A connectionist model of auditory word perception in continuous speech

Richard Shillcock — Centre for Cognitive Science, University of Edinburgh, 2 Buccleuch Place, Edinburgh, rcs@cogsci.ed.ac.uk

Joe Levy — HCRC, University of Edinburgh, 2 Buccleuch Place, Edinburgh, joe@cogsci.ed.ac.uk

Nick Chater — Department of Psychology, University of Edinburgh, 7 George Square, Edinburgh, nicholas@cogsci.ed.ac.uk

Abstract

A connectionist model of auditory word perception in continuous speech is described. The aim is to model psycholinguistic data, with particular reference to the establishment of lexical percepts. There are no local representations of individual words: feature-level representations are mapped onto phoneme-level representations, with the training corpus reflecting the distribution of phonemes in conversational speech. Two architectures are compared for their ability to discover structure in temporally presented input. The model is applied to modelling the phoneme restoration effect and phoneme monitoring data.

Introduction

The recognition of a spoken word involves correctly matching some representation derived from the acoustic input, often of poor quality, with some stored representation. The nature of the recognition problem is partly determined by the frequency distribution of elements in spoken language. This paper presents a model of how these distributional statistics can effect the recognition of spoken words. The aim is to capture as much of the relevant psycholinguistic data as possible in terms of statistical properties of streams of phonetic and phonemic input, without adverting directly to explanation at the word level[1].

The process of lexical access

Speech sounds arrive over time and must be matched against some kind of stored representation. Psycholinguistic theories have variously proposed that such input is represented in terms of features, phonemes, morphemes and words. The Cohort Model (Marslen-Wilson & Welsh 1978; Marslen-Wilson 1987) has generated a "lexicalist-localist" tradition in which the incoming signal is seen as directly contacting specific lexical representations, which are activated in proportion to their match with the input. These activated representations then compete until one of them is uniquely distinguished by the input and/or is integrated into the ongoing interpretation of the utterance. That is, it is assumed that contacting the lexical entry for the word *automatic*, for instance, makes available all of its associated information (pronunciation, orthography, semantics, and so on) and that once *automatic* is alone in the cohort (at some point during its third syllable), then it exclusively determines processing from that point until the end of the word.

Within this approach, less has been said about the development of the representation which makes initial contact with the lexical entry (/ɔ/, /ɔt/, /ɔtə/... for *automatic*). There are different claims concerning the types of information which may influence the activation of lexical representations, the effect of non word-initial partial matches (/ɔt/ in the input matching *porter* or *short* in the lexicon), the mechanism which mediates competition between activated lexical representations, and the continuing role of representations which cease to match the input. However, one computationally explicit account, the TRACE model (McClelland & Elman 1986), based on the early Cohort Model, has captured many aspects of human spoken word recognition in a principled way and represents a coherent stance on the issues mentioned above. For instance, constraining the activation of lexical representations and segmenting the continuous input are two major issues in spoken word recognition, and TRACE provides a computationally explicit answer to both.

Three aspects of TRACE suggest avenues for further research. First, the implementation of TRACE is limited to 15 different phonemes; it is desirable to be able to model the full scale and richness of human word recognition both to handle real discourse and to be able to assess the model's performance on actual stimulus materials taken from psycholinguistic experiments. Second, TRACE does not learn; TRACE's knowledge of the language is confined to the word frequency values which are built into its lexicon. Third, TRACE's lexical level mediates influence between adjacent phonemic material. An alternative possibility is that processing may be adequately captured just by statistical dependencies at the phonemic level. For example, there is a TRACE/Cohort prediction that monitoring of a word-medial phoneme, like /t/ in *curtail* should be facilitated to the extent that it lies on, or close to, the word's

[1] We would like to thank the following: Geoff Lindsey for advice concerning the phonemic and featural descriptions, Steve Finch for assistance and advice with bigram and trigram statistics, Alex Monaghan for the use of the CSTR text-to-phonemes program, Ellen Bard, Henry Thompson and Richard Rohwer for valuable discussion.

uniqueness point. At this point only one lexical representation remains in the cohort, the listener has all the information necessary to identify the word and its constituent phonemes. In a TRACE simulation, *curtail* would become the most highly activated word node at or just beyond the uniqueness point and would supply top-down information to the phoneme-nodes representing its constituent phonemes. If, however, it could be demonstrated that the perception of the /t/ in *curtail* could be captured by a model which had no local lexical nodes, and had had no previous experience of the word *curtail*, then we would be justified in preferring this simpler explanation.

The model advanced below is motivated by the belief that it is necessary to account for as much of the data as possible on the basis of processing in which there are no explicit exclusive lexical representations. The model makes no distinction between representations of any frequent sequence, whether it be specifically a morpheme, a syllable, a word or an idiom. This model builds on the recent departure from the lexicalist-localist view of lexical access, involving a distributed connectionist model, as described in the next section . This perspective does not rule out the possibility that explicit specific lexical representations might be necessary to account for certain data. Only after investigating exhaustively how much of the data can be accounted for by a model which does not possess such representations can the role of lexical representations in explaining psycholinguistic data be properly assessed.

More recently a second computationally explicit model of spoken word recognition has been presented by Norris (1988), which employs an architecture also investigated by Elman (1988, 1990). In Norris' model feature-level representations of consecutive phonemes are mapped onto local representations of words via one layer of hidden units. Recurrent connections from the hidden units copy their pattern of activity to a set of state nodes which then re-present this pattern to the hidden units at the next time-step. Thus the network has the potential to respond to patterns of phonemes across time. (This approach is described in more detail below.)

Norris's model is architecturally more elegant than TRACE. It learns the frequency of the words it can recognize from its training set, and, as Norris (1988) reports, it captures a range of "cohort behaviours". In simulations with miniature lexica the model generally assigns a spread of activity to all words which are congruent with the input up to and including the current phoneme. At the uniqueness point of a word the model generally opts overwhelmingly for that word and maintains its level of activation until the end of that word in the input. Again, three aspects of the model suggest possible further research. First, the simulations which Norris reports are all with small scale lexica (each word which the model can recognize is given a specific output node). Second, there is a considerable volume of psycholinguistic data which addresses infra-lexical processing (*e.g.* phoneme-monitoring) and which it is not feasible to model using the activation levels assigned to whole words which are the output of Norris' model. Third, the inclusion of a specifically lexical level, the output level of Norris' model, is unduly constraining on any wider model of sentence processing. It prevents the model from learning sub-word regularities, except to the extent that word-initial similarity occasions activation of a cohort, and also super-word regularities, unless strings like *that's*, or *out of* or *good morning* are lexicalized by being given a dedicated output node.

The model we present in Section 4 resembles Norris' model in network architecture and in involving a mapping from feature-level descriptions. Crucially, however, the model does not contain local representations of words; a number of advantages spring from this fact. Before describing the model, however, it is necessary to motivate further the exclusion of local lexical representations.

Seidenberg and McClelland's model of pronunciation

Considerable coverage of psycholinguistic data has been achieved with Seidenberg and McClelland's (1989) connectionist modelling of word naming and visual word recognition. In modelling naming, orthographic representations are mapped onto phonological representations; there are no local representations of words, only weights between the three layers of nodes. An extension of the model, involving an identity mapping of the orthographic level, captures aspects of visual word recognition. Crucially the training regime reflects the frequency with which the words appear in the language.

How can word recognition data be modelled when there are no local representations of words and therefore no activation levels which might be assigned to specific words? As an example, Seidenberg and McClelland argue that in tasks in which subjects are required to discriminate between orthographically regular words (*e.g. fellow, tanker*) and orthographically irregular nonwords (*e.g. fnrkte, jplerhn*), their performance may be accurately modelled by the accuracy with which an identity mapping may be made between two orthographic levels mediated by a layer of hidden units and trained by back-propagation.

In the model of spoken word recognition described below, an analogous approach is taken within the auditory domain: feature-level representations are mapped onto phoneme-level representations. The training regime is taken from spoken discourse and reflects the frequency with which speech sounds corresponding to phonemes occur and co-occur in spoken language.

The problem of learning the structure of temporal sequences

Neural network methods have been developed largely to learn to classify static patterns. Since many important aspects of cognition involve processing temporally structured sequences, there has been considerable attention devoted to extending network methods to learn the structure of time-varying sequences. One strategy is

simply to represent a "moving window" of past inputs explicitly (used, for example in the NETTALK model of reading (Sejnowski & Rosenberg 1987)). Explicit buffering of past input can be avoided by using recurrent connections which recirculate past input so that it may continue to have an influence on network performance, rather than being "flushed through" the network. There has been much recent work on various ways in which back-propagation can be generalized to recurrent networks (Rumelhart, Hinton & Williams 1986; Almeida 1987; Pineda 1987; Pearlmutter 1990). In some of these regimes, learning occurs when the net has settled into a stable pattern, and in some continuous time signals are used. Rumelhart, Hinton & William's original suggestion, "back-propagation through time", which simply unfolds a recurrent network into many copies arranged in a feed-forward architecture and applies standard back-propagation to the result, is the most appropriate for this kind of task. The more time-steps back the network is unfolded, the better the network will be able to learn to respond to temporally remote information, but at greater computational expense.

The model

The model, illustrated schematically in Fig. 1, simply consists of a mapping between two levels of representation within the auditory modality – a feature level and a phoneme level.

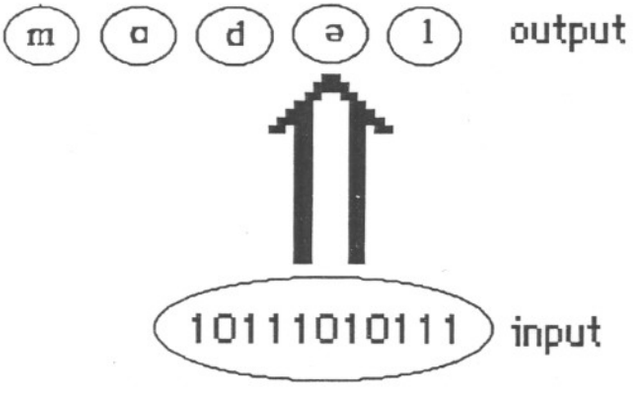

Figure 1: Feature to phoneme mapping. The input is a binary feature-level description of the current segment. The output is the identity of the current phoneme, the predicted next phoneme and a number of previous phonemes.

Input is a bundle of features corresponding to the segment at that point in time. This input is replaced by successive bundles of features corresponding to consecutive segments across time. The output is a phoneme-level description of the current segment, together with a prediction of the identity of the next segment and a confirmation of the identity of the last several phonemes. The simplest version of the model possesses a three-segment output window – current, predicted and last.

Two phonemes may share a number of features (/p/ and /t/ are identical on all but one). In the hypothetical case where two phonemes possess the same feature-level instantiation (as a result of noise, for instance), the only way in which they may be distinguished is by their surrounding context: thus the phoneme designated * will be classified as /s/ in /y e */ and as /f/ in /i */, given that its feature-level description is ambiguous between /s/ and /f/ and it has been trained on the words *yes* and *if*. In the model, each phoneme possessed a unique feature-level description. This resulted in "current" identifications approximating to 100% and remaining at that level in the "last" position, in which confirmation was expected. In human speech perception, however, segments are often underspecified since the signal is noisy, and information relevant to the identification of any one segment is often spread over several surrounding segments. In some of the simulations below we rely on the addition of noise to the signal during training to encourage the network to rely on the phonemic context since the information may no longer be encoded in a single segment.

The mapping in Fig. 1 may be achieved by means of any one of a family of networks which are sensitive to structure across time. The minimum sensitivity to such structure involves replicating simple bigram probabilities. The extent to which the networks under study are sensitive to more than bigram probabilities is an empirical question. Comparison with simple bigram and trigram statistics is a powerful means of assessing the performance of the models (although detailed analysis is not reported here). Finally, it is important that the network be tested with the full extent of the feature to phoneme mapping in the language. Below we report results using the Elman/Norris net and a feedforward net incorporating a moving window.

The networks

The basic mapping between the two levels of representation was achieved using a "cut-down" version of back-propagation through time (Rumelhart, Hinton & Williams 1986), unfolding the network once rather than many times (Servans-Schreiber, Cleeremans & McClelland 1989; Chater 1989) and thus sacrificing the ability reliably to pick up long distance dependencies, in exchange for speed of training. This "copyback" structure (Fig. 2) was introduced by Elman (1988, 1990) and Norris (1988) .

There were 11 input units, 15 hidden units, and hence 15 corresponding "copyback" units (which retained the hidden unit activations from the previous timeslice) and 108 output units (coding the 36 phonemes at the previous, current and next time-step). For a qualitative comparison, a simple "moving-windows" architecture, also with 15 hidden units (with 22 additional input units representing the phonetic features at the previous two time steps), was implemented. The architecture was that of a standard feedforward network, trained with back-propagation, where the input layer represents not just the current phonetic input, but also the phonetic input at

previous time steps. In our simulations, the moving window extended over three time-steps.

Three simulations are reported: the recurrent network architecture was trained both on noisy and non-noisy inputs, and the moving window architecture was trained only on non-noisy input data. Using only the bigram statistics of the stream of input data, and assuming that error on the previous and current phoneme is 0, the best total sum of squares error over the training corpus is 8044 steps. Of course, by considering higher order statistics, better performance is possible, though a large corpus is required to obtain reliable higher order statistics.

This figure is comparable with the performance of the networks: 8023 for the recurrent network trained on the noise-free corpus, 9788 when the training data is noisy and 7892 for the moving windows architecture with noise-free input. The slight difference in performance between the recurrent network and the moving windows architecture results from better performance of moving windows at outputting the previous phoneme – in this architecture the phonetic input at the previous time step is presented as part of the input, whereas the recurrent network must learn to buffer this information. That network performance is comparable with the results of a bigram analysis does not of course imply that the network is responding only to bigram structure – in fact, it seems more likely that it is picking up some higher order statistics of the input, while not perfectly accounting for bigram statistics. This is currently being tested by comparing network performance of the net on real data versus *n*th-order approximations to that data.

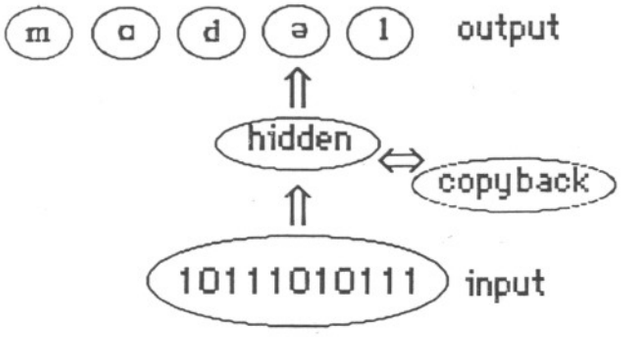

Figure 2. Structure of the Elman/Norris network. Input is the feature-level representation of the current segment. Output is the classification of the current segment, prediction of the next segment, and confirmation of the last several segments. The copyback units re-present to the hidden units the pattern of activation on those units at t-1.

In all the simulations reported, a learning rate of 0.1 was used, and momentum was not used. Each network was trained until it began to show signs of overfitting – training that resulted in a decrease of error for the training set but led to increasing error for a separate test set was disregarded. This required between 500 and 600 epochs. In some simulations, the phonetic input was made noisy, by randomly changing 9% of the input values from 0 to 1 or *vice versa*. The noise was generated on-line and was different for every epoch of training. The learning phase of the simulations was quite computationally intensive, using 30-40 CPU hours on a variety of SUN SPARC-based machines, using a customized version of the Rumelhart and McClelland (1988) simulation package.

The corpus and training regime

The initial, limited training data was derived from some 3490 words of spoken discourse, taken largely from the LUND Corpus (Svartik & Quirk 1980). The discourse was transcribed at the word level and included filled pauses, false starts and corrections. The training set was made up of 9097 phonemes and a test set of 3285 phonemes was used to test generalisation/overfitting. In some later simulations the training data have been generated from a 33,000 word phonemic dictionary, containing frequency information; this allowed better exposure to open-class words while sacrificing some of the character of the distribution of closed-class words.

The phoneme-level descriptions

The utterances were converted to idealized phonemic representations using the CSTR text-to-phoneme program and employing 36 different phonemes based on those of the CSTR Machine Readable Phonetic Alphabet. The eight dipthongs were each converted to sequences of two phonemes.

The feature-level descriptions

The phonemic transcription was then converted to an idealized feature-level representation, consisting of the following 11 features based on those of Jakobson, Fant and Halle (1952): *vocalic/non-vocalic, consonantal/non-consonantal, voiced/unvoiced, discontinuous/continuant, strident/mellow, nasal/oral, diffuse/non-diffuse, compact/non-compact, tense/non-tense, grave/acute, flat/plain*. Thus the phonemes /ə/ and /I/ were represented on the respective features as below.

ə=	1	0	1	1	0	0	0	0	0	1	0
I =	1	0	1	1	0	0	1	0	0	0	0

All 36 phonemes were given a value of 1 or 0 for each of the 11 features. The final form of the training corpus was of a continuous stream of such feature-level descriptions of segments, with no information being given concerning word boundaries.

Behaviour of the networks

To illustrate the behaviour of the networks with unseen discourse, consider the test string *this is a test of the model*. When this was converted into sequences of feature descriptions and given to the networks, all three networks demonstrated sensitivity to phonotactic constraints, with greater probabilities of predicting the

next phoneme in the short sequences of closed-class words and in the unstressed syllable of *model*. The models all predict the next phoneme more accurately when the sequence is from normal discourse than when the same phonemes are presented in random order. The effect of training with noise was to depress the scores given to phonemic hypotheses and to increase the number of hypotheses which received a non-zero score. In the simulation reported here, this increased noise did not result in better performance in terms of the ranking of the correct hypothesis within the total list of hypotheses.

It is easier to see the networks' use of context in the case of the classification of the current phoneme, where all of the information necessary for the unique identification of the phoneme is present on that presentation. For the Elman/Norris net trained with noise, performance was worse in the scrambled phonemes case compared with the normal discourse case (mean phoneme scores were 66.6 and 79.1, respectively; $t = 2.35$, $df = 18$, $p = .031$). We may conclude that in the noise condition this network was relying on previous context to identify the current phoneme; when this context was aberrant, it hindered correct recognition. The Elman/Norris net trained without noise and the moving-window net both generated reduced mean "current" phoneme scores for the scrambled input but neither of the differences was significant.

Human listeners employ context both before and after the phoneme in question. The aim of training with noise was to force the network to rely on both "left" and "right" context. The scores for the phoneme in "past" position were compared on the normal and abnormal discourse. While there was no significant difference between the two mean scores for the Elman/Norris net trained without noise, the version trained with noise was significantly worse on the abnormal discourse (means were 82.7 and 63.2, respectively; $t = 3.88$, $df = 18$, $p = .001$.). The network was sensitive to right context in classifying phonemes, and was misled by abnormal right context.

Modelling psycholinguistic data

Phoneme restoration

Listeners' perception of degraded individual speech sounds in words is often restored (Warren 1970). Restoration is strongest when the intended phoneme and the replacing sound (*e.g.* white noise, a click, silence) are similar, and when replacement occurs after the uniqueness point of the word. Otherwise the effect is not compelling.

This was modelled by putting minimally different test words, like *got* and *gop*, in the carrier sentence *...and the next word is x and the next word is y and the....* For frequent words like *got*, *this*, and *yes*, there was no substantial restoration. For example, when the current phoneme was /i/ in *thif*, the prediction for /s/ was 14 and the prediction for /f/ was 4 (*if* is a frequent word); when the current phoneme became /f/, it was scored at 98, compared with 2 for /s/. The frequency of *this* was not enough to overturn the bottom-up information.

Restoration was observed, however, when the value of the critical feature distinguishing /f/ and /s/ was replaced by 0.5. Input, for the *this/thif* case, was then /ð i */ where * was completely ambiguous between /f/ and /s/. In this case, the current scores were 99 for /s/and 5 for /f/, changing to 98 and 0 respectively at confirmation.

The model respects the input. It does not hallucinate phonemes on the basis of word frequency. This captures the effect more accurately than TRACE, in which lexical level reinforcement restores any degraded phoneme and even overturns bottom-up evidence, converting *vocabulaty* to *vocabulary*

Phoneme restoration also occurred based purely on the right context. When the input was /* e s/, in which * was ambiguous between /y/ and /r/, the model restored the /y/.

Monitoring for word-medial phonemes

Simulations were run to test whether the model predicted the data from an experiment (Shillcock *submitted*) in which subjects monitored for word-medial phonemes like /p/ in *repel* or /p/ in *lapel*. Subjects in the experiment took significantly longer to respond to phonemes in monomorphemic words compared to matched prefixed words.

The stimulus materials from the experiment were embedded in the context *...and the next word is...and the next word is...* and presented to the three trained networks. Activations for the critical phoneme in each word (/p/ in *repel*) were recorded when that phoneme was in "next", "current" and "past" position. Only the moving window network gave significantly different mean activations for the monomorphemic words and the prefixed words, mirroring the human data (97.7 and 98.7, respectively; $df = 14$, $t = 2.137$, $p = .05$). This difference occurred in the "current" position, reflecting the fact that response times were facilitated if the sequence of phonemes up to and including the critical phoneme represented a prefix as opposed to a monomorphemic word beginning. When activations in the simulations for the individual stimulus items were compared with the mean response times from the experiment, there was no significant correlation. Shillcock (*submitted*) reports that the best predictor ($r = -.477$) of the phoneme-monitoring data was the (frequency-weighted) number of times the sequence of phonemes up to and including the critical phoneme (*i.e.* /r I p/ for *repel*) occurred in a large phonemically transcribed dictionary (*i.e.* /r I p/ in *script, report, unrepentant...*, all weighted by word frequency). The networks were therefore not employing information as relevant as the bigram and trigram information available from a large phonemic dictionary. This may reflect most on the size of the training corpus.

Conclusions

Initial testing of the model gives encouraging results, with the various simulations from the different architectures demonstrating desirable behaviours. Many of its limitations may be due to the modest size of the training corpus: the 3490 words in the corpus represented 905 different words. A very large corpus will be required

to ensure adequate coverage of the open-class vocabulary. Training with a corpus of transcribed discourse opens the possibility of studying what special processing of the closed-class vocabulary may emerge; the literature contains numerous claims concerning the special status of the closed-class vocabulary compared with the open-class vocabulary. Yet further work will be required concerning the amount and nature of (idealized) phonological reduction in the corpus.

Regarding network architecture, it may be that the most promising avenue for future research lies in a network which allows back-propagation through time for several rather than just one time-step. We are currently exploring this avenue.

The model described is seen as part of a larger model incorporating semantic representations. The absence of explicit, localist lexical representations is crucial. We envisage a mapping from the phonemic output of the model to semantic representations, mediated only by a layer of hidden units, although this clearly raises a serious binding problem in the absence of word boundary information. This arrangement gives more scope than a lexicalist-localist model for modelling the details of effects in which homophones and partial homophones produce brief erroneous priming.

In conclusion, it may be that many psycholinguistic phenomena which have been taken to involve access to specific representations of spoken words may be explained in terms of the low-level statistical structure of the phonetic/phonemic input, as picked up by a simple neural network account. We are currently applying this model to a range of other experimental phenomena. There is a methodological imperative within psycholinguistic research to allow "higher level" interpretation of empirical data only when low level explanations can be ruled out.

References

Almeida, L. B. 1987. A learning rule for asynchronous perceptrons with feedback in a combinatorial environment.. In Proceedings of the IEEE First International Conference on Neural Networks, San Diego, California, June, 1987.

Chater, N. 1989. Learning to respond to structures in time, RIPRREP 1000/62/89 Research Initiative in Pattern Recognition, St Andrews Road, Malvern, Worcs., U.K.

Chater, N. Responding to temporal structure. Submitted to *Cognition*.

Chater, N. & Ganis, G. Double dissociation in distributed systems. Submitted to *Psychological Review*

Elman, J. L. 1988. Finding structure in time. Technical Report, CRL TR 8801, Centre for Research in Language, UCSD.

Elman, J. L. 1990. Finding structure in time. *Cognitive Science*, 14:179-211.

Jakobson, R., G. Fant and M. Halle 1952. Preliminaries to speech analysis. Technical Report 13, M.I.T. Acoustics Laboratory, Mit Press.

Jordan, M. 1986. Serial order: a parallel distributed approach. Institute for Cognitive Science Report, 8604, University of California, San Diego.

Juang, B. H. 1984. On the hidden Markov model and dynamic time warping for speech recognition - a unified view. *Bell Systems Technical Journal*, 63, no.7, 1213-1243.

Marslen-Wilson, W. 1987. Functional parallelism in spoken word-recognition. In U. H. Frauenfelder & L. K. Tyler eds *Spoken word recognition*, 71-102, Special Issue of Cognition.

Marslen-Wilson, W. & Welsh, A. 1978. Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*, 10, 29-63.

McClelland, J. L. & Elman J. L. 1986. Interactive processes in speech perception: the TRACE model. In D. E. Rumelhart & J. L. McClelland eds. *Parallel Distributed Processing*, Vol. 2., 58-121, Cambridge, Mass: MIT Press.

McClelland, J. L. & Rumelhart, D. E. 1988. *Explorations in Parallel Distributed Processing: Models, Programs and Exercises*. Cambridge, Mass: MIT Press.

Minsky, M. & Papert, S. 1969. *Perceptrons*. Cambridge, Mass: MIT Press.

Norris, D. G. 1988. A dynamic-net model of human speech recognition. Talk given at the Cognitive Models of Speech Processing Workshop, Sperlonga 1988. See G. Altmann ed. 1990, *Cognitive Models of Speech Processing*, MIT Press.

Patterson, K. E., Seidenberg, M. S. & McClelland, J. L. 1989. Connections and disconnections: acquired dyslexia in a computational model of reading processes. In R. G. M. Morris ed., *Parallel distributed processing: implications for psychology and neurobiology*. Oxford, Oxford University Press.

Pineda, F. 1987. Generalization of back-propagation to recurrent and higher order neural networks. *Neural Information Processing Systems*, New York.

Seidenberg, M. S. & McClelland, J. L. 1989. A distributed, developmental model of word recognition and naming. *Psychological Review*, 96, 523-568.

Sejnowski, T. J. & Rosenberg, C. R. 1987. Parallel networks which learn to pronounce English text. *Complex Systems*, 1, 145-160.

Servans-Schreiber, D., Cleeremans, A. & McClelland, J. L. 1989. Learning sequential structure in simple recurrent networks in D. Touretsky ed. *Advances in Neural Information Processing Systems*, Vol 1, Morgan Kaufman, Palo Alto, 643-653.

Svartvik, J, & Quirk, R. 1980. *A corpus of English conversation*. Lund: Gleerup.

Shillcock, R. C. 1990. The processing of prefixed words: a connectionist account. Submitted to *Memory & Cognition*

Waibel, A., Hanazawa, T., Hinton, G. E., Shikano, K. & Lang, K . 1987. Phoneme recognition using time-delay neural networks. ATR Technical Report TR-I-0006 Japan: ATR Interpreting Telephony Research Laboratory.

Warren, R. M. 1970. Perceptual restoration of missing speech sounds. *Science*, 167, 392-393.