An Attractor Network Model of Serial Recall

Matt Jones (mattj@umich.edu) and Thad A. Polk (tpolk@umich.edu)
Department of Psychology, 525 E. University
Ann Arbor, MI 48109 USA

Abstract

We present a neural network model of verbal working memory which attempts to illustrate how a few simple assumptions about neural computation can shed light on cognitive phenomena associated with the serial recall of verbal material. We assume that neural representations are distributed, that neural connectivity is massively recurrent, and that synaptic efficiency is modified based on the correlation between pre- and post-synaptic activity (Hebbian learning). Together these assumptions give rise to emergent computational properties that are relevant to working memory, including short-term maintenance of information, time-based decay, and similarity-based interference. We instantiate these principles in a specific model of serial recall and show how it can both simulate and explain a number of standard cognitive phenomena associated with the task, including the effects of serial position, word length, articulatory suppression (and its interaction with word length), and phonological similarity.

Introduction

Working memory is among the most intensively studied cognitive processes in both cognitive psychology and neuroscience, and yet results from the two fields have not made as much contact with each other as one might hope. For example, cognitive psychology has discovered a host of robust empirical phenomena associated with verbal working memory and has developed elegant theoretical models, such as Baddeley's phonological loop, that can explain the empirical results (Baddeley, 1986). Nevertheless, the details of how these psychological hypotheses are instantiated in the brain is an open question (but see Burgess & Hitch, 1999, for one recent proposal). Similarly, there is a substantial body of neuroscientific research investigating the neural substrates of working memory in both animals (Fuster, 1973; Funahashi, Bruce, & Goldman-Rakic, 1989) and humans (Smith & Jonides, 1999), but this work has typically only addressed a small subset of the rich behavioral data and theories available in cognitive psychology.

In this paper, we attempt to illustrate that a simple and independently motivated model of neural computation can make contact with, and even shed light on, the cognitive psychology of verbal working memory. We begin by describing a few widely accepted assumptions about neural computation. Next, we discuss some of the emergent computational properties of these assumptions that are relevant to verbal working memory (e.g., maintenance, decay, interference). We then illustrate how these assumptions can be instantiated in a specific computational model that simulates and explains many of the major psychological phenomena associated with the serial recall task.

A Simple Model of Neural Computation

We begin with three simple and widely accepted assumptions about neural computation. The first is that representations in the cortex are generally distributed across a population of neurons, rather than being localized to individual cells. The second is that there is massive connectivity among neurons within local areas of cortex and that this connectivity is recurrent rather than unidirectional. The third assumption is that synaptic efficiency is modified based on the correlation between pre- and post-synaptic activity (Hebbian learning; 'cells that fire together wire together').

Taken together, these assumptions give rise to networks with interesting emergent properties, many of which are relevant to working memory. For example, such networks are known to be capable of maintaining an activation pattern via internal reverberatory activity even after the input to the network has been removed (Hopfield, 1982). Those patterns which the network can maintain in this way are termed attractors, and under the Hebbian learning rule they tend to become those patterns to which the network is repeatedly exposed. Furthermore, when presented with a noisy or incomplete version of a previously trained pattern, the activity of the network will tend to converge upon that attractor state which is most similar to the input, thereby retrieving the original pattern.

Another property of attractor networks that is relevant to working memory is that they naturally exhibit similarity-based interference. Attractor networks are capable of storing multiple patterns as attractor states, but if those patterns are similar to each other (overlap substantially) then there is a greater likelihood of error. In particular, we have found that these networks often retrieve a pattern that in some sense represents a group of similar patterns, but from which it is not possible to recover a single specific pattern unambiguously.

Finally, we have also found that attractor networks can be easily extended to exhibit time-based decay. In the original formulation of attractor networks, each unit was binary (either ON or OFF) and activation patterns could be maintained for indefinite periods of time (Hopfield, 1982). Hopfield (1984) subsequently showed that networks using more realistic continuous-valued units could also exhibit similar computational properties. We have found that such continuous-valued attractor networks are also capable of exhibiting time-based decay once external input is removed.

The Serial Recall Task

In the standard serial recall task, a subject is presented, either visually or auditorially, with a sequence of items, most often words, letters, or digits. Once presentation of the list has been completed, the task of the subject is to repeat back the list in its original order, either by speaking or by writing.

This task has been intensively studied and a large number of robust behavioral phenomena have been identified. Below are some of the major phenomena which we will address in this paper. For a more thorough review of the literature see Gathercole (1997).

Serial Position The effects of an item's position within the presented list are generally described as two separate phenomena (see, e.g., Crowder, 1972). *Primacy:* Items from the start of the list tend to have a higher probability of recall than those from the middle of the list. *Recency:* Items from the end of the list tend to be recalled better than those from the middle.

Word Length Lists composed of items which take a longer time to articulate tend to be associated with poorer recall (Baddeley, Thompson, & Buchanan, 1975).

Articulatory Suppression Requiring subjects to overtly articulate irrelevant verbal material during presentation of a list tends to impair their performance (Murray, 1968).

Word Length x Articulatory Suppression The effect of word length is significantly reduced under conditions of articulatory suppression, provided that suppression continues throughout recall (Baddeley et. al., 1984).

Phonological Similarity Recall of a list tends to be decreased when the items of the list are phonologically similar to or confusable with each other (Conrad & Hull, 1964). Furthermore, when phonological similarity is limited to a subset of the items, e.g. those in the even positions, then performance on that set is selectively impaired as compared to the non-confusable items (Baddeley, 1968).

An Attractor-Based Model of Serial Recall

The goal of the present model is to demonstrate that the basic assumptions about neural computation outlined previously are relevant to our understanding of some of the behavioral phenomena associated with serial recall. To do so, we show how these computational principles can be instantiated in a specific model of serial recall that exhibits many of these phenomena.

The model is composed of a number of separate yet interconnected attractor networks of the type described previously (Figure 1).

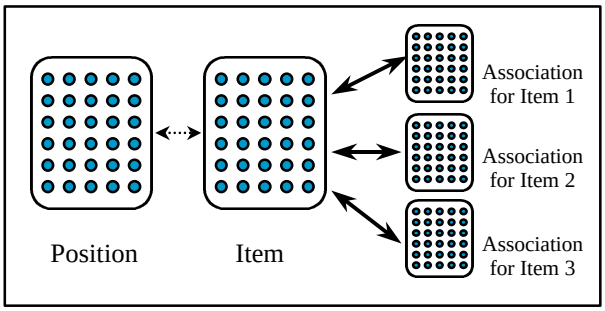

Figure 1. The architecture of the model. Circles represent individual units, rectangles represent individual attractor networks, arrows represent connections between units in different networks. Units within each network are all interconnected (not shown). Both these connections and connections between the Item and Association networks (thick arrows) are assumed to reflect long-term memory and do not change in the simulation. Connections between Position and Item (dashed arrow) reflect short-term position-item associations and are modified according to a Hebbian learning rule as each item is presented or rehearsed.

Position Network

This network encodes position within an arbitrary list of items. As currently modeled, each position corresponds to an activity pattern in which 10 out of 100 units are active. Patterns for different positions are pairwise disjoint, although this assumption could be changed to model more detailed data on positional confusions. Although the network itself does not draw a distinction, we interpret these patterns as encoding relative, rather than absolute, position in the list. Note that the Position network itself encodes no information about item identity; this knowledge will be stored in connection weights, learned during list presentation and rehearsal, between the Position and Item networks.

Item Network

The Item network is where the actual elements of the list are represented, again as distributed patterns comprising 10 active units each. Individual units are meant to correspond to various phonological or otherwise acoustic properties of the stimulus (a word or letter), and the network is presumed to have learned

these patterns via Hebbian learning over repeated exposure to each of them. Thus units which are active in the same pattern(s) have mutual excitatory connections between them while units which are active in different patterns tend to inhibit each other. The result is that the network, when given external input, e.g. from the Position network, and then allowed to evolve its activity over time, will settle into the learned pattern that most closely matches the pattern of the input.

Association Networks

One crucial aspect of the Item network is that it is competitive. This means that whenever one item is represented by the network, representations of all other items are wiped out, so that as far as the Item network is concerned, all information about which items have been recently encountered is lost. This property allows the network to select a response. However the fact that competitive dynamics wipe out past information also implies that there must be some other source of item information in the system. This other source of item information is provided by the Association networks associated with each item.

Each Association network has a single attractor whose constituent units share permanent excitatory connections with those comprising the corresponding attractor in the Item network. Crucially though, the Association networks don't interact with each other, allowing multiple Association networks to be active at the same time. Consequently the item information in these networks is not erased by the representation of later items, but rather it remains and slowly decays. This residual activity provides another source of (non-position-specific) information to the system to be used at the time of recall.

The assumption then is that when a new item is represented, it partially overwrites the activation associated with other items, but that it does not do so completely. For example, if presented with the list "K B", the assumption is that presentation of "B" partially overwrites the representation of "K", but that some aspects of the representation of "K" are preserved. In the model, this distinction is captured by the distinction between the Item network (in which previous activity is overwritten) and the Association networks (in which it is not).

Model Operation

Simulation of the serial recall task in the model consists of three phases: list presentation, rehearsal (which is interleaved with presentation), and recall. During presentation of each item, the Item, Association, and Position networks are put into the attractor patterns corresponding to the present item and list position. The source of the input that generates these patterns is not modeled but is presumed to be early sensory processing, as well as perhaps some executive input in the case of the Position network. Co-activation of units in the Position and Item networks now leads to formation of excitatory connections via a Hebbian learning rule, so that later activation of the same pattern in the Position network will under suitably favorable conditions lead to the corresponding pattern appearing in the Item network.

Between presentations of each successive list item, the model rehearses already presented items in order to further strengthen the Position to Item connections that have been learned. This is accomplished by putting the Position network into the attractor pattern corresponding to a given position, and allowing the connections from there to the Item network, along with inputs coming from the Association networks, to generate a pattern in Item. After allowing activity to evolve for a short period of time (reflecting the time constraints during this portion of the task), the system uses the resultant pattern of activity to rehearse. Rehearsal is presumably accomplished via covert articulation generating a sensory-level input to the Item network of the same type as it receives at presentation, after which the same Hebbian learning rule as was used during presentation is applied to update the Position to Item connections.

Because rehearsal is restricted to items that have already been presented, we have by the termination of presentation a gradient in number of rehearsals across serial positions which favors the earlier items. This gradient translates into an advantage for the earlier positions in two ways. First, the extra learning of associations between early position patterns and their corresponding item patterns leads to stronger connections and thus a stronger memory trace. Second, the additional learning has a significant effect on proactive interference: leftover connections from position patterns to item patterns from previous lists get attenuated with each application of the learning rule (because those old item patterns are not active when the rule is applied), thus leaving less potential for interference during recall.

Also worth noting at this point is another positional gradient in the state of the system at the conclusion of presentation, this time in the level of activity in the Association networks. Because each Association network is activated at the time of presentation of its corresponding item and then decays after that, the networls for items most recently presented, i.e. those at the end of the list, will be most active at the start of recall.

The process of recall is quite similar to the retrieval processes that operate in rehearsal. For each list position starting with the first, the Position network is

placed into the attractor pattern corresponding to that position (presumably by some executive process). Activity in the Item network is then allowed to evolve until it stabilizes, with inputs from both the Position and Association networks tending (in ideal conditions) to drive that activity towards the pattern for the correct response. Once the network has stabilized the system probabilistically chooses an item for response based on the similarity of all known patterns to the actual pattern.

Experiment 1: Simulation of Standard Phenomena

The following set of simulations provides a demonstration of the model's ability to predict many of the standard phenomena associated with the serial recall task. The data we attempted to simulate were taken from Baddeley et. al. (1984; Experiment 5), which explores the effects of serial position, word length, and articulatory suppression.

Experimental Design

As in the design of Baddeley et. al. (1984), we ran the model on lists of both short and long words, both with articulatory suppression and without. The short and long word lengths used allowed for 5 and 9 item rehearsals per presentation, respectively (note Baddeley's presentation rate was 1.5 sec/word). Proportion of correct responses (or rather mean probability of responding correctly) were recorded for each serial position in each condition.

Results

The results of 150 runs on each condition are presented in Figure 2, along with the empirical data. Both empirical and simulated data exhibit the initial increase in error percentage over the first few serial positions (primacy effect), as well as a decrease on the final position (recency effect). In both cases performance is impaired for longer words and in conditions of articulatory suppression, with an interaction between these two effects indicated by a smaller effect of word length under the suppression conditions.

Discussion

Closer inspection of the model's performance and inner workings during the task reveal the following explanations for the phenomena:

Primacy Effect As described previously, increased rehearsals for earlier position-item pairs, and thus more applications of the Hebbian learning rule, lead to better quality of information encoded in the connections from the earlier position patterns to the Item network. This in turn lead to higher rates of correct recall for earlier items in the list.

Recency Effect In keeping with the other positional gradient described previously, the Association networks for the final items on the list were more active at the time of recall. As a result the additional information encoded by their inputs to the Item network acted to increase rates of correct recall at the end of the list.

Word Length Rehearsal was assumed to take place via covert articulation (which provides the source of the simulated sensory input to the Item network), and thus the time to rehearse should be dependent on the articulation time of the items in question. Lists of longer words were therefore allowed fewer rehearsals, and so were given less opportunity for learning associations between positions and items, thus leading to lower overall performance.

Articulatory Suppression Articulatory suppression was modeled as reducing the probability that each attempt at rehearsal was successful, rather than being interrupted by the process of overt articulation. As with the word length effect, this reduction in rehearsals led to less learning and in turn lower performance.

Word Length x Articulatory Suppression Under suppression rehearsals were less likely to be successful, and thus reducing the number of attempts at rehearsal by increasing word length had less of an effect on learning. Conversely, with longer words there were fewer rehearsals than with shorter ones, and thus interfering with them by imposing articulatory suppression made less of a difference.

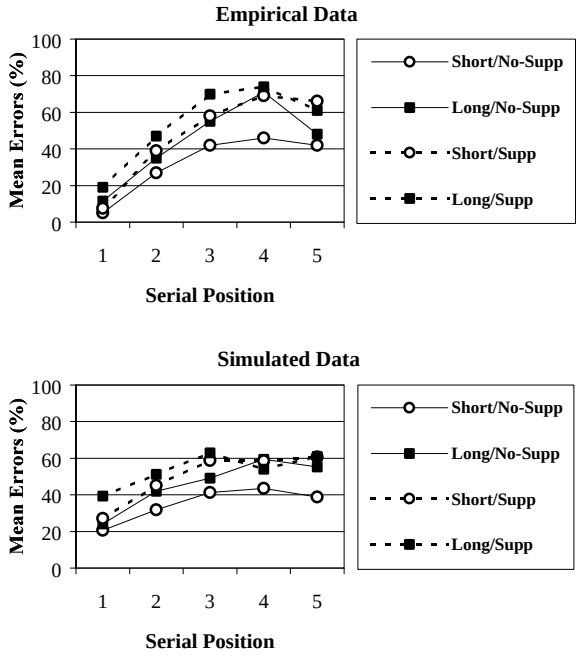

Figure 2: Mean percent error on the serial recall task in both empirical (Baddeley, et. al., 1968) and simulated studies. Data in each graph are divided into condition according to word length and articulatory suppression.

Experiment 2: Phonological Similarity

The last effect we attempted to model was that of phonological similarity between list items.

Experimental Design

The experiment followed the design of Experiment V of Baddeley (1968). In that experiment, lists of length 6 were taken from a pool of 12 letters, 6 of which were acoustically similar to each other (B,C,D,P,T,V) while the other 6 were all dissimilar (J,K,L,R,W,Y). In one condition only the even positions had confusable letters, and in another only the odd positions did. In both cases the resultant serial position curves had a characteristic sawtooth shape, with greater percentages of errors on confusion positions than on non-confusion positions.

Our approach in modeling phonological similarity was to assume that it is reflected by increased similarity between representations in the Item network. Our hypothesis was that similarity-based interference would lead to conditions in which the network failed to retrieve a single item but rather retrieved a pattern that was a combination or superposition of multiple items. The main change made to the model in order to capture this idea was to include a set of units in the Item network that were shared by the representations of all 6 acoustically similar items. Other changes included reducing the level of inhibition in the network in order to facilitate superpositional patterns.

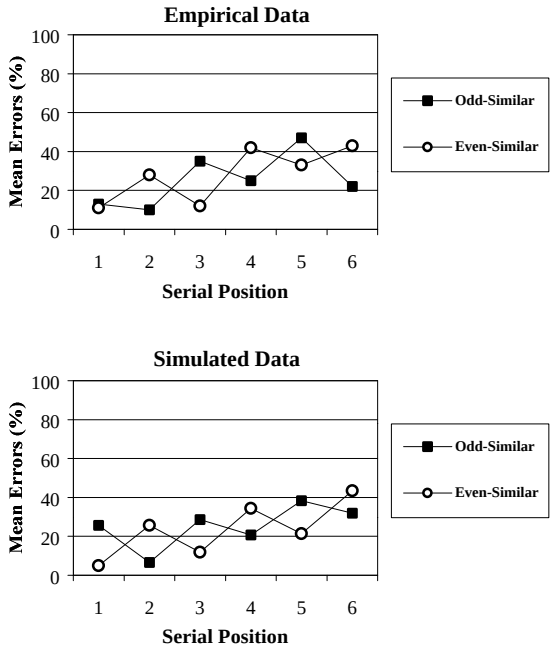

Figure 3: Mean error rates in the phonlogical similarity experiment in both empirical (Baddeley, et. al., 1968) and simulated studies. Data in each graph are divided into condition according to which list positions (evens or odds) contain the phonologically similar items.

Results

The model was run for 100 lists in both the Odd-Similar and Even-Similar conditions. Mean probability of a correct response was calculated for each serial position in each condition and is shown in Figure 3 along with the empirical data from Baddeley (1968). Both graphs clearly show the effects of phonological similarity, with greater error rates on the acoustically confusable items.

Discussion

As hypothesized, the model produced phonological similarity effects by often falling into spurious attractor states representing the combination of two or more similar items. When this happened, the system was left with only partial information about the identity of the correct item, and had to guess based on similarity between the actual and idealized patterns. This explanation differs from classical theories about acoustic confusion among items, but may be better seen as a theory of redintegration (Schweickert, 1993).

General Discussion

Psychological theories of verbal working memory, such as Baddeley's (1986) phonological loop model, have had great success in explaining serial recall at a cognitive level. These models have identified a core set of cognitive constructs (e.g., similarity-based interference, information maintenance with time-based decay, reactivation by articulatory rehearsal, etc.) that have proven extremely useful in explaining human behavior in this task. Nevertheless, these models do not typically address how those cognitive constructs are realized computationally in the brain. Conversely, research on neural computation has shown how many of these same cognitive constructs can arise as emergent properties in neural networks inspired by properties of the brain. However, these findings have not previously been exploited to explain detailed behavioral data regarding verbal working memory. In this paper, we have tried to show that ideas from cognitive psychology and neural computation can be fruitfully combined to produce an integrated model of verbal working memory that begins to bridge the gap between the cognitive and neural levels.

Most of the assumptions incorporated in the model are already well supported and widely accepted. For example, in keeping with many other models of verbal working memory, we assume that participants rehearse the items in an effort to keep their representations active (and that early items are rehearsed more), that rehearsal is related to covert articulation, that articulation suppresses the ability to rehearse, that similar-sounding items interfere with each other, etc. Similarly, the simple assumptions about neural computation that are incorporated in the model are well

established and their emergent computational properties are well known.

Incorporating assumptions from both psychology and neural computation in a single, integrated model has a number of benefits. For example, most psychological theories have little to say about some fundamental issues regarding the mechanisms underlying verbal working memory. For example, how is information actually maintained, why does it decay over time if not rehearsed (and how does rehearsal refresh it), and how do similar items interfere with other? Indeed, even computational models of verbal working memory often build in these assumptions rather than simulating them (e.g., by explicitly weakening memory traces as a function of time or by assuming that similar-sounding items are occasionally confused with each other). By exploiting a few independently motivated assumptions about neural computation, the current model is able to provide computationally explicit answers to these kinds of questions.

Considering constraints from both fields also led to a model with a number of novel theoretical features. For example, assuming that the neural representation of a stimulus/concept corresponds to a specific distributed activity pattern suggests that different instances of the same item involve the same units. This contrasts with model such as the Phonological Loop which allow for multiple independent instances of a repeated item.

Learning also plays a much more important role in the attractor model than it does in the phonological loop model and its variants. With each presentation of the item, the attractor model learns an association between a position representation and an item representation. These associations interfere with the learning of new position-item associations and therefore allow the model to predict intrusion errors from similar positions on previous lists and, more generally, substantial proactive interference (learning previous lists impairs the model's ability to learn subsequent lists). Furthermore, Hebbian learning within the Item network can provide a natural account of long-term learning of new vocabulary.

There are a number of aspects of serial recall that the model has not yet accounted for. Among some of the most important of these are the effects of visual presentation (articulatory suppression reduces the phonological similarity effect with visual presentation, unlike with auditory presentation) lexicality (memory for words is better than nonwords), temporal grouping (presenting items in groups that can be chunked improves performance), and positional similarity (errors often involve transposing items that are nearby in the list). The lack of coverage for these phenomena is among the most important limitations of the current model and work is underway investigating whether it can be extended to address them.

References

Baddeley, A. D. (1968). How does acoustic similarity influence short-term memory. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 20A*, 249-264.

Baddeley, A. D. (1986). *Working memory.* Oxford: Clarendon Press.

Baddeley, A. D., Lewis, V., & Vallar, G. (1984). Exploring the articulatory loop. *Quarterly Journal of Experimental Psychology, 36A*, 233-252.

Baddeley, A. D., Thompson, N., & Buchanan, M. (1975). Word length and the structure of short-term memory. *Journal of Verbal Learning and Verbal Behavior, 14*, 575-589.

Burgess, N. & Hitch, G. J. (1999). Memory for serial order: A network model of the phonological loop and its timing. *Psychological Review, 106,* 551-581.

Conrad, R. & Hull, A. J. (1964) Information, acoustic confusion and memory span. *British Journal of Psychology, 55*, 429-432.

Crowder, R. G. (1972). Visual and auditory memory. In J. F. Kavanaugh & I. G. Mattingly (Eds.), *Language by ear and by eye.* New York: McGraw-Hill.

Funahashi, S., Bruce, C. J., & Goldman-Rakic, P. S. (1989). Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. *Journal of Neurophysiology, 61*, 331-349.

Fuster, J. M. (1973). Unit-activity in prefrontal cortex during delayed-response performance: Neuronal correlates of transient memory. *Journal of Neurophysiology, 36*, 61-78.

Gathercole, S. E. (1997). Models of verbal short-term memory. In M. A. Conway (Ed.), *Cognitive models of memory*. Cambridge, MA: The MIT Press.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences, 79*, 2554-2558.

Hopfield, J. J. (1984). Neurons with graded response have collective computational properties like those of 2-state neurons. *Proceedings of the National Academy of Sciences, 81*, 3088-3092.

Murray, D. J. (1968). Articulation and acoustic confusability in short-term memory. *Journal of Experimental Psychology, 78*, 679-684.

Schweickert, R. (1993). A multinomial processing tree model for degradation and redintegatration in immediate recall. *Memory and Cognition, 21(2)*, 168-175.

Smith, E. E. & Jonides, J. (1999). Storage and executive processes in the frontal lobes. *Science, 283*, 1657-1661.