Processing time-warped sequences using recurrent neural networks: Modelling rate-dependent factors in speech perception

Mukhlis Abu-Bakar*
Neural Networks Research Group
Department of Psychology
University of Edinburgh
Edinburgh EH8 9JZ
U.K.
mukhlis@cogsci.ed.ac.uk

Nick Chater
Neural Networks Research Group
Department of Psychology
University of Edinburgh
Edinburgh EH8 9JZ
U.K.
nicholas@cogsci.ed.ac.uk

Abstract

This paper presents a connectionist approach to the processing of time-warped sequences and attempts to account for some aspects of rate-dependent processing in speech perception. The proposed model makes use of recurrent networks, networks which take input one at a time and which can pick up long-distance dependencies. Three recurrent network architectures are tested and compared in four computational experiments designed to assess how well time-warped sequences can be processed. The experiments involve two sets of stimuli, some of which reflect aspects of rate dependent processing in speech; one where the sequences are distinguished by the way their constituent elements are sequentially ordered, and another where the sequences share similar arrangement of the constituent elements but differ in the duration of some of these elements. The results establish certain conditions on rate-dependent processes in a network of this type vis-a-vis the obligatory use of rate information within the syllable, and throw some light on the basic computer science of recurrent neural networks.

1. The problem of time-warped sequences

Time-warping of utterances occur frequently in conversational speech. This results from speakers' tendency to speed up and slow down when they talk rather than maintain a constant rate of speech. The variation in rate that occurs in conversational speech can be substantial (Miller, Grosjean & Lomanto 1984). At the lexical level, such time-warping can be seen as a distortion in the temporal structure of words so that some parts of the word may be compressed, others stretched, and some remain durationally invariant to changes in the speech rate. Yet listeners appear to have little difficulty making the appropriate perceptual adjustments for these variations.

The problem is more complex at the phonetic level. As articulation time is altered due to changes in the speech rate, certain acoustic properties that specify the identity of phonetic segments are modified, since they are themselves temporal in nature. For instance, a short duration of some property may specify one phonetic segment while a longer duration specifies another (Lisker & Abramson 1964). This may cause a problem for deriving the phonetic structure of an utterance. Again, listeners are able to maintain perceptual constancy in the face of such changes (Summerfield 1981).

To-date, there have been various standard attempts at solving time-warping problems (e.g., hidden Markov modelling (Huang, Ariki & Jack 1990) and dynamic programming (Sakoe & Chiba 1971)). However, these are essentially engineering in design and claim little psychological relevance. The present work attempts to fill this gap by offering a processing account of the rate adjustment process using connectionist tools.

2. Using recurrent backpropagation

Recurrent neural networks have been widely used in modelling sequence processing (e.g. Elman 1990). The presence of recurrent connections gives the network the opportunity to store information about past items, and thus to respond on the basis of the sequence as a whole, rather than just the present input item. In the present application, the network must ascertain the rate of speech from the sequence of past input, and use this information to classify later material. The networks are trained to classify input sequences into a small number of categories, in some cases corresponding to different syllables.

These networks are trained by recurrent backpropagation (Rumelhart, Hinton & Williams 1986) in which the recurrent network is unfolded into a feedforward network (Fig 1). This network is trained using conventional backpropagation, with the constraint that the weight changes for each link in the original recurrent network are the sum of the weight changes for each copy of that link in the unfolded network, so that the feedforward network can be folded back up into a recurrent form. This is a slightly different training regime from that used in related work (Elman 1990; Norris 1990), but this method is preferred since recurrent backpropagation appears to be better at encoding information across many items in a sequence (see Chater and Conkey 1992 for dis-

* Also Department of Linguistics, University of Wales, Bangor, Gwynedd LL57 2DG, U.K.

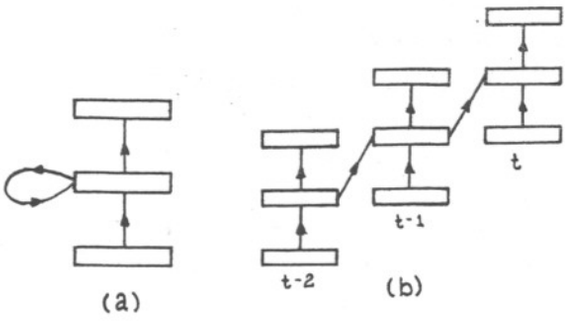

Figure 1. (a) A folded recurrent network
(b) Network unfolded through three time-steps.

cussion). All simulations were implemented using the Xerion simulator (van Camp, Plate & Hinton 1992).

2.1. Comparing architectures

Perhaps the main advantage of using recurrent networks is their ability to treat speech as a sequence of events and take input one at a time. We trained the network by feeding it with the relevant sequences one element at a time and keeping the target output pattern present throughout the presentation of each sequence. The production of the correct output when the sequence is presented indicates that the sequence has been classified successfully. If performance is optimal, correct classification should occur after the "recognition point" of the category is reached - that is, when enough of the sequence has been encountered that it can be classified unambiguously (Norris 1990 uses a model of this kind to capture cohort effects in word recognition).

To see how well networks can make such classifications with time-warped stimuli, we compared this basic architecture (Network A) with two minor variants. These networks contain additional output windows of different lengths at the output layer (Fig 2). In one (Net B), this output window contain nodes representing inputs at the past and future two time slices and the current input (cf. Maskara & Noetzel 1992; Shillcock, Lindsey, Levy & Chater 1992). For the other variant of the network (Net C), the nodes in the extra output window represent input at t+2 time step only. In contrast to the target output which remains constant, these additional outputs change with the presentation of each input. The idea is to force the network to pay attention to the individual elements being presented in succession for a specified window and/or to prepare the net to accept inputs that arrive at a specified time in the future.

3. Non-duration-based stimuli

The interesting question we asked ourselves was: how might a recurrent network solve the problem of

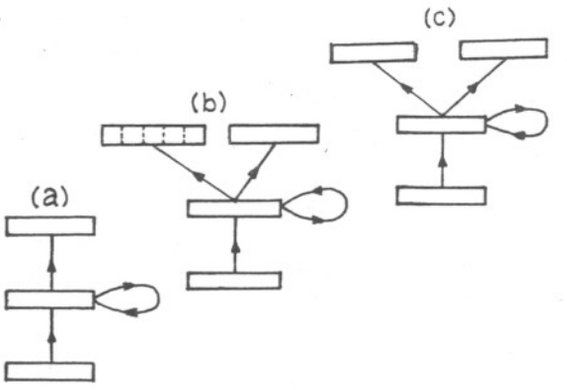

Figure 2. Folded versions of (a) Network A,
(b) Network B, (c) Network C

learning to classify a set of sequences presented at a rate it was not familiar with? In our first two series of studies, we used sequences which were unique in the sequential order of their constituent elements and whose respective identities remained unaffected by changes in the duration of these elements. Two training versions and one test version of 27 sequences were built from all possible combinations of three numbers; each version representing different rates of input. The rate at which the test sequences were presented were intermediate between the two training sets. The numbers were implemented as three-bit binary elements.

3.1. Simple variation of input

The stimuli for the first experiment were prepared following the procedure used by Norris (1990). The stimuli comprised a 'fast`, 'medium' and 'slow' series, the 'medium' series being set aside as test items. In the 'fast' series each element of a sequence remained constant for one time slice, in the 'medium' series, this is extended to two units of time each, and in the 'slow' series, each element lasted for three units of time. Table 1 illustrates the temporal composition of a pattern across the three rates.

In this and the next set of simulations, the basic network consists of an input layer of 3 input nodes, a single hidden layer of either 30 or 36 nodes, and an output layer of 27 nodes. The two variants of the network contain an additional 15 and 3 output nodes respectively. The networks were unfolded for 13 cycles during training. We ran each simulation twice with a different weight start. Batch learning was employed.

Table 1. Breakdown of rate sequences for a temporal pattern ABC

Time	t1	t2	t3	t4	t5	t6	t7	t8	t9
Fast	A	B	C						
Slow	A	A	A	B	B	B	C	C	C
Test	A	A	B	B	C	C			

Table 2. Results from the best of two weight starts.

	Net A		Net B		Net C	
no. h.u.s	30	36	30	36	30	36
no. targets correct	25	27	24	26	25	26
no. identified on time	20	21	22	22	21	21

Results Training ceased when the sum-squared error of the training set no longer decreased by at least a ten thousandth of a fraction. Total training time was usually between 1,000 to 2,000 iterations. The network learned the training set to a total sum-squared error of less than 250. In interpreting network response, we used a winner-take-all criterion. Since the uniqueness point of each training sequence can be determined a priori, the nets' performance in identifying the winning node with respect to these uniqueness points was also compared.

All the three nets correctly classified all the training stimuli (Table 2). The nets also performed surprisingly well with the test stimuli. A large proportion of these correct responses were achieved at the point the sequences became unique. Figure 3 demonstrates the recognition process of the 'medium' version of pattern 132 (appeared as 113322). Notice that the sequence of input up to the third time step is an exact copy of the 'fast' version of pattern 113. Since the net has learned to recognize this pattern during training, it responded by exciting the output node of this pattern. However, as more information became available with the presence of subsequent input, this earlier decision was revoked; activations of pattern 113 began to decrease while that of pattern 132 increase. At the fifth time-step, pattern 132 was declared the winner. Two other competing patterns are also shown in the figure. The nets' strategy was to accumulate just enough information to make the final decision; this usually coincided with the uniqueness point of the relevant sequence.

3.2. Complex variation of input

In real speech, changes in rate do not result in a simple compression and expansion of the speech signal as modelled in the previous section. Rather the time warping is quite complex. One case in point concerns

Table 3. Breakdown across rates for a temporal pattern ABC

	Time	t1	t2	t3	t4	t5	t6	t7	t8
Version Z	Fast	A	B	C					
	Medium	A	B	B	C	C			
	Slow	A	B	B	B	C	C	C	C
Version X	Fast	A	B	C					
	Medium	A	A	B	B	C			
	Slow	A	A	A	A	B	B	B	C

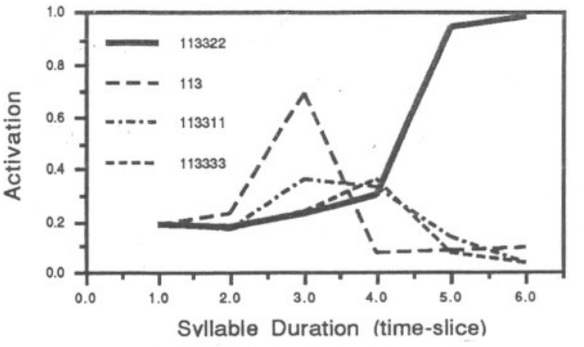

Figure 3. On-line recognition of the test sequence 113322 (obtained by network C)

the absolute and relative durations of vowels. Evidence has shown that an increase in speech rate reduces the duration of a long vowel ([i]) more than a short vowel ([I]), so that the absolute difference between the two vowels is reduced at the faster rate of speech (Miller 1981). In this section we modified the earlier stimuli to incorporate such complexity.

Instead of varying the duration of all the elements of the sequences linearly, only some were varied linearly, while others nonlinearly (Table 3). Two versions of these stimuli were constructed (versions Z and X) by switching the rate at which the durations of the first and third elements of each sequence grow in length. As in the previous experiment, the medium series in both versions served as the test set.

The motivation behind having two versions of the complex variation of sequences was the intuition that a left-to-right processing model of this kind will exact a higher cognitive cost if the transition from one element to another occurs much later in the sequence than if it occurs earlier in time, since the system must learn to pay attention to temporally more distant information. We wanted to confirm this intuition.

Results Again all the nets were successful in correctly classifying the training stimuli. As shown in Table 4, the nets also handled quite well the test stimuli,

Table 4. Results from the best of two weight starts.

Version Z	Net A		Net B		Net C	
no. h.u.s	30	36	30	36	30	36
no. targets correct	27	27	27	27	27	27
no. identified on time	27	27	27	26	27	27
Version X						
no. h.u.s	30	36	30	36	30	36
no. targets correct	22	25	21	22	26	23
no. identified on time	22	25	15	17	24	19

recognizing a large number of them at their unique points. However, as expected, version X proved more difficult for the networks as compared to version Z.

Comparatively, network B performed less better than the other two networks. A look at the activity of its target nodes suggested that the net has a much more powerful discriminant decision capability than the other networks, but this held only for the training phase; the net, in fact, wasn't as good at generalizing for this group of stimuli.

4. Duration-based stimuli

In the last two experiments, the duration of the constituent elements of a sequence made no difference to the identity of that sequence. In this section, we consider a set of sequences whose identity depends on the duration of these very elements. This occurs in real speech and is extensively discussed in the speech production and perception literature (see Miller 1981 for a review). One commonly cited example involved the voicing distinction between /bi/ and /pi/ as specified by the voice onset time (VOT). These syllables can be differentiated simply by the duration of this property: the VOT of /b/ being typically shorter than that of /p/. More importantly, however, as speaking rate changes from fast to slow and the individual words become longer, the criterion VOT value that distinguishes /b/ and /p/ also move toward longer values (Miller, Green & Reeves 1986). Interestingly, the perceptual system seems to adjust accordingly, as though taking into account the change in rate and treating VOT in a rate-dependent manner when categorising voiced and voiceless stop consonants (Summerfield 1981). Our goal is to work towards a first approximation of this rate normalization process.

In general, the magnitude of the boundary shift obtained for production data was greater than that typically found in perceptual experiments (Miller, et al. 1986). Although we remain agnostic about which data our stimuli were modelled after, for consistency, they can be taken as modelling production data.

The stimuli were loosely patterned after the synthesized syllables used by Volaitis & Miller (1992). We represent the /bi/ and /pi/ syllables as follows:

/bi/ --> 21133333333333444444
/pi/ --> 21111111111333444444

where the number correspond to various acoustic properties, in this case, 2 refers to the release burst, 1 silence, 3 transition, and 4 steady-state. Each property is repeated depending on how long we want to represent the duration of that property in the syllable[1]. The representation for /pi/ was derived simply by lengthening the VOT (counted from the onset of burst till the offset of silence) of the /bi/ syllable. This meant replacing the transition portions with a lengthening silence.

The properties were implemented as 4-bit patterns. The basic network consisted of 4 input units, 5 or 10 hidden units, and 2 output units. The two variant networks have additional output windows of 20 and 4 nodes respectively. In this and the next set of simulations, the nets were unfolded for 36 time cycles during training.

4.1. Non-overlapping stimuli

From the speech production data of Miller et al. (1986), it appears that within place of articulation, there is some overlap in the distribution of VOT values for voiced and voiceless consonants across different speech rates[2]. In this section, however, we assume no overlap of VOT values across rates. This should be a straightforward task from the processing point of view: a property that lasts for a certain time range specifies one segment, and another if it extends beyond that range. Several /bi/-/pi/ pairs were constructed across six rates[3]. Figure 4 shows how the VOT values for the syllables vary across these rates. One of the (vertical) pairs was set aside as test material.

Results All the nets were able to handle both the training and test stimuli. The fact that the net was able to make appropriate generalisations was interesting. As expected, the mechanism employed by the nets

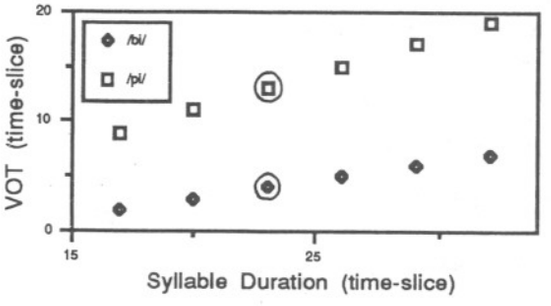

Figure 4. VOT against syllable duration (Items circled used as tests)

[1] In synthetic stimuli, the standard way of incrementing VOT is by replacing periodic portions of the transitions with aspirated portions instead of silence but we didn't consider this minor variation as crucial for the simulations.

[2] Overlap here is defined as the range of VOTs between that value obtained in a voiceless token produced at the fastest speech rate and that value obtained in a voiced token produced at the slowest speech rate. This differs from the common use of the term which defines overlap locally in relation to a particular speech rate specified by the syllable duration.

[3] In natural speech, there is no one-to-one mapping between speech rate and VOT for both voicing categories, as assumed in this paper. Instead, for any given speech rate, listeners often encounter tokens of both categories produced with not only just one VOT each but a range of VOTs.

operates in terms of a criterion VOT boundary range, such that stimuli with VOTs lower than the criterion are classified as /b/ and those with higher VOTs are classified as /p/. Syllable duration is therefore irrelevant in cueing phonetic distinction and has no influence over the perceptual task for this group of non-overlapping stimuli. The offset of the VOT was the critical point in the syllable that triggered the contrast.

4.2. Overlapping stimuli

In this section, we asked how the nets might perform in the face of stimuli whose VOT values overlap over a certain range, as are in real speech. Figure 5 shows the distribution of the /bi/-/pi/ syllables across VOTs and syllable durations. Three of the /bi/-/pi/ (horizontal) pairs are within the overlap range; they share VOT values but differ in syllable duration, as illustrated below. Sequence A is a /pi/ syllable presented at a faster speech rate (as specified by a shorter syllable duration) than sequence B, a /bi/ syllable, but their VOT values are identical. To recover the intended phonetic segment specified by the VOT value, one has to consider the entire syllable.

A /pi/ 21113333344444
B /bi/ 211133333444444444444

Two (vertical) pairs were set aside as test material. Of these, one (horizontal) pair (with identical VOTs) assumes the form A and B above.

Results All the networks were successful in learning to classify the training stimuli including those within the overlap range. However, only networks B and C were able to generalize appropriately all the test stimuli. The syllables in the overlap range proved difficult for network A. It classified both /bi/ and /pi/ as /bi/. Nevertheless, the fact that the other nets can make appropriate generalisations with this kind of stimuli was encouraging.

The on-line processing by the network revealed that the identification of voiced and voiceless tokens lying outside the overlap region was straightforward, with performance reaching optimal point at the offset of the VOT (Fig 6a). The processing of the tokens in the overlap region, however, proceeded in two stages. In the first stage, the VOT was calculated. Given that in the stimulus set the VOT values for the voiceless tokens are located at the higher end of the continuum, the net showed a first preference for /pi/ by gradually increasing the activation of /pi/ through the entire length of the VOT. Upon reaching the end of the VOT, however, the activation for the voiced and voiceless tokens switched direction, triggered by the possibility that a short VOT might indicate the presence of a /bi/. In the second stage, syllable duration was considered. Given that for each VOT value in the overlap case the overall syllable duration

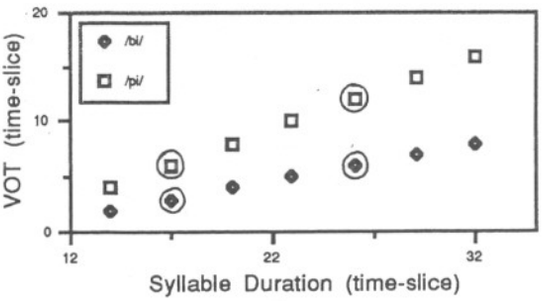

Figure 5. VOT against syllable duration (Items circled used as tests)

of the voiced tokens is always greater than the voiceless tokens, the net assumed from the start of transition that it perceived a /bi/. It thus reinforced the activations of /bi/ from that point on until it was classified unambiguously. For the voiceless tokens, their activations began to pick up only at the offset of the syllable; that is, syllable duration was, in this case, critical for the identification of these tokens (Fig 6b).

5. Discussion

The basic computer science of the recurrent networks emerged in an interesting way. Without additional output windows, the networks work wonderfully well in accomodating shorter non-duration-based time-warped sequences as well as longer duration-based sequences whose constituent elements do not overlap in time. With additional windows, the networks perform better with duration-based stimuli but otherwise with non-duration-based ones. However, if the output window is limited to serve only one input at a time, in this case given a look-ahead of 2 time-steps, the recurrent network can be made to accomodate all types of time-warped sequences.

In terms of speech processing, the present finding is significant in that it offers a plausible account for the correspondence between the way in which a contextual variable alters VOT values and the way in which the variable seeks relevance in the restructuring of the

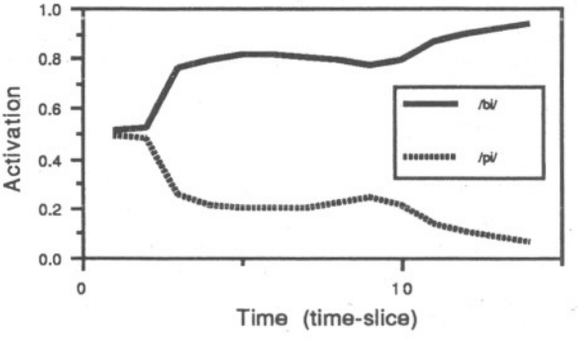

Figure 6a. On-line recognition of a fast /bi/ syllable (duration 14), a token outside the overlap range.

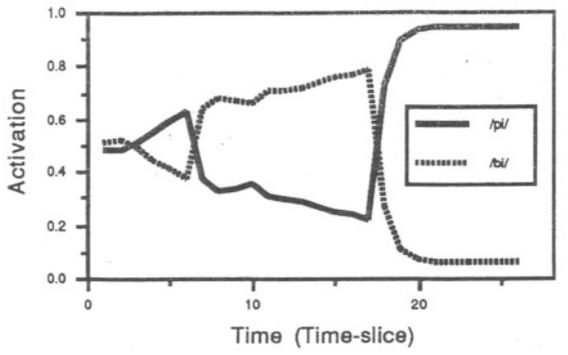

Figure 6b. On-line recognition of the test syllable /pi/ (duration 17), a token from the overlap range. Its activation shoots up at the offset of the syllable, at which point it differs from the /bi/ syllable of identical VOT but longer syllable duration (26 time-slice)

phonetic category in perception. We have shown a mechanism whose strategy is to pick up information early in the syllable, and use only those that are relevant to the contrast being judged, which means they don't always have to consider all information within the syllabic unit. Spcifically, where no overlap is present, and the range of VOTs is distinct between /bi/ and /pi/ across different speech rates, syllable duration is an unnecessary aid to phonetic distinction. But where there is overlap in the VOT distribution as one would find in real speech, the mechanism discriminates between stimuli on the basis of whether they are within or outside the overlap region of the VOT continuum; the use of rate information as provided by the syllable duration is obligatory only when processing tokens from the overlap range (cf. Miller, 1987). This raises some questions about the nature of the human speech processing system. Firstly, in the face of changing speech rates, is the system sensitive to the structural distribution of temporal properties such as the VOT that provide cues to phonetic contrasts? In particular, does the system treat differently tokens that belong to the overlap region and those that do not? Secondly, assuming that the system can make a voicing decision partway through the syllable, is the initial decision made on the VOT and then changed due to later rate information, or is the decision postponed until the rate information is available? Apparently, these are questions beyond the scope of this paper.

6. Conclusions

We have described a useful way of modelling some rate-dependent factors in speech perception. This was possible through the use of recurrent neural networks whose behaviour brought out the aspects relevant to understanding time-warping problems. Although we cannot determine precisely how the representational compromises that have been made contributed to this behaviour, we were nevertheless encouraged by the results we obtained. Future work will necessarily have to address this issue if more factors linked to effects of speaking rate are to be accounted for.

References

Chater, N., & Conkey, P. (1992). Finding linguistic structure with recurrent neural networks. *Proceedings of the 14th Annual Meeting of the Cognitive Science Society,* Hillsdale, NJ: LEA.

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

Huang, X.D., Ariki, Y., & Jack, M. A. (1990). *Hidden Markov Models for Speech Recognition.* Edinburgh: Edinburgh University Press.

Lisker, L. & Abramson, A. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, **20**, 384-422.

Maskara, A. & Noetzel, A. (1992). Forcing simple recurrent neural networks to encode context. *Proceedings of the 1992 Long Island Conference on Artificial Intelligence and Computer Graphics.*

Miller, J. L. (1987). Rate-dependent processing in speech perception. In Ellis (Ed) *Progress in the Psychology of Language*, Vol. 3. London: Erlbaum.

Miller, J. L. (1981). Effects of speaking rate on segmental distinctions. In Eimas & Miller (Eds) *Perspectives on the study of speech.* Hillsdale, NJ: LEA.

Miller, J. L., Grosjean, F. & Lomanto, C. (1984). Articulation rate and its variability in spontaneous speech: A reanalysis and some implications. *Phonetica*, **41**, 215-255.

Miller, J. L., Green, K. P. & Reeves, A. (1986). Speaking rate and segments: A look at the relation between speech production and perception for the voicing contrast. *Phonetica*, **43**, 106-115.

Norris, D. (1990). A dynamic-net model of human speech recognition. In G. Altmann (Ed) *Cognitive Models of Speech Processing: Psycholinguistic and Cognitive Perspectives.* Cambridge: MIT Press.

Rumelhart, D. E., Hinton, G. E. & Williams, R. J. (1986). Learning internal representations by error propagation. In Rumelhart & McClelland (Eds) *Parallel Distributed Processing: Explorations in the Micro-structures of Cognition*, Vol. 1. Cambridge, Mass: MIT Press.

Sakoe, H. & Chiba, S. (1971). A dynamic programming approach to continuous speech recognition. *Proceedings of the International Congress on Acoustics.* Budapest, Hungary, Paper 20 C-13,.

Shillcock, R., Lindsey, G., Levy, J. & Chater, N. (1992). A phonologically motivated input representation for the modelling of auditory word perception in continuous speech. *Proceedings of the 13th Annual Meeting of the Cognitive Science Society,* Hillsdale, NJ: LEA.

Summerfield, A. Q. (1981). Articulatory rate and perceptual constancy in phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance*, **7**, 1074-1095.

van Camp, D., Plate, T. & Hinton, G. (1992). Xerion Neural Network Simulator. Dept. of Computer Science, University of Toronto.

Volaitis, L. E. & Miller, J. L. (1992). Phonetic prototypes: Influence of place of articulation and speaking rate on the internal structure of voicing categories. *Journal of the Acoustical Society of America*, **92**, 723-735.