Connectionist Models of Memory and Language

Edited by

Joseph P. Levy, Dimitrios Bairaktaris,
John A. Bullinaria, Paul Cairns

CHAPTER 14

Time-warping tasks and recurrent neural networks

Mukhlis Abu-Bakar & Nick Chater

Introduction

Finding structure in real-world acoustic signals, whether from the perspective of engineering or psychology, is difficult because not only must an underlying sequence of elements be discerned, but, frequently, the duration of those elements may be "warped" in complex ways. Furthermore, in some contexts, such as speech recognition, this warping is especially problematic, since the very identity of sequence elements may be given by duration-based cues, which warping will distort. The ability to cope flexibly and successfully with time-warped material is crucial if neural networks, or any other computational technique, are to be applicable to a wide range of real-world temporal processing tasks.

Utterances are frequently time warped because speakers 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 quite substantial (Miller et al. 1984). More often, time warping distorts the temporal structure of words: a segment of a word may be compressed, another stretched, while others remain durationally invariant to changes in the speech rate. The problem is more complex at the phonemic 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). Thus, time warping potentially confounds temporal cues to phoneme identity.

Within engineering, there have been various approaches to solving time-warping problems (e.g. hidden Markov models (Huang et al. 1990), dynamic time warping (Lipmann et al. 1987) and dynamic rate adaptation (Nguyen & Cottrell 1993)). The present work attempts to apply connectionist tools to the problem of time warping. To the extent that connectionist methods are psycho-

logically plausible, this work also gives an attractive approach to modelling aspects of human speech perception.

Using recurrent back-propagation

Recurrent neural networks have been widely used in modelling sequence processing (e.g. Elman 1990), including a wide range of problems drawn from speech processing (e.g. Norris 1990, Shillcock et al. 1991, Watrous et al. 1990). Recurrent networks are attractive for such problems since their behaviour depends on the entire sequence of inputs, rather than just the current input, although there are various ways in which feedforward networks can be modified in order to handle sequential material (see Chater (1989) for a review).

We use a standard recurrent neural network architecture, in which the units in the hidden layer are connected to all other hidden units by weights which operate with a delay of one time step. This kind of recurrent network is often thought of as involving an additional set of units, the "context" units, to which the hidden units at the previous time step are copied. According to this conception, the context units are treated simply as additional input units to the network. This kind of recurrent network can be trained in a variety of different ways, the most common being Elman's (1990) "copy-back" scheme, which uses a computationally cheap approximation to gradient descent in error to change the weights. We use recurrent back-propagation (Rumelhart et al. 1986), which computes gradient descent more exactly by "unfolding" the recurrent network into a sequence of serially connected feedforward networks, and then trains the resulting network using standard back-propagation. The only additional constraint on learning is that the weights in each "incarnation" of the recurrent network in the unfolded feedforward network are constrained to be the same, so that it is possible to fold the trained feedforward network back up into a recurrent network.

In general, the larger the number of unfoldings used, the more exactly the network computes true gradient descent, although the benefits of additional unfoldings begin to tail off after some point, because very deep feedforward networks are very slow to train. It is also important to note that the number of unfoldings used in training does not place a strict limit on the distance back in the sequence to which the network can learn to be responsive. Even if the network is trained as a feedforward network unfolded through n time steps, the "context" units in the final unfolding are likely to contain information about the inputs at earlier time steps, and the network may therefore learn to become sensitive to this information. Nonetheless, although under certain circumstances networks can learn to respond to information which is very much more temporally distant than the number of unfolded time steps, in practice, performance is generally rather poor for such distant items (see Chater & Conkey (1994) for discussion). Training used conjugate gradient descent, and was implemented on the Xerion simulator (van Camp & Plate 1993).

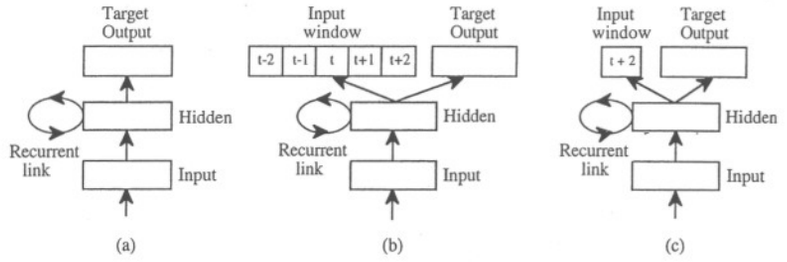

Figure 14.1 Folded versions of (a) a standard recurrent network (network A), (b) a network with five additional input windows at the output layer (network B), and (c) a network with a single additional input window at the output layer (network C).

Comparing architectures

Perhaps the main advantage of using recurrent networks, as opposed to other neural network methods, is their ability to treat temporal material, such as speech, as a sequence of events and take input one at a time. We trained the networks by feeding them with the sequences one input pattern at a time and keeping the target output pattern present throughout the presentation of each sequence. The production of the correct output as the sequence is presented indicates that the sequence is classified successfully. If performance is optimal, correct classification should occur after the "recognition point" of the category is reached, i.e. 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 the basic recurrent network architecture (network A) with two minor variants (networks B and C). These latter networks contained additional input windows of different sizes at the output layer (Fig. 14.1). In one (network B), this window contained nodes representing inputs at the previous two and next two time slices and the current input (cf. Maskara & Noetzel 1992, Shillcock et al. 1992). For the other variant of the network (network C), the nodes in the window represented input at the $t+2$ time step only. In contrast to the target output, which remained constant, these additional outputs changed with the presentation of each input. The idea was 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. The following set of experiments was designed to test how well these various networks classify sequences presented at a rate they are not familiar with.

Non-duration-based stimuli

For the experiments in this section, we used sequences which are unique in the sequential order of their constituent elements and whose respective identities remain unaffected by changes in the duration of these elements. Two training versions and one test version of 27 sequence types were built from all possible combinations of the numbers 1, 2 and 3, with each version representing different rates of input (fast, medium, slow). The set of stimuli presented at the intermediate rate was the test set. The three numbers were implemented as three-bit binary elements. Thus, 100 stood for 1, 010 for 2 and 001 for 3.

The basic network consisted of an input layer of three 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 contained an additional 15 and three output nodes, respectively. The networks were unfolded for 13 cycles during training. We ran every simulation twice with a different weight start for each attempt. Batch learning was employed.

Experiment 1: simple variation of input

We varied the duration of the constituent elements of the sequence types following the scheme used by Norris (1990). Table 14.1 illustrates the temporal composition of a model sequence type across the three rates. The period over which each constituent member of a sequence type appears (or the number of time it is successively presented) is captured by the relation $N_i = i$, where i is the rate, and N_i the number of successive presentations of each constituent member at the specified rate. Thus, in the "fast" series, each member of a sequence type remains constant for one time slice. In the "medium" series, this is extended to two units of time each, and in the "slow" series, each constituent member lasts for three units of time.

For all the 54 training stimuli, it is possible to determine by hand at which point in a stimulus the sequence type is recognizable. With a few exceptions, this does not normally require that the entire stimulus be processed. It would be interesting if the nets can capture this sequence structure by identifying the sequence type at the point in time when the stimulus item becomes unique. The crucial test, however, is how to generalize from this sequence structure to new stimuli presented at an intermediate rate.

All stimuli were preceded by an input pattern in which all three bits were set to 0. In the long version of the stimuli, where each constituent member of a

Table 14.1 A model sequence type at the three presentation rates (experiment 1).

Rate	t1	t2	t3	t4	t5	t6	t7	t8	t9
1: Fast	A	B	C						
2: Medum	A	A	B	B	C	C	C		
3: Slow	A	A	A	B	B	B	C	C	C

sequence type was repeated three times, they were followed by a final 0 input pattern. In the short version where each constituent member of a sequence type appeared only once, they were followed by seven input patterns which were set to 0. In the test stimulus where the constituent members appeared twice, they were followed by four of these 0 input patterns. As an illustration, sequence type 321 would be presented to the network in the slow and fast modes as follows:

Slow mode			Fast mode			
0	0	0	0	0	0	at time t1
0	0	1	0	0	1	at time t2
0	0	1	0	1	0	at time t3
0	0	1	1	0	0	at time t4
0	1	0	0	0	0	at time t5
0	1	0	0	0	0	at time t6
0	1	0	0	0	0	at time t7
1	0	0	0	0	0	at time t8
1	0	0	0	0	0	at time t9
1	0	0	0	0	0	at time t10
0	0	0	0	0	0	at time t11

Results

Training ceased when the sum-squared error of the training decreased by less than 0.0001. Total training time was usually between 1000 and 2000 iterations. To assess network response, we looked for the node with the highest activation at the point the sequence type can be identified. The activation of this node must be higher than that of other competing nodes by a criterion value of at least 0.25; otherwise there will be no winner, and the search for the winning node moves to the next time step. A stimulus is accepted as being correctly classified if the winning node corresponds to the target node and that its distance from the other nodes is maintained right up to the end of the stimulus and one time step after.

All three nets successfully fulfilled these criteria for all the training stimuli, whatever the hidden unit population. With the test stimuli, however, the nets

Table 14.2 Successful identification of test stimuli for each network type and hidden unit population (experiment 1).

Network type	Hidden units	Targets correct	On time
A	30	25	15
	36	27	21
B	30	23	15
	36	24	15
C	30	25	16
	36	26	18

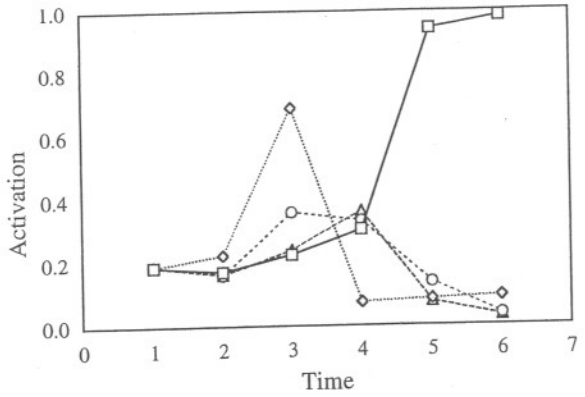

Figure 14.2 Activity of test stimulus 113322 (solid line) with 132 as the underlying sequence type. Activations of three other competing sequence types are also shown (experiment 1).

achieved a slightly lower rate of success (Table 14.2). Consider the results from simulations with 36 hidden units since these show better performance. Only network A maintained 100% correct response for the test stimuli, whereas networks B and C had three and one wrong classifications, respectively. In terms of the number of timely responses, network A managed a high 21, while networks B and C only 15 and 18, respectively. Thus, network A performed more consistently with both sets of stimuli than networks B and C.

Figure 14.2 demonstrates the recognition process of the test (medium) version of sequence type 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 sequence type 113. Since the net has seen this stimulus item during training, it responds appropriately by exciting the output node of this sequence type at its unique point (i.e. at the third time step). However, when another 3 comes along at the following time step, the net is forced to revoke its decision, for this creates a novel string 1133. This change is depicted by the downward shift in the activation value of the sequence type 113. Activations of other sequence types continue to remain low. Interestingly, when 2 is presented at the fifth time step, the net promptly and correctly activates sequence type 132 without waiting for the last input pattern to arrive. This is an optimal performance where the strategy is to accumulate just enough information to make the final decision.

Experiment 2: complex variation of input

For many complex temporal stimuli, including speech, changes in rate do not result in a simple compression and expansion of the signal as modelled in the previous section. Rather, the time warping is quite complex. One case in point concerns the absolute and relative durations of vowels in conversational speech. An increase in speech rate has been shown to reduce the duration of a long vowel (e.g. [i]) more than a short vowel (e.g. [I]), so that the durational difference between the two vowels is reduced at the faster rate of speech (Miller 1981). In this experiment, we modified the earlier stimuli to reflect such complexity.

Instead of varying the duration of all the constituent members of a sequence type equally via a single linear function, three different functions were used. Two versions of these stimuli were constructed. In the first version, X, the duration of the first constituent member remained constant across rates following the relation $N_i = 1$. For the second member, the duration was specified by the same linear function used previously, namely, $N_i = i$; and for the last member, a nonlinear function $N_i = 2^{(i-1)}$ was used. For all the functions, i is the rate and N_i the number of successive presentations of the constituent member at the specified rate. The resulting temporal configuration can be found in Table 14.3. A second version, Y, was created from the first version by switching the functions associated with the first and third constituent members. The motivation behind having two versions of the complex variation of sequence types 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 attend to temporally more distant information. We wanted to confirm this intuition. The nets were trained with the two versions separately. As in the previous experiment, the medium series in both versions served as the test set.

Table 14.3 Two versions (X and Y) of a model sequence type at the three presentation rates (experiment 2).

Version	Rate	t1	t2	t3	t4	t5	t6	t7	t8
X	1: fast	A	B	C					
	2: medium	A	B	B	C	C			
	3: slow	A	B	B	B	C	C	C	C
Y	1: fast	A	B	C					
	2: medium	A	A	B	B	C			
	3: slow	A	A	A	A	B	B	B	C

Results

All the nets classified both versions of the training stimuli correctly, irrespective of hidden unit population. Generalization to the test stimuli, however, was not uniform across stimulus versions and hidden unit population (Table 14.4). With the exception of network C, simulations with 36 hidden units for the other two

Table 14.4 Successful identification of each version of test stimuli for each network type and hidden unit population (experiment 2).

Network type	Hidden units	Targets correct	On time
Version X			
A	30	27	27
	36	27	27
B	30	27	27
	36	27	26
C	30	27	27
	36	27	27
Version Y			
A	30	22	22
	36	25	23
B	30	20	9
	36	22	13
C	30	26	21
	36	22	18

nets produced better results. And, as expected, version Y, as opposed to version X, proved more difficult for all three nets. Simulations with version X produced perfect scores on correct classification, and almost perfect scores on getting the classifications correct on time, but with version Y the recognition rate varied between 20 and 26 while that of timely responses between 9 and 23. Comparatively all round, network B performed less well than the other two networks.

Duration-based stimuli

In the preceding experiments, the duration of the constituent members of a sequence type made no difference to the identity of that sequence type that we altered by time warping. In this section, we consider a set of sequences whose identity depends on the duration of these very elements. This occurs in a variety of ways in natural speech, and is extensively discussed in the speech production and perception literature (see Miller (1981) for a review). One commonly cited example involves 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: /b/ having typically shorter VOTs than /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 moves towards longer values (Miller et al. 1986). Potentially, due to this variation, the mapping from acoustic signal to phonetic percept is difficult, but, interestingly, listeners adjust for these variations with apparent ease. Our goal is to work towards a first approximation of this "rate normalization" process.

Although primarily intended as an abstract test, our stylized stimuli were loosely patterned after the synthesized syllables used by Volaitis & Miller (1992). Two contrasting stimuli, /bi/ and /pi/, took the form

/bi/ → 21113333333333444444
/pi/ → 21111111111333444444

where the numbers represent the states of various acoustic properties; in this case, 2 may be taken to refer to the release burst, 1 to F1 cutback, 3 to transition, and 4 to steady state. The duration of a particular property is specified by the number of times the corresponding state is repeated. /pi/ is derived from /bi/ simply by lengthening the latter's VOT (counted from the onset of the burst till the onset of transition). This involves extending the F1 cutback by cutting into the transitions. Localist representations of 4-bit patterns were used for the states. The basic network consisted of four input units, five or ten hidden units, and two output units. The two variant networks, B and C, contained an additional 20 and four output nodes, respectively. In this and the next set of simulations, the nets were unfolded for 36 time cycles during training.

Experiment 3: non-overlapping stimuli

From the production data of Miller et al. (1986), it appears that within a place of articulation, there is some overlap in the distribution of VOT values for voiced and voiceless consonants across different speech rates. However, in this section, we assume no overlap of VOT values between categories across rates. Thus, recognition should be a straightforward task from the processing point of view: a VOT that lasts for a certain time range specifies one segment, and another if it extends beyond that range. Six /bi/–/pi/ pairs were constructed across six rates, as shown in Figure 14.3. One pair (23 time steps syllable duration) was set aside as test material.

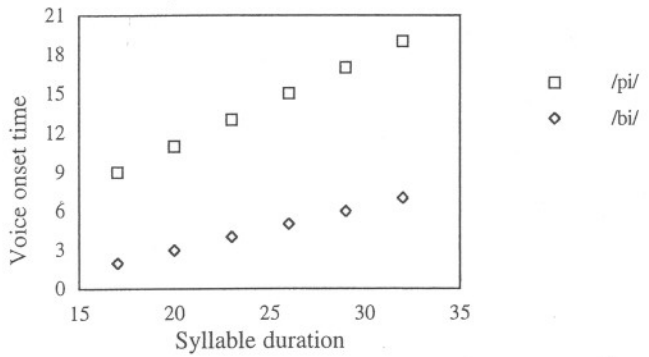

Figure 14.3 Distribution of /bi/ and /pi/ tokens for the non-overlapping case (experiment 3).

Results

All three nets were able to handle effectively both the training and test stimuli. As expected, the strategy employed by the nets operates in a straightforward fashion: information about VOT alone triggers the contrast between /b/ and /p/ for this group of non-overlapping stimuli. Syllable duration is thus irrelevant in the distinction and exerts no influence over the processing task.

Experiment 4: overlapping stimuli

We now consider the more realistic case in which VOT values overlap over a certain range, as in natural speech. Figure 14.4 shows the relationship between VOT and syllable duration for 14 /bi/ and /pi/ stimuli. Six of these stimuli are within the overlap range (/pi/-1 and /bi/-3, /pi/-2 and /bi/-5, and /pi/-3 and /bi/-7, where /pi/-*n* denotes /pi/ at rate *n*, with rate 1 being the fastest rate and 7 the slowest). Every /bi/–/pi/ pair in this range has an identical VOT but different syllable duration, as illustrated below. Sequence U is a /bi/ syllable presented at a slower speech rate (as specified by a longer syllable duration) than sequence V, a /pi/ syllable, but their VOT values are the same. To recover the intended voicing feature specified by the VOT value, the nets have to consider the entire stimulus.

U /bi/ 21113333344444444444
V /pi/ 21113333344444

Two /bi/–/pi/ pairs (rates 2 and 5) were set aside as test material. Of these, /bi/-5 and /pi/-2 have identical VOT values but different syllable duration.

Results

All three 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 to all the test stimuli appropriately. The stimuli in the overlap

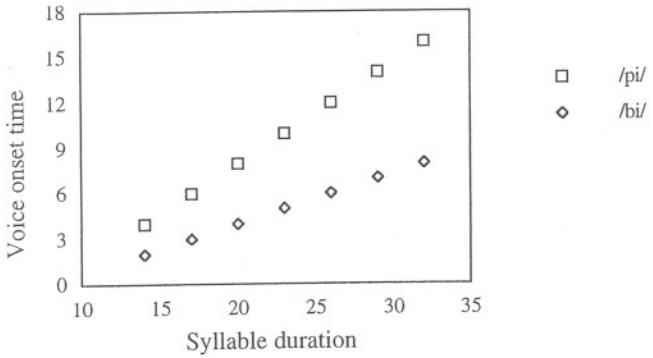

Figure 14.4 Distribution of /bi/ and /pi/ tokens for the overlapping case (experiment 4).

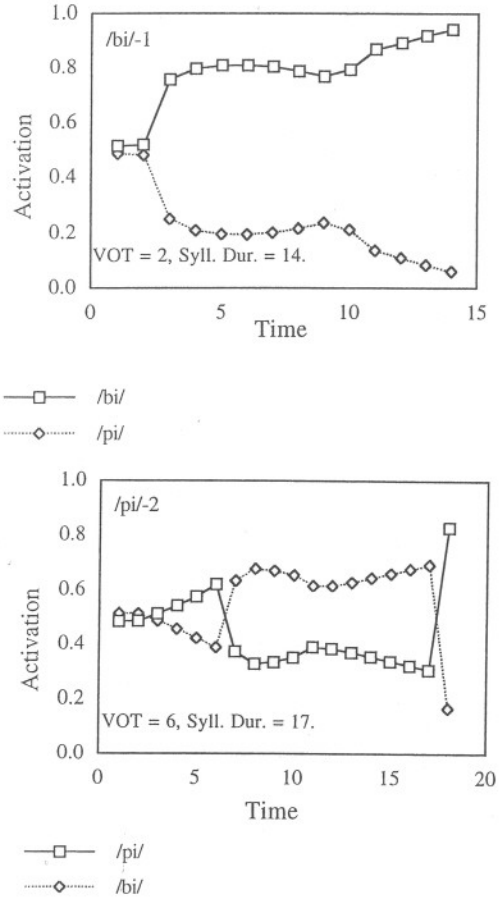

Figure 14.5 Activity of a fast /bi/ syllable (/bi/-1), a token outside the overlap range (experiment 4).

Figure 14.6 Activity of a test /pi/ syllable (/pi/-2), a token from the overlap range (experiment 4).

range proved difficult for network A: it classified both /bi/ and /pi/ as /bi/. Nevertheless, the fact that the other nets can make appropriate generalizations with this kind of stimuli was encouraging.

The networks' on-line processing reveals that the process of identifying voiced and voiceless tokens that lie outside the overlap region is straightforward, as it was with the previous set of stimuli, with performance nearly optimal at the offset of the VOT (Figure 14.5). The processing of the tokens in the overlap region, however, proceeded differently, and in two stages. Figure 14.6 shows the on-line processing of a voiceless token from the overlap range (/pi/-2). In the first stage, VOT is calculated. The nets show an initial 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 activations for the voice and

voiceless tokens switch direction, triggered by the possibility that the given VOT duration is uncharacteristically short for a /pi/ stimulus thus favouring /bi/ instead. This is the second stage where vowel duration is considered. At the end of the vowel, the activations again switch direction, this time cued by the possibility that the given vowel duration against the earlier information about VOT can only match a /pi/ rather than a /bi/ stimulus. Thus, syllable duration, in this case, is critical for the identification of the voiceless tokens.

Discussion

What these experiments reveal about the basic computer science of recurrent networks is encouraging. The type of problem dealt with here is tractable using relatively simple networks. Without additional input windows at the output layer, the networks work well in accommodating the shorter non-duration-based time-warped sequences as well as the longer duration-based sequences whose constituent elements do not overlap in time. With additional windows of input units trained to remember and predict inputs, the networks handle well duration-based stimuli (overlap and non-overlap) but not non-duration-based stimuli. However, with a single input window at $t + 2$, the recurrent network can accommodate the full range of test cases we have considered with an appreciable degree of accuracy. We might speculate that having to predict the future forces the network to encode the structure of the input material more carefully, and that this, as a side-effect, results in a representation which is useful for the classification task. However, having too many time steps to predict and remember might have forced the network to devote too much attention to these tasks, and therefore to neglect the classification task.

In all four experiments the nets needed only to interpolate from training data in order to perform well on the test data. It is therefore not clear if the nets can cope with input which is more extreme than that in their training sets. The ability to *extrapolate* from training data is seemingly crucial with respect to real speech data, for there is no guarantee that the training set will contain stimuli covering the wide range of speaking rates. A follow-up experiment was thus carried out with network C, and the findings suggest that extrapolation is an achievable task for this network. In the experiment, we used stimuli from experiment 4, but instead of training the net with stimulus pairs at rates 1, 3, 4, 6 and 7, and testing it on pairs at rates 2 and 5, we trained the net on stimulus pairs at rates 2, 3, 4, 5 and 6 and tested it on pairs at the extreme rates (1 and 7). A particularly interesting case is the test item /pi/-1 which has the same VOT value as training item /bi/-3. Despite being trained on the longer /bi/-3 stimulus, the net successfully classified /pi/-1 on the basis of its shorter overall duration. Thus, the net was able to extrapolate to data at rates not in the training set.

The strategies arrived at by the nets are interesting as psychological models, even if it can be argued that recurrent networks, trained by algorithms like back-propagation through time, are not particularly psychologically plausible. The

present findings are significant in that they offer a plausible account of the correspondence between the way in which a contextual variable alters VOT values and the way in which the variable is used to restructure phonetic categories in perception. We have shown a mechanism whose strategy is to pick up early in the stimuli any information that is relevant to the contrast being judged, and to alter any initial decision if later information proves important for the distinction. Specifically, where no overlap is present, and the range of VOT is distinct between /bi/ and /pi/ across different speech rates, syllable duration is an unnecessary aid to phoneme 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; syllable duration is critical only when processing tokens from the overlap range. This raises some questions about the nature of the human speech-processing system. First, in the face of changing speech rates, is the system sensitive to the structural distribution of temporal properties such as 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 once the syllable has been processed, or is the decision postponed until processing of the entire syllable is completed? These questions require empirical study which is beyond the scope of this chapter.

Extension and application to speech perception

In this section, we describe simulations that apply the network's strategy (using network C) on another set of contrast, namely /b/ and /w/. These consonants are distinguished by the abruptness of their onsets or changes in transition duration (hereafter referred to as TD). In the studies of the production of these consonants, the onset for /b/, as in the syllable /ba/, was reported to be more abrupt than that for /w/ (Dalston 1975). Perceptually, the standard contrast effect has also been reported for these phonetic categories: as syllable duration increases, the /b/–/w/ boundary moves towards transitions of longer duration (Miller & Liberman 1979). This boundary, however, shifts in the opposite direction when the increase in syllable duration is effected by adding a final transition corresponding to a third phonetic segment, as in /bad/. The simulations here demonstrate how the network responds to the combined effect of syllable duration and syllable structure.

Stimuli

Eleven pairs of "speech-like" stimuli that resemble /ba/ and /wa/ syllables co-varying in syllable duration and TD values were constructed. For every pair

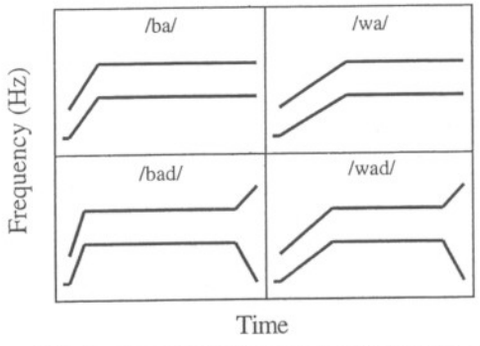

Figure 14.7 Schematic two-formant representations of /ba/, /bad/, /wa/ and /wad/ at a single rate. Each consists of a brief prevoicing (first formant only), a variable duration of formant transition appropriate for /b/ or /w/, a subsequent period of steady state formants, and, for the CVC stimuli, a final period of transition corresponding to /d/. Notice the variations in TD across category and syllable structure.

of /ba/ and /wa/ syllables of a certain syllable duration, we created a corresponding pair of /bad/ and /wad/ syllables of similar overall duration. The timing of the phonetic segments was constrained such that, for any given syllable duration, the TD value for syllables with a CVC (consonant–vowel–consonant) structure was always shorter than for those with a CV (consonant–vowel) structure, as shown in Figure 14.7. This production pattern was directly derived from the perceptual findings of Miller & Liberman (1979) with respect to the /b/–/w/ distinction when the rate and syllable structure were altered (cf. Summerfield 1981, Volaitis & Miller 1991). The "formant frequencies" which we modelled the stimuli from can be found in Abu-Bakar & Chater (1993a).

The relationship between TD and syllable duration for all stimuli is shown in Figure 14.8. Notice that CVC syllables are located along individual distributional curves separate from syllables of the CV type. However, curves which hold syllables with the same syllable–initial consonants are pulled closer together. Twelve tokens of varying syllable duration and structure were reserved as test items. Of these, /wa/-2, /wad/-4, /ba/-7 and /bad/-9 have identical TD values but different syllable duration.

In the simulations reported here, 58 unfoldings were used. The input layer which consisted of 31 units can be thought of as falling into two groups. One group represents the frequency of the first formant, and the other represents the frequency of the second formant. We use a simple localist-style coding to represent this frequency information. Each unit in a particular bank represents a particular frequency, and if a formant has frequency F, then all and only the units which represent frequency values F and less will be active. Sixty hidden units were used, which seems to be approximately the smallest number of units that can learn the task successfully. The target output window of the net has six units; one each for /ba/, /wa/, /bad/ and /wad/, and another for /b/ and /w/. The last two can be conceived of as phoneme detectors. They were included to allow for some independence between the identification of the syllable–initial phonetic segments and the classification of the syllables. This target output is in addition to another bank of units which is trained to continually predict the next but one pattern in the input sequence (recall Fig. 14.1c).

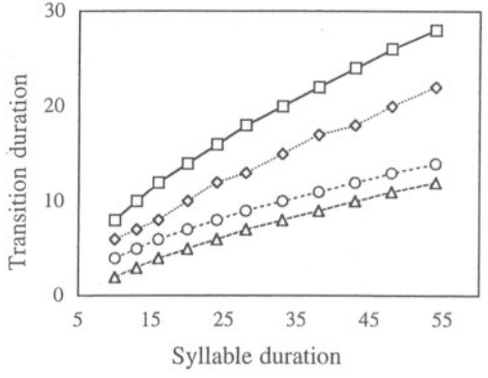

Figure 14.8 /ba/–/bad/–/wa/–/wad/ stimuli used in the final experiment.

Results

Training stopped after about 400 iterations, by which time the network had successfully classified all the training stimuli and correctly generalized to the test tokens. Figure 14.9(left) shows the on-line recognition of a /wa/-2 stimulus, one of the four test tokens with identical TD values. Since the TD is ambiguous between /b/ and /w/, the activation of the /w/ detector remains very low for much of the duration of the syllables. It shoots up only at the offset of the vowel segment, which also happens to be the end of the syllable. It is also at this point that the /wa/ syllable is distinguished from the other syllables (/wad/, /ba/ and /bad/) by way of an abrupt increase in the activation of the node representing the /wa/ syllable. In Figure 14.9(right), we have a /wad/-4 stimulus of the same TD value as /wa/-2. As with the latter, the activation of the /w/ detector is low initially and shoots up at the offset of the steady state section of the syllable. But here, the recognition point does not coincide with the end of the syllable, for there is still the final transition following the vowel. Thus, a parsimonious explanation for the syllable–initial distinction is one that takes into account the relationship between the TD and the adjacent vowel and not overall syllable duration. Indeed, we found that for all the stimuli in the overlap region, irrespective of whether they are CV or CVC syllables, the pattern is the same, i.e. the identity of the syllable–initial phoneme as well as the syllable is processed in relation to the syllable's CV component.

Figure 14.10 illustrates the processing of a group of fast /ba/ and /bad/ syllables (/ba/-2 and /bad/-2) whose TD values are outside the category overlap. As

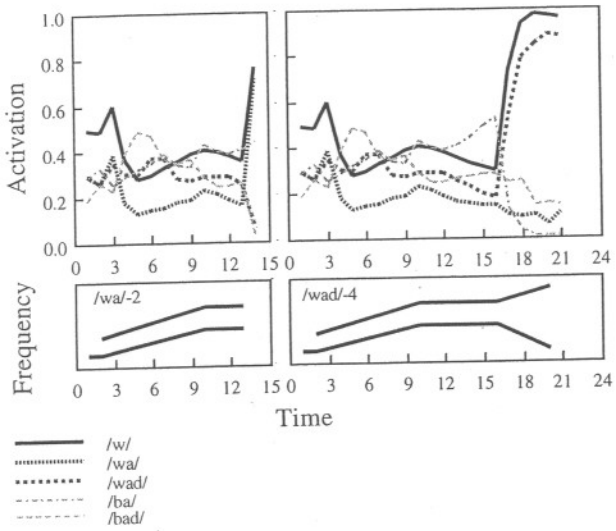

Figure 14.9 Activity of /w/ in /wa/-2 (top left) and /wad/-4 (top right), and the corresponding temporal representation of the stimuli at their respective rates (bottom panels). The activity of the /b/ detector is purposely omitted here while those of /ba/ and /bad/ are included.

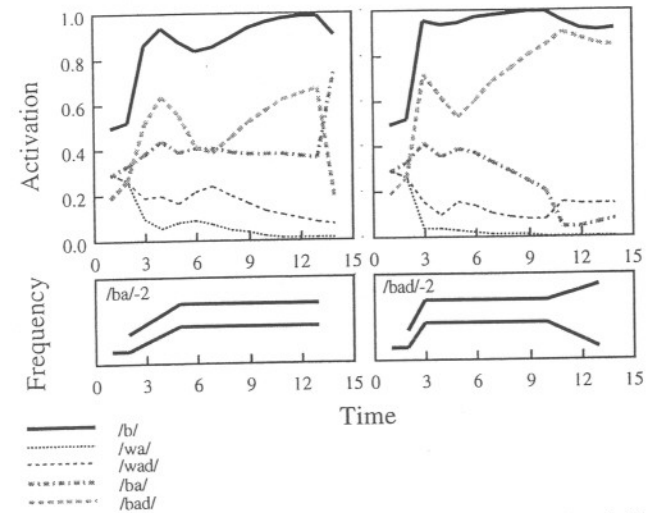

Figure 14.10 Activity of /b/ in /ba/-2 (top left) and /bad/-2 (top right), and the corresponding temporal representation of the stimuli at the given rate (bottom panels). The activity of the /w/ detector is purposely omitted here while those of /wa/ and /wad/ are included.

the figure shows, the identity of the syllable-initial /b/ is obvious right from the start since the TD values are exclusive to the /b/ category. In contrast to syllable-initial phoneme identification, however, syllable recognition occurs late. For the /bad/ syllables, activation increases gradually and reaches the maximum just before the onset of the final transition, whereas for the /ba/ syllables, activation remains low almost throughout the syllable but increases abruptly at the end of the vowel (which also coincides with the offset of the syllable). The recognition of other stimuli outside the category overlap follows the same description, i.e. syllable-initial phoneme recognition occurs as soon as the TD is determined, while syllable identification takes place only at the end of the vowel.

Discussion

There are two main results of the present simulations. First, the present findings are similar to those obtained from the preceding series of simulations in that rate information, specified by the overall stimulus duration, is important for the phonetic distinction between the syllable-initial segments, but only for items in the category overlap. Second, and more importantly, the network picked up the durational constraints we imposed on syllable structure. In the way that we have manipulated the CV and CVC stimuli, syllable structure, apart from giving information about overall syllable duration, also provides specific information about the duration of the phonetic segments that constitute the syllable. Our experiments show that this was crucial to the network in two instances: first, when making the relevant phonetic contrast for those items in the category overlap, and, secondly, when distinguishing between syllables of different structure but identical syllable-initial phoneme for all tokens along the TD continuum.

The work also speaks on the issue of whether information provided by syllable structure is used by listeners during perception. Miller & Liberman (1979) have shown that this was indeed the case for their subjects. Interestingly, however, their proposal that listeners calculate the number of phonetic segments to determine the articulation rate of a given syllable and use this to influence the phonetic categorization of an initial consonant does not fit with the network performance. The behaviour of our network suggests a more general strategy which involves making a durational contrast between the cue provided by the TD and the following adjacent segment, a vowel in this instance. Syllable structure, as noted earlier, reconfigures the vowel duration and TD with respect to the overall syllable duration, as is the case when a third segment is appended to the CV syllable (see Volaitis & Miller 1991). Possibly, this is geared to maintain perceptual intelligibility in much the same way that language communities intentionally regulate vowel length in order to enhance perceptually the closure duration cue for voicing distinctions (Kluender et al. 1988). The resulting configuration, particularly the relationship between a vowel and the initial TD, is learned by the network, and this information is used both to capture the relevant phonetic contrast and to generalize to new stimuli.

This explanation is consistent with the results obtained by Newman & Sawusch (1992). In examining the effects of adjacent and non-adjacent phonemes on the perception of the syllable-initial contrast between /sh/ and /ch/, cued by the duration of friction, in the /shwaes/–/chwaes/ series, they demonstrated that varying the /w/ duration produced the standard contrast effect, i.e. a longer /w/ made the initial segment seem shorter, or sound more like a /ch/, while a shorter /w/ made the initial segment seem longer, or sound more like a /sh/. Variation in the duration of the non-adjacent vowel, on the other hand, had no contrastive effect. In relation to the /bad/–/wad/ stimuli used in our computational experiments, the distant segment that does not contribute to the contrast effect is the final transition. The same explanation can also be used to interpret the data obtained by Volaitis & Miller (1991). In studying the role of syllable structure on the perception of VOT in /di/–/ti/ and /dis/–/tis/ syllables in the context of changing speech rate, they found that listeners adjusted for changes in VOT in relation to the syllable's CV duration, and not to its overall duration, which is consistent with our model.

Conclusion

Motivations for dealing with time-warped sequences have been discussed, and a successful recurrent network model has been described that can accommodate a range of rate-varying stimuli. Application of this model to a specific problem in phonetic perception has also been examined with encouraging results. To the extent that recurrent networks embody a learning account, one remaining issue raised by the model is whether the strategy of contrasting segmental durations is dependent on what is learned about the properties of actual speech, or is the strategy that is hardwired in the auditory system (Diehl & Walsh 1989). Recently, we completed another series of studies with these networks (Abu-Bakar & Chater 1994a,b) to investigate the viability of the model in simulating other phenomena in speech perception, namely shifts in category boundaries due to rate, experience and selective adaptation, and alteration to the internal structure of phonetic categories as a consequence of changes in speaking rate. The results from this work bring further implications for spoken language processing and models of perception and categorization of human speech (see Abu-Bakar & Chater 1994c).

Acknowledgements

Most of the material contained in this chapter has appeared elsewhere (Abu-Bakar & Chater 1993a,b). We gratefully acknowledge the helpful comments of an anonymous reviewer on an earlier draft of the manuscript. We thank the Department of Psychology, University of Wales, Bangor, and the Centre for Cognitive Science, University of Edinburgh, for the extensive use of computer facilities. Correspondence should be addressed to the second author.

References

Abu-Bakar, M. & N. Chater 1993a. Studying the effects of speaking rate and syllable structure on phonetic perception using recurrent neural networks. *Irish Journal of Psychology* **14**, 426–441.

Abu-Bakar, M. & N. Chater 1993b. Processing time-warped sequences using recurrent neural networks: modelling rate-dependent factors in speech perception. *15th Annual Conference of the Cognitive Science Society, Proceedings*, 191–7. Hillsdale, New Jersey: Lawrence Erlbaum.

Abu-Bakar, M. & N. Chater 1994a. Distribution and frequency: modelling the effects of speaking rate on category boundaries using recurrent neural networks. *16th Annual Conference of the Cognitive Science Society, Proceedings*, 3–8. Hillsdale, New Jersey: Lawrence Erlbaum.

Abu-Bakar, M. & N. Chater 1994b. Phonetic prototypes: modelling the effects of speaking rate on the internal structure of a voiceless category using recurrent neural networks. *3rd International Conference on Spoken Language Processing, Proceedings*. Tokyo: The Acoustical Society of Japan.

Abu-Bakar, M. & N. Chater 1994c. A recurrent neural network model of rate effects in phonetic perception. Unpublished paper, Department of Linguistics, University of Wales, Bangor.

Chater, N. 1989. *Learning to respond to structure in time*. Research Initiative in Pattern Recognition, RSRE Malvern, Technical Report RIPRREP/1000/62/89.

Chater, N. & P. Conkey 1994. Sequence processing with recurrent neural networks. In *Neurodynamics and psychology*, M. Oaksford & G. D. A. Brown (eds), 269–94. London: Academic Press.

Dalston, R. M. 1975. Acoustic characteristics of English /w, r, l/ spoken correctly by young children and adults. *Acoustical Society of America, Journal* **57**, 462–9.

Diehl, R. L. & M. A. Walsh 1989. An auditory basis for the stimulus-length effect in the perception of stops and glides. *Acoustical Society of America, Journal* **85**, 2154–64.

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

Huang, X. D., Y. Ariki, M. A. Jack 1990. *Hidden Markov models for speech recognition*. Edinburgh: Edinburgh University Press.

Kluender, K. R., R. L. Diehl, B. A. Wright 1988. Vowel-length differences before voiced and voiceless consonants: an auditory explanation. *Journal of Phonetics* **16**, 153–169.

Lippmann, R. P., E. A. Martin, D. P. Paul 1987. Multi-style training for robust isolated-word speech recognition. *IEEE International Conference on Acoustics, Speech, and Signal Processing, Proceedings*.

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

Maskara, A. & A. Noetzel 1992. Forced simple recurrent neural networks and grammatical inference. *14th Annual Conference of the Cognitive Science Society, Proceedings*, 420–7. Hillsdale, New Jersey: Lawrence Erlbaum.

Miller, J. L. 1981. Effects of speaking rate on segmental distinctions. In *Perspectives on the study of speech*, P. D. Eimas & J. L. Miller (eds), 39–70. Hillsdale, New Jersey: Lawrence Erlbaum.

Miller, J. L. & A. M. Liberman 1979. Some effects of later-occurring information on the perception of stop consonant and semivowel. *Perception and Psychophysics* **25**, 457–465.

Miller, J. L., F. Grosjean, C. Lomanto 1984. Articulation rate and its variability in spontaneous speech: a re-analysis and some implications. *Phonetica* **41**, 215–55.

Miller, J. L., K. P. Green, A. Reeves 1986. Speaking rate and segments: a look at the relation between speech production and perception for the voicing contrast. *Phonetica* **43**, 106–15.

Newman, R. S. & J. R. Sawusch 1992. Assimilative and contrast effects of speaking rate on speech perception. *Acoustical Society of America, Journal* **92**(suppl. 2), SP11.

Nguyen, M. & G. W. Cottrell 1993. A technique for adapting to speech rate. *IEEE Workshop on Neural Networks for Signal Processing, Proceedings*.

Norris, D. 1990. A dynamic-net model of human speech recognition. In *Cognitive models of speech processing: psycholinguistic and computational perspectives*, G. T. M. Altmann (ed.), 87–104. Cambridge, Mass.: MIT Press.

Rumelhart, D. E., G. E. Hinton, R. J. Williams 1986. Learning internal representations by error propagation. In *Parallel distributed processing: explorations in the microstructures of cognition*, vol. 1. *Foundations*, D. Rumelhart & J. McClelland (eds), 318–62. Cambridge, Mass.: MIT Press.

Shillcock, R., J. Levy, N. Chater 1991. A connectionist model of word recognition in continuous speech. *13th Annual Conference of the Cognitive Science Society, Proceedings*, 340–3. Hillsdale, New Jersey: Lawrence Erlbaum.

Shillcock, R., G. Lindsey, J. Levy, N. Chater 1992. A phonologically motivated input representation for the modelling of auditory word perception in continuous speech. *14th Annual Conference of the Cognitive Science Society, Proceedings*, 408–13. Hillsdale, New Jersey: Lawrence Erlbaum.

Summerfield, A. Q. 1981. Articulatory rate and perceptual constancy in phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance* **7**, 1074–95.

van Camp, D. & T. Plate 1993. *Xerion neural network simulator*. Department of Computer Science, University of Toronto.

Volaitis, L. E. & J. L. Miller 1991. Influence of a syllable's form on the perceived internal structure of voicing categories. *Acoustical Society of America, Journal* **89**(suppl. 2), SP10.

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

Watrous, R., B. Ladendorf, G. Kuhn 1990. Complete gradient optimisation of a recurrent neural network applied to /b/, /d/, /g/ discrimination. *Acoustical Society of America, Journal* **87**, 1301–9.

CHAPTER 15

Bottom-up connectionist modelling of speech

Paul Cairns, Richard Shillcock, Nick Chater, Joseph P. Levy

Introduction

Low-level phonological information plays an important role in spoken word recognition. It is certain that listeners are highly sensitive to simple sequential phonological patterns – thus any speaker of English can instantly say which of the following are possible words in their language: /snarp/, /mplaf/, /krad/, /sakf/. Furthermore, one intuitively rates legal examples such as /sfɪp/ as being less "normal" than items such as /stɪp/. Although it might be possible for such judgements to be mediated by rapid calculation of some lexical intersection, for reasons of computational efficiency it would be advantageous for this type of information to be represented in summary form in a component of the human language processor. We believe that this sublexical information impinges upon higher-level processes such as lexical access, and there is evidence to show that this is the case (e.g. Jakimik 1979, Foss & Gernsbacher 1983). In this chapter we describe a system that learns to encode this type of information. Our system is not a model of any particular psycholinguistic process in itself; rather, it can be used as a tool to represent and assess the possible influence of sublexical phonological information in specific cognitive processes.

In recent years, statistically based models have come to dominate computational psychological modelling, with neural networks playing an increasingly prominent role. However, it is frequently the case that such work is prone to a particular misunderstanding about the nature of statistical modelling: in order to stand as a valid model of human behaviour, and particularly *learning* behaviour, a statistical system must be derived from input that is representative of genuine natural language input – yet this fact seems to be forgotten in much connectionist modelling work. Thus, a model which employs a "toy" training set, of say a dozen lexical items, can only make defensible claims about human behaviour to the extent that its input is statistically representative of natural input –