle 6 that is higher than that of a phoneme at alignment position 2 at the same time cycle. If the only relevant factor is the alignment position, then it does not matter that the phoneme aligned at position 10 has early activity that is higher than that of the phoneme aligned at position 2. In order to know what is perceived at time-cycle 6, one has to consider only the activity of the phonemes aligned at position 2, not just at the given time cycle, but over the whole course of the simulation (preventing the possibility of ignoring late cycle activations).

On the other hand, if the phoneme that is perceived at time-cycle 6 only depends on the activity of phonemes, independent of alignment position, then a large number of problems arise. For example, in the restoration case, if word-medial /S/ is replaced by noise, the phoneme representation for /S/ aligned at position 8 will show positive activation, but phoneme representations at earlier alignment positions (e.g., /k/ aligned at position 6) will still have higher activity than the activity level of /S/ at the correct position. This would mean that, in all the simulations of restoration, a phoneme from a different alignment position that received full bottom-up acoustic input would maintain higher activation levels than the restored phoneme at the correct alignment position.

VIII. CONCLUSIONS

As shown in the above sections, whereas ART is a principled cognitive and neural theory, TRACE is a computational metaphor whose properties exhibit major foundational problems in its representations of time, temporal order, silence, and top-down processing. The phenomenon
of phonemic restoration highlights ART’s strengths and TRACE’s weaknesses in representing these fundamental processes.

In particular, in order to argue that the TRACE model can simulate phonemic restoration phenomena, Magnuson made a series of changes to the model inputs and representations. Despite these changes, his simulation results fail to exhibit the most basic properties of phonemic restoration. Even when all changes have been made, phonemes that are replaced by silence can become more active than phonemes that are replaced by noise. Indeed, replacing more phonemes by silence only makes their activations bigger, and hallucinations of future phonemes for which there is no bottom-up evidence can occur. These problems can be traced to fundamental representational deficiencies in how both the original TRACE model, and Magnuson’s variations thereof, represent time, temporal order, silence, and top-down processing.

In particular, Magnuson’s simulations fail to show how restoration occurs when the word-initial phoneme has been replaced by noise. Furthermore, by removing all lexical entries aside from “luxury”, the single remaining word node that represents “luxury” becomes active in response to minimal bottom-up inputs. Masking Fields were introduced and developed (Cohen & Grossberg, 1986; Grossberg, 1978a) to solve this temporal chunking problem (Sections V.B and V.C), thereby ensuring that list chunks in a Masking Field network can only be activated by bottom-up inputs from working memory when there is sufficient evidence for them. Because TRACE has not solved the temporal chunking problem, in the case of word medial and final replacements, top-down feedback is available prior to the arrival of the replaced phoneme. This means that Magnuson's simulations not only fail to show restoration on their own, but they fail the additional constraint that a model be able to show how future context can influence conscious percepts of earlier noise-occluded phonemes; that is, they fail to show backward effects in time.

The driving top-down feedback used by TRACE defeats any attempt to use TRACE to explain phonemic restoration. This strong conclusion follows from the fact that increasing the duration of a silence replacement robustly results in greater activation of the deleted phoneme representation, than does a shorter silence replacement. The idea that replacing a portion of acoustic input with a larger amount of silence, rather than a smaller amount, results in an increased likelihood or strength of a percept for the removed phonemes, is prima facie implausible. This is not a parametric failure, or a failure of linking hypotheses, but rather a fundamental weakness of the TRACE model.
In addition to the problems with silence and top-down signals, the temporal order representations in TRACE are not only theoretically untenable, but force contradictory interpretations of model simulations. The use of temporally aligned phoneme and word representations, each of which is responsible for processing acoustic inputs in a predetermined time window, makes it impossible to say what is perceived, and when it is perceived. If, as Magnuson argues, only phoneme activations up to approximately 30 time cycles should be considered after the onset of its corresponding input, what allows us to ignore the fact that the previously activated phoneme is much more strongly activated than a noise-replaced phoneme in that 30 cycle window?

The phenomenon of phonemic restoration thus deeply probes issues about the representation of time, temporal order, silence, and top-down processing that expose fundamental weaknesses of TRACE, while also illustrating how design principles and mechanisms of cARTWORD, and temporal ART models more generally, can naturally explain percepts that are highly context-dependent, including percepts that depend for their conscious representations on future disambiguating contextual information. In particular, unlike TRACE, cARTWORD simulations include a clear representation of what is perceived, and when it is perceived, as well as of why silence-replaced portions of an acoustic signal do not lead to a percept of the removed phoneme, while replacement of that phoneme by broadband noise leads a listener to perceive the phoneme as though it was intact, even when the disambiguating context occurs after these silence and noise intervals. These key issues are all ones to which the TRACE model provides either no answer, or an answer that is contradicted by the data. In contrast, in ART, these properties emerge from its psychophysically and neurobiologically supported explanations of how top-down expectations and attention dynamically stabilize the rapid and stable learning of auditory representations in real time, and upon which future theoretical developments can build with confidence.

References
Proceedings of the National Academy of Sciences USA, 90, 5718–5722.


Gao, E., and Suga, N. (1998). Experience-dependent corticofugal adjustment of midbrain...


**Figure Captions 1 and 7 (including mathematical symbols)**

**Figure 1.** (a) ARTWORD model macrocircuit. (b) cARTWORD model macrocircuit. cARTWORD includes a hierarchy of two cortical processing levels that model different cortical regions. Each level is organized into laminar cortical circuits that share a similar laminar organization, with cells organized into layers 5/6, 4, and 2/3, and with a similar distribution of inter-laminar connections. In both levels, deep layers (6 and 4) are responsible for processing and storing inputs via feedback signals between them. Superficial layers (2/3) respond to signals from layer 4 to categorize, or chunk, distributed patterns across these deeper layers into unitized representations. The first level is responsible for processing acoustic features (cell activities $F_i$ and $E_i$) and item chunks (cell activities $C_i^{(I)}$), whereas the second level is responsible for storing of sequences of acoustic items in an Item-and-Order working memory (activities $Y_i$ and $X_j$), and representing these stored sequences of these items as unitized, context-sensitive list chunks (activities $C_j^{(L)}$). List chunks are selected in a Masking Field, which is a multiple-scale recurrent on-center off-surround network whose self-similar and shunting properties enable its list chunks to selectively represent sequences of multiple lengths. Top-down connections exist both within and between levels. Intra-level connections enable item chunks in layer 2/3 of the first level to send top-down attentional matching signals to their distributed features in layer 5/6, and list chunks in layer 2/3 of the second level to send top-down signals to their working memory item chunks in layer 5/6. Both types of signals can modulate, but not fire, their target cells when acting alone. Inter-level top-down signals are the ones that can trigger resonance.
They occur from list chunks in layer 2/3 of the second level to a basal ganglia gate (triangle), and from stored item chunks in layer 5/6 of the second level to the features in layer 5/6 of the first level. The basal ganglia gate opens when a list chunk in layer 2/3 of the second level is chosen in response to a sequence of item chunks in level 4 of the cognitive working memory. Once the gate opens, top-down feedback from the cognitive working memory in layer 5/6 of the second level can resonate with active item features in level 5/6 of the first level, thereby triggering a coordinated resonant wave that can propagate through bottom-up and top-down signal exchanges throughout both levels of the cortical hierarchy and give rise to conscious percepts. [Reprinted with permission from Grossberg and Kazerounian (2011).]

**Figure 7.** Left two columns: SEF rank-direction cell that responds phasically before rightward saccades that occur at the third ordinal position in a sequence of saccades. Each row, from left to right, shows the cell’s responses when the first, second, and third saccades occur. Symbols U, L, and R denote upward, left, and right saccades. The cell responds vigorously only when the third saccade in the sequence moves to the right, as summarized in the left column and last two rows, and the right column and top row. [Data adapted with permission from Isoda and Tanji (2002).]

Right two columns: SEF rank-direction model cell activity $S^X_\nu$ of a cell that codes the same properties. [Simulations reprinted with permission from Silver et al. (2011).]