A connectionist model of phonological short-term memory

Gordon D. A. Brown
Department of Psychology
University College of North Wales
United Kingdom

ABSTRACT

A connectionist model of phonological short-term memory is described. The model makes use of existing connectionist techniques, developed to account for the production and perception of speech and other sequential data, to implement a model of the articulatory rehearsal involved in short-term retention of verbal information. The model is shown to be consistent with a wide range of experimental data, and can be interfaced with existing connectionist models of word recognition. The model illustrates, within a connectionist framework, how the mechanisms of speech perception and production can be recruited for the temporary storage of information. Advantages of this strategy are discussed.

INTRODUCTION

The inclusion of some limited-capacity speech based temporary store is near-universal within cognitive models of language processing, and the properties of this store have been extensively investigated by psychologists over the past three decades. Recent connectionist modelling work has naturally been concerned with the temporary storage of information, but a large body of existing experimental evidence from cognitive psychology cannot readily be interpreted in terms of existing connectionist models. This is partly because of the difficulty of dealing with certain types of temporal phenomena in connectionist models, and also because earlier cognitive psychological models have not always taken account of the temporal dimension in any explicit way (Elman, 1988). There is a need, then, for a psychologically well-motivated model of the temporal characteristics of human short-term memory.

Previous connectionist approaches to short memory have generally been concerned to characterize the types of architecture that can give rise to temporary information storage, either at a neural level or in terms of a cognitive-level working memory system (e.g. Grossberg & Stone, 1986; Schneider & Detweiler, 1987). Schreter and Pfeifer (1989) describe a simple localist architecture which gives rise to primacy and recency serial position curve effects, but their architecture is not intended to account for the detailed experimental results of the type outlined below. Our own approach focuses specifically on the phonological short-term memory store, which is normally viewed as just one subpart of a more complex working memory system (e.g. Baddeley & Hitch, 1974).

PSYCHOLOGICAL APPROACHES TO STM

Many early theorists held the view that short-term memory contained a constant number of "slots" that could be filled by material to be remembered. More than approximately seven items could not

be held in short-term storage, but "item" came to be interpreted broadly, allowing for the possibility that large amounts of information could be chunked together in such a way that each slot could hold much information- a character, a word, even a well-learned sentence.

An alternative class of explanation of limited STM capacity comes from the time-limited **trace decay** model (e.g. Baddeley, 1986; Schweickert & Boruff, 1986). In this type of model, a trace is registered in immediate memory when each stimulus item is encountered, and this trace is subject to decay over time. The trace can be refreshed by using a subvocal rehearsal procedure, but if the traces of all the items are to be maintained then it must be possible to rehearse all the items to be remembered within the time taken for the trace of any item to decay to threshold. Thus, as Schweickert and Boruff make clear, the probability that a list will be correctly recalled will be equal to the probability that the time taken to recite the list is less than the variable duration of the trace. Many researchers have suggested that subjects' immediate memory span for familiar materials such as words and digits will be equal to the amount of that material that can be rehearsed subvocally in a fixed time interval. Estimates of this constant time interval vary, but average out at around two seconds.

There is considerable experimental evidence for the trace decay model. A correlation between articulation rate and span has been observed in a variety of contexts, across and within both languages and individuals. Developmental increases in memory span are paralleled by an increase in speech rate (Hulme & Muir, 1985), and adult span correlates with rate of articulation (e.g. Baddeley, Thomson & Buchanan, 1975). Memory span for long words is smaller than span for shorter words in the same language, where "length" is measured in terms of articulation duration (Baddeley et al., 1975). This word length effect is abolished when subjects are required to suppress articulation and are therefore unable to make use of the subvocal rehearsal procedure (Baddeley et al., 1975; Baddeley, Lewis & Vallar, 1984). A similar pattern of results is observed across languages: subjects using languages in which materials (usually digits) are more slowly articulated show reduced memory spans. These ubiquitous correlations between rate of articulation and memory span have been taken to support some version of the verbal trace decay model. In one specific version, Salame and Baddeley (1982) claim that the "articulatory loop" component of the working memory system consists of a phonological store (which gives rise to phonemic confusability effects in STM tasks) and an articulatory rehearsal process (which gives rise to word length effects). Information in the phonological store will decay unless rehearsed. Access to the store when material is presented visually can only be gained via the articulatory rehearsal procedure, and use of the rehearsal procedure will be prevented by articulatory suppression. The connectionist model we report here may be seen as an implementation of a phonological store and speech-based rehearsal process.

It can be seen that these models, which have received a great deal of support from the psychological literature, rely heavily on the temporal characteristics of both information decay and the speech-based articulatory rehearsal procedure. In order to implement this type of model using connectionist methodology, it is therefore necessary to have a way of representing the temporal flow of information. There have been considerable recent advances in the ability of connectionist models to account for temporal phenomena in plausible ways. Previous attempts involved recoding the temporal dimension as a spatial one (Elman, 1988), and sometimes required a reduplication of the entire network for each time-slice of input. However, a different approach involves making some of the input units to a network sensitive to the recent activation history of the network (McClelland & Rumelhart, 1988). In the following section we show how this type of architecture can be extended to produce a psychologically plausible model of the temporal characteristics of human short-term memory.

THE MODEL ARCHITECTURE

The heart of the model of STM is a model of the articulatory rehearsal process used to refresh traces in the phonological store. We model this by taking two separate connectionist networks, one designed for speech production and one designed for speech perception, and interfacing these two nets. The first net, based on an architecture developed by Jordan (1986), can take a temporally static, unordered plan (e.g. a representation of whole-word phonology) and translate this unchanging input into a temporal *sequence* of outputs (e.g. an ordered list of phonemes or articulatory commands). This type of architecture is illustrated by the left half of Figure One: the "production net plan units" are held constant throughout a given output sequence, and the "production net state units" or "context units" have their activations set on the basis of the previous network output. This architecture has been modified by Norris (1989) to *recognize* temporal sequences as single items (as occurs in speech perception).

The Norris model has the same basic architecture as the Jordan net, but is trained to associate a temporally constant pattern of activation on the output units with a time-varying sequence of inputs to the "plan" units. Thus a sequence of items can be input to the network, which will compute a single appropriate output. The resulting network has a number of attractive characteristics, including the ability to generalize in the time domain (e.g. to recognize words spoken at varying rates) and the ability to recognize items within a constant stream of sequential input without the need for reduplication of the net at every point where an item to be recognized might begin (see Norris, 1989, for a discussion of these issues).

When a speech production net and a speech recognition net of the types discussed above are interfaced, so that the output of the production net provides a source of input to the perception net, the architecture in Figure One results. This may be interpreted as a model of the subvocal articulatory rehearsal process in STM, in that the speech production system may direct its output into the speech perception system without overt spoken output ever resulting. Note that not all connections to other parts of the cognitive system are shown: for example, we assume that input and output lexica are separate but connected, and that production and perception mechanisms are connected at various stages (see Ellis & Young, 1988; and Monsell, 1987, for discussion of relevant architectural issues). Input to the rehearsal procedure is provided from a set of input nodes (those in the bottom left-hand corner of Figure One): these represent knowledge about word pronunciations and could be computed for example from the position-independent orthographic trigram units in the network discussed by Mozer (1987). In a complete model there would be input to the speech production net from both high-level and low-level spelling-to-sound correspondences (Brown, 1987a). Note that our "lexical input units" are labelled as input units simply because they provide input to the network we are modelling; in a complete model of the cognitive system they would be more properly characterized as output units. All that matters for present purposes is that there is a set of units that provides input to the speech production network, and that is all that is implemented at present. These units would themselves receive input from a number of different sources- the semantic system and visual short-term memory as well as the spelling-to-sound translation process.

In the present small-scale version of the model, there are 10 nodes in each oval drawn in Figure One- thus in each of the perception and the production nets there are 10 input/plan units, 10 context/state units, 10 hidden units and 10 output units. Individual nodes (other than hidden units) in the present model represent single phonemes, although in some of our experimental (and psychologically more plausible) versions of the model, nodes stand for individual articulatory features. Finally, and crucially for the present model, there are 10 phonological storage nodes which take as their (sequential) input the (sequential) output from the speech perception network (the right half of Figure One). These phonological storage nodes are partly responsible for

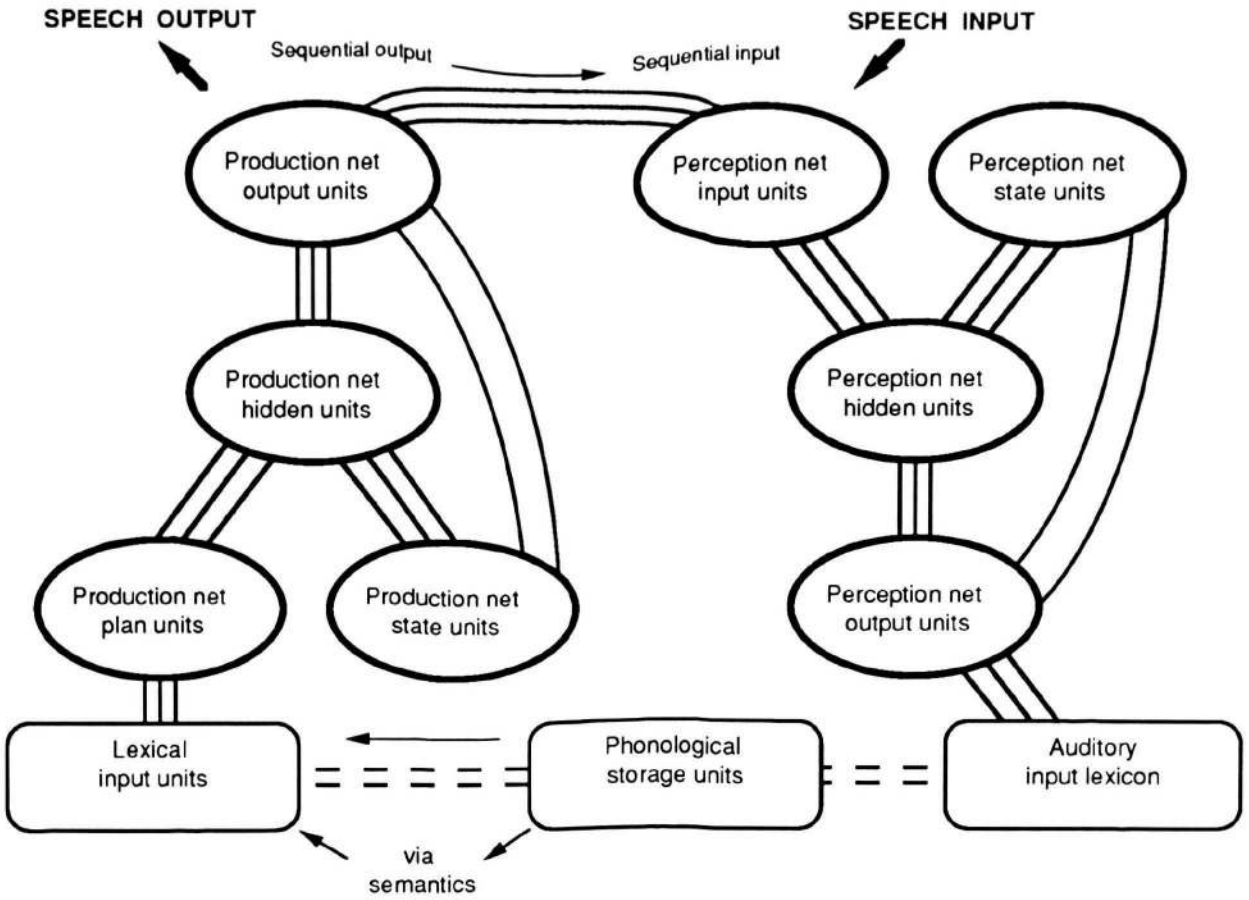

Figure One

temporary storage in the model, but they cannot be accessed directly from visual lexical input. As in the Salame and Baddeley (1982) account, it is assumed that the phonological store gives rise to phonemic confusability effects in short-term memory, and the articulatory rehearsal process gives rise to word length effects (see below).

The simulation of short-term memory processes in the model involves two quite separate phases. The model is first given "long-term memory" about the sequences of phonemes that make up words, and this information is assumed not to change during the later simulation of short-term memory for lists of whole words. Thus, in the first phase, the learning phase, the perception net and the production net separately learn to recognize and produce the same set of phoneme sequences. This learning takes place using the standard back-propagation algorithm described in Rumelhart, Hinton & Williams (1986); the precise training procedure for these nets is described in Jordan (1986) and Norris (1989). At present the nets are trained with a small vocabulary of items which vary in length from 2 to 5 phonemes, with the phonemes being drawn from the very limited (due to computational resource restrictions) pool with which the model currently operates. No psychological reality is claimed for this process.

The main phase of the simulation is the retention in STM of a sequence of items represented at the lexical input level. The input of the sequence of items to be remembered is given by clamping on sets of the lexical input units in sequence: this is analogous to the presentation of a sequence of words. While it is clamped on, each word in the input sequence acts as a (temporally constant) input to the production net side of the articulatory rehearsal process in Figure One. Thus each item

in the list of material to be remembered can be input into the "production" network, emerging as a temporal output sequence, and this sequence can then be directed as input to the sequence "recognition" net that effectively re-recognizes the item in question and hence re-activates, or refreshes, the phonological storage nodes over which the item is represented. This process is repeated for each word in the sequence, and the process for each word takes an amount of time that depends on the spoken duration of the item in question, because a complete pass through the production and perception system is required for each time-slice of the item to be rehearsed.

During the rehearsal of each word as described above, information in the non-ordered input nodes, and in the phonological store, decays. This is the primary cause of forgetting in the model. The output of the rehearsal process may refresh the phonological representation of the rehearsed item in the phonological store, as described above, and this in turn can refresh the nodes at the lexical input level that initiated the rehearsal process and hence make the item available for spoken output or another rehearsal. We make no commitment as to whether the phonological storage units can gain this access to the speech production system directly or only via the semantic system (not illustrated or implemented). In the current version of the model the phonological storage nodes can excite the lexical input nodes in a linear fashion. This is a unidirectional link: lexical input nodes on the left hand side of Figure One can neither excite nor inhibit the phonological storage units directly. This is consistent with the experimental evidence (Baddeley, Lewis & Vallar, 1984).

It is assumed that only those items whose activation in the input nodes is above a certain threshold will be available for recall. In assessing the performance of the model, it is simply assumed that recallable items are those whose entries in the lexical input units are above threshold at the time of recall. Activation of an item represented in the input level may be reinforced either by incoming activation from other cognitive modules, such as the semantic system or visual STM, or by activation from the phonological storage units. (At present we are not concerned to model these cognitive modules, and in our simulations we simply assume a small but constant amount of activation arriving at the input to the speech production system from other sources, such as visual memory, while items are being rehearsed. Only a fixed amount of such activation is assumed to be available for all the items to be remembered.) If the level of activation for an item in the input lexical nodes decays below a certain level before the item can be rehearsed, that item will be forgotten. Thus, as in the trace decay model, short-term memory span is limited in capacity to those items whose activations can be refreshed by the articulatory rehearsal process described above before their activations decay to below threshold. During the continuous sequence of rehearsal, the next item to be rehearsed is always selected on the basis of which item's representation is most decayed while still being above threshold. Note that the phonological store in this model can be viewed as both pre-production and post-production, in that material in the store has been processed by much of the speech production apparatus, but can also, indirectly, be part of further sequences of speech production. Note also that the model incorporates both long-term and short-term storage without using both fast and slow weights as in some other accounts.

THE EXPERIMENTAL DATA

There is a wide range of empirical data relevant to evaluation of the model, not all of which can be covered here (see Baddeley, 1986, for a review). Most of our investigations to date have examined the model's ability to remember various sequence-lengths of items, where the items themselves can vary in length. The performance of the model with simulated visual and auditory input can readily be tested, with and without portions of the articulatory rehearsal procedure being made unavailable. For the sake of simplicity it is assumed that all possible phonemes take the same length of time to produce, and that a word containing six phonemes will take twice as long to articulate as words with only three phonemes. These simplifying assumptions are not critical to the operation of the model.

There are widely-observed **word length effects** in STM tasks (Baddeley, Thomson & Buchanan 1975): subjects can remember more items when the items to be remembered have a short spoken duration. Like the Salame and Baddeley (1982) model, our connectionist model behaves in the same way as human subjects because of the temporal characteristics of the rehearsal procedure: long items (those with many phonemes) take longer to rehearse, for rehearsal time is proportional to the number of phonemes (one pass through the network is necessary for each time-slice of the material to be remembered). And the longer the rehearsal time before an item can be refreshed, the more likely it is that the traces of earlier items will have decayed to the extent that they cannot be retrieved. The experimental manipulation **articulatory suppression**, which requires subjects to recite irrelevant material aloud at the same time as remembering a sequence of auditorily or visually presented items, has its effect in the model by making the speech production net unavailable. Thus the word length effect, reflecting the rehearsal procedure, is abolished by articulatory suppression. There is some residual memory capacity even under suppression conditions, arising from visual and semantic coding; we have not yet modelled these sources of capacity in any detail. (There is a need, for example, to account for the fact that articulatory suppression has differential effects across varying serial position.) **Phonemic confusability effects**, which are widely assumed, as here, to reflect the operation of the phonological store rather than the articulatory rehearsal procedure, are also abolished by suppression when material is visually presented, because visually presented material can only gain access to the phonological store via the rehearsal procedure. In contrast, auditorily presented material can show phonemic confusability effects, because this material can gain access to the phonemic store via the recognition side of the rehearsal procedure. This modality-dependent behaviour of the model is consistent with the observations and model of Baddeley et al. (1984). We have not yet examined the mechanisms of confusability effects in the model in detail, due to computational resource constraints and the need for a larger vocabulary, but they are assumed to arise due to interference in the phonological store. As in the interactive activation model of word recognition, the probability of being able to identify an item is assumed to reflect the level of activation of that item's units *relative to* the activation of units for other items. And when items share phonemes, their total levels of activations over phonemes are relatively more similar, leading to difficulty in identification. The **retention of order information** is generally believed to be an important function of STM (Healy, 1974); in our model (as in other models of STM) order information is represented simply in terms of the extent to which the activation if an item code has decayed in the phonological store. The model has ready access to this information for other purposes, and can use the decay levels as order markers without the need for further mechanisms inside the phonological store. This appears to provide a relatively efficient method of encoding order information for humans, for such information is more likely to be lost whenever phonological STM is made unavailable (but cf. Grossberg & Stone, 1986). Effects of **lexicality**, **imageability** and **visual confusability** on STM capacity are assumed in the model to result from non-phonological sources of activation that help to maintain the activation level of lexical input units. Thus, they simply provide an alternative source of input in addition to refreshment by the output of the articulatory rehearsal procedure. **Chunking** effects have a similar source, in that they are assumed to arise from the (as yet underspecified) coalitions of units that can be brought to bear on the recall process. It has been suggested that **item identification time** may be independently related to memory span for those items (Dempster, 1981): Our implementation assumes that the input to the rehearsal process can be seen as the output of a word identification process, and so if items take a long time to load into the rehearsal procedure, there will be correspondingly more time for the codes of other to-be-remembered items to decay. Indeed, the model reported here was designed as an extension and development of an earlier computational model of single word reading (Brown, 1987a, 1987b).

While the model can account for a wide range of data as it stands, it is assumed that a more complete model, which includes more sub-components of the working memory system, will be

required to account for suffix effects and aspects of retroactive and proactive inhibition as well as the use of retrieval cues. Those parts of the **serial position curve** that are sometimes assumed to reflect rehearsal and transfer to LTM are consistent with the current version of the model, and it is assumed that there is an additional, passive storage mechanism responsible for recency effects. The model as it stands also assumes an outside source of strategic control (deciding when to rehearse), and some binding mechanism so that the model can distinguish different tokens of the same word. In addition, we note that the model builds on speech processing mechanisms that have been criticized for requiring a segmented input stream.

DISCUSSION

The model provides a connectionist, psychologically plausible account of the way in which mechanisms of speech perception and production can be recruited to serve as a temporary storage system. The suggestion that temporary phonological storage capacity is available as a by-product of the language processing system is a well-established one (Ellis, 1979), but computationally explicit mechanisms have been lacking. The model is essentially a connectionist implementation of the Baddeley model of the articulatory loop (Baddeley, 1986; Salame & Baddeley, 1982; cf. also Schneider & Detweiler, 1987). Our model accounts in a similar way for the limited capacity of human short-term memory, in that it is only possible for a temporally limited amount of material to be rehearsed by the network before information decays beyond recall. Similar reasoning can be used to explain the developmental increases observed in temporary memory capacity, as well as providing an explanation of word length effects and the ability of STM to encode order information. We are currently extending the model and investigating its ability to account for developmental phenomena in particular. The model is being trained with a larger vocabulary, represented in terms of acoustic features rather than phonemes as at present, for empirical evidence demonstrates that confusions in STM can occur at sub-phoneme levels.

Our approach is motivated by the belief that rehearsal processes, as characterized in current cognitive models, are a ubiquitous feature of human cognition, and there are good reasons for this which are illustrated by reference to our model. If a trace is refreshed via the normal perception and production mechanisms, which are available at no extra cost to the organism, the maintenance of the trace can take advantage of what is know about the perceptual structure of the world, for these regularities are encoded in the perception and production mechanisms. This contrasts with the case of simple resonance, where units can remain active simply by passing activation backwards and forwards without making use of perception and production mechanisms and the regularities implicit therein.

ACKNOWLEDGEMENTS

This research was supported by grants from the Medical Research Council (U.K.) (1989) and the Leverhulme Trust (1988-1990). I thank Dennis Norris for useful discussions.

REFERENCES

BADDELEY, A.D. (1986). *Working memory*. Oxford: OUP.

BADDELEY, A.D. & HITCH, G.J. (1974). Working memory. In G. Bower (Ed.), *Advances in the psychology of learning and motivation 8,* New York: Academic Press.

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

BADDELEY, A.D., LEWIS, V. & VALLAR, G. (1984). Exploring the articulatory loop. *Quarterly Journal of Experimental Psychology,* 36A, 281-289.

BROWN, G.D.A. (1987a). Resolving inconsistency: A computational model of word naming. *Journal of Memory and Language* , 23, 1-23.

BROWN, G.D.A. (1987b). Constraining interactivity: Evidence from acquired dyslexia. *Proceedings of the Ninth Annual Conference of the Cognitive Science Society* , 779-793. Hillsdale, NJ: Lawrence Erlbaum Associates.

DEMPSTER, F.N. (1981). Memory span: Sources of individual and developmental differences. *Psychological Bulletin,* 89, 63-100.

ELLIS, A.W. (1979). Speech production and short-term memory. In J. Morton & J.C. Marshall (Eds.), *Psycholinguistic series Vol 2: Structures and processes*. Cambridge, Mass: MIT Press.

ELLIS, A.W., & YOUNG, A.W. (1988). *Human cognitive neuropsychology*. Hillsdale, NJ: Lawrence Erlbaum Associates.

ELMAN, J.L. (1988). *Finding structure in time*. CRL Technical Report 8801, University of California, San Diego.

GROSSBERG, S., & STONE, G. (1986). Neural dynamics of attention switching and temporal order information in short term memory. *Memory & Cognition*, 14 (6), 451-468.

HEALY, A.F. (1974). Separating item from order information in short-term memory. *Journal of Verbal Learning and Verbal Behavior*, 13, 644-655.

HULME, C. & MUIR, C. (1985) Developmental changes in speech rate and memory span: A causal relationship? *British Journal of Developmental Psychology,* 3, 175-181.

JORDAN, M.I. (1986). Attractor dynamics and parallelism in a connectionist sequential machine. *Proceedings of the Eighth Annual Conference of the Cognitive Science Society*, Hillsdale, NJ: Lawrence Erlbaum Associates.

McCLELLAND, J.L., & RUMELHART, D.E. (1988). *Explorations in parallel distributed processing: A handbook of models, programs and exercises*. Cambridge, Mass: MIT Press.

MONSELL, S. (1987). On the relation between lexical input and output pathways for speech. In A. Allport, D. MacKay, W. Prinz & E. Scheerer (Eds.), *Language Perception and Production*. New York: Academic Press.

MOZER, M.C. (1987). Early parallel processing in reading: A connectionist approach. In M. Coltheart (Ed.), *Attention and performance XII: The psychology of reading*. Hillsdale, NJ: Lawrence Erlbaum Associates.

NORRIS, D. (1989). Dynamic net model of human speech recognition. In G.T. Altmann (Ed.) *Cognitive models of speech processing: Psycholinguistic and computational perspectives.* Cambridge, Mass: MIT Press (in press).

RUMELHART, D.E., HINTON, G.E., & WILLIAMS, R.J. (1986). Learning internal representations by error propagation. In D.E. Rumelhart and J.L. McClelland (Eds.), *Parallel Distributed Processing Vol 1*. Cambridge, Mass: MIT Press.

SALAME, P., & BADDELEY, A.D. (1982). Disruption of short-term memory by unattended speech: Implications for the structure of working memory. *Journal of Verbal Learning and Verbal Behavior*, 21, 150-164.

SCHNEIDER, W., & DETWEILER, M. (1987). A connectionist/control architecture for working memory. In G.H. Bower (Ed.) *The psychology of learning and motivation vol 21*. New York: Academic Press.

SCHRETER, Z., & PFEIFER, R. (1989). Short-term memory/long-term memory interactions in connectionist simulations of of psychological experiments on list learning. In L. Personnaz and G. Dreyfus (Eds.), *Neural networks: From models to applications*. Paris: I.D.S.E.T.

SCHWEICKERT, R., & BORUFF, B. (1986) Short-term memory capacity: Magic number or magic spell? *Journal of Experimental Psychology: Learning, Memory and Cognition,* 12 (3), 419-425.