The Irish Journal of Psychology, 1993, **14**, 3, 410-425

Studying the effects of speaking rate and syllable structure on phonetic perception using recurrent neural networks

Mukhlis Abu-Bakar
University of Wales, Bangor
&
Nick Chater
University of Edinburgh

We apply recurrent neural networks to the processing of time-warped sequences, particularly, modelling how listeners distinguish between phonetic categories in the context of changing speech rate. In an earlier paper (Abu-Bakar & Chater, 1993), we modelled the effects of speaking rate on the perception of voicing contrasts specified by voice-onset-time (VOT) in syllable-initial stop consonants using a simple coding procedure. In the present investigation, we apply a more detailed speech representation to model the effects of both speaking rate and syllable structure on the syllable-initial distinction between a stop consonant /b/ and a semivowel /w/ cued by the duration of the formant transitions. In the first set of experiments, we constructed nine pairs of /ba/-/wa/ syllables varying in syllable duration and transition values. In another set of experiments, we compressed these syllables and added syllable-final transitions appropriate for a stop consonant /d/ to produce a second set of syllables (/bad/-/wad/). On both occasions, information about speaking rate provided by the syllable duration and structure was relevant to the contrast, but only for tokens with overlapping transition values. In adjusting for changes in transition duration, the network relied on the duration of the syllable's CV component, instead of the entire syllable. We discuss the implications of these results for our understanding of how human listeners make precise use of rate information during phonetic perception.

Address for correspondence: Muklis Abu-Bakar, University of Wales, Bangor. Nick Chator, Centre for Cognitive Science, University of Edinburgh.

In normal conversation, speakers often do not maintain a constant rate of speech (Miller, Grosjean & Lomanto, 1984). Importantly, the change in rate involves not only a change in the number and duration of pauses in the utterance but also a change in the temporal characteristics of the speech signal itself (Miller & Baer, 1983; Miller, Green & Reeves, 1986; Volaitis & Miller, 1992). This makes the mapping between signal and phonetic percept a potentially complex task.

One temporal property that has been thoroughly studied vis-a-vis changes in speaking rate is voice-onset-time (VOT). It is well established that voiced consonants have shorter VOT values than do voiceless consonants and that listeners can use this difference to identify a given consonant as voiced or voiceless (Lisker & Abramson, 1964). However, a change in speaking rate can result in a considerable shift in the distribution of VOT values, especially for voiceless stops. Miller *et al.* (1986) examined this effect in some detail for the contrast between /b/ and /p/. They found that as syllable duration increased, there was only a slight increase in the VOT values for /b/, whereas for /p/ there was a substantial increase in these values. The result was a fair amount of overlap in the VOT distributions for the voiced and voiceless tokens around the region of small VOT values. The question is how does the perceptual system maintains accurate voicing decisions in the face of these variations, especially when confronted with tokens produced with short VOT values.

One account is based on the idea that the perceptual system categorizes accurately by adjusting the criterion VOT value used to make the voicing decision with respect to speaking rate (Green & Miller, 1985; Summerfield, 1981). Findings from experiments that mainly employ identification procedures have been taken as evidence to support this rate-dependent processing hypothesis. For example, Green and Miller (1985) demonstrated that the overall duration of a target syllable, which varies as a function of speaking rate, influences listeners' voicing judgement on the syllable's initial stop consonant as either /b/ or /p/. They edited from natural speech /bi/-/pi/ VOT series that differed from each other in overall syllable duration. These stimuli were presented to listeners who were asked to identify each stimulus as either /b/ or /p/. The main finding was that as the syllables became longer, the perceptual category boundary shifted toward a longer VOT value.

The relationship between syllable duration and voicing boundary is, however, not as straightforward. As Summerfield (1981) has shown, extending the overall duration of a target syllable by adding acoustic information corresponding to a third phonetic segment (as in /biz/& /piz/) actually has the effect of shifting the voicing boundary toward shorter VOT values. This finding was also obtained by Miller and Liberman (1979) with respect to the stop-semivowel distinction of /bad/ versus /wad/. They concluded that articulation rate is not calculated solely on overall syllable duration, but also on the number of phonetic segments in the syllable. Information about articulation rate thus obtained is then used by the perceptual system to influence the phonetic categorization of the initial consonant.

This explanation is indeed plausible. However, the assumption that the normalization process that leads to the boundary shift is dependent on listeners' sensitivity to variation in articulatory rate has not been uniformedly supported in the literature. Other investigators have shown that such boundary shifts can also be accounted for by the general auditory principle of durational contrast. This account proposes that perceived length of a given acoustic segment is affected contrastively by the duration of adjacent segments (Pisoni, Carrell & Gans, 1983; Kluender, Diehl & Wright, 1988; Diehl & Walsh, 1989; Newman & Sawusch, 1992). According to this account, a long vowel will make the formant transitions seem shorter and, hence, more stop-like. This principle of durational contrast applies equally to the perception of speech and nonspeech signals, unlike that of the rate normalization account offered by Miller and Liberman (1979).

While the effect of speaking rate on perception continues to be a large area of research, studies that examine the precise use of rate information in phonetic perception are relatively rare. Recently, we embarked on a computational study of this issue using 'speech-like' stimuli (Abu-Bakar & Chater, 1993). The focus was on the overlap that exists in the distribution of contrasting phonetic categories along some critical continua. This notion of overlap can be conceived of in at least two ways. One can speak of overlap with respect to a particular speech rate specified by the syllable duration; this normally manifests itself at fast rates of speech. The concept of overlap can also be seen from a more global perspective. Taking the VOT as an example, global overlap 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. In our simulations, we tested the network's sensitivity towards this global overlap. Specifically, the rate processing account appears not to hold for stimulus items outside the range of overlap since, for these items, taking rate into account is unnecessary. Such a strategy simplifies the perceptual problem and reduces computational overheads since a large chunk of the category tokens can be dealt with straightforwardly leaving the system free to attend to the heavier demands of the ambiguous items.

Extending beyond voicing contrast

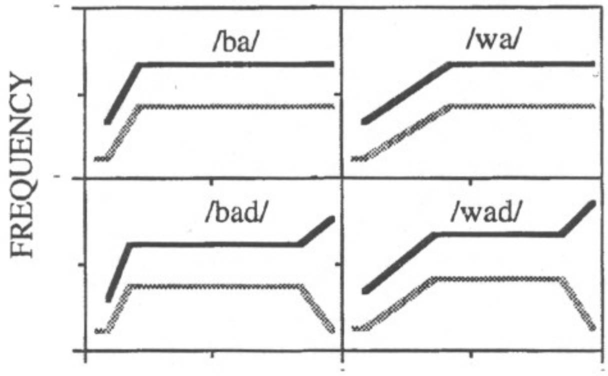
The /b/-/p/ distinction is but one of a family of contrasts whose production and perception is affected by a change in rate. If our model is truly robust, it should be able to handle all contrasts, regardless of the nature of the acoustic information underlying the phonetic distinction. An interesting case in point is the stop-semivowel contrast, /b/ versus /w/ in this instance, specified by a change in transition duration (hereafter referred to as TD) or the abruptness of their onsets. In the studies of the production of these consonants, the onset for /b/ 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 increased, the /b/-/w/ boundary moved toward transitions of longer duration (Miller & Liberman, 1979). As indicated earlier, however, this boundary moves in the opposite direction when increase in syllable duration is effected by adding a final transition corresponding to a third phonetic segment.

In what follows, we will give a description of our model in detail. We then describe two computational experiments that test the performance of the model with respect to the /b/-/w/ contrast. One way in which the present set of experiments differ in complexity from the previous study is in the use of distributed codes to represent each time fragment of the stimuli. This enables us to capture the changes in formant values across the TD segment. In the first experiment, we ask if the same phenomenon observed for the voicing contrast also applies to the stop-semivowel contrast. We expect tokens outside the category overlap to be recognised immediately once the rate of formant movements is determined, while for those inside the category overlap, their recognition will be delayed relative to the duration of the syllable. In the second experiment, we want to establish the way in which the combined effect of syllable structure and syllable duration affects the recognition of the syllable-initial segments (see Figure 1).

Recurrent neural networks

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, Lindsey, Levy & Chater, 1992). 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).

Figure 1. A schematised two-formant representation of syllables /ba/-/bad/-/wa/-/wad/ over time. Notice the difference in TD across category and syllable structure - three are outside the category overlap [Test and training tokens].



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 is Elman's (1990) "copy-back" scheme which uses a computationally cheap approximation to gradient descent in error to change the weights. We use "backpropagation through time" (Rumelhart, Hinton & Williams, 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 backpropagation. 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.

An important parameter in backpropagation through time is how many time steps the recurrent network is unfolded. The larger this number, the more exactly the network computes true gradient descent, although the benefits of additional unfoldings rapidly tail off, because very deep feedforward networks are very slow to train. It is also important to note that

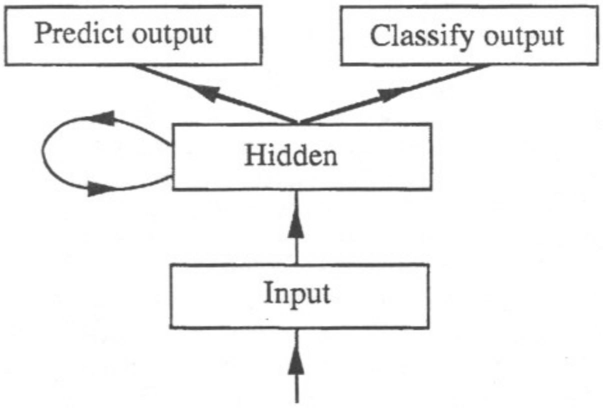
'bad/-/wa/-
e structure

the units
ıts which
k is often
units, to
ɔrding to
put units
a variety
ɔy-back"
gradient
through
gradient
equence
esulting
traint on
network
:, so that
ecurrent

is how
ıumber,
ɔugh the
y deep
ıote 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 (1993) for discussion). Training uses conjugate gradient descent, and is implemented on the Xerion simulator (van Camp, Plate & Hinton, 1992).

Figure 2. The architecture of the recurrent network used (see text for explanation).



In the simulations reported here, between 47 and 58 unfoldings are used. The architecture used (shown in Figure 2) comprises a three layer network, with 31 input units representing the current input. These 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 (at intervals of 50 Hz), and if a formant has frequency *F*, then all and only the units which represent frequency values *F* and less will be active. In the first experiment, there are 30 hidden units with recurrent connections (60 in the second experiment), which seems to be approximately the smallest number of units that can learn the task successfully. The output units can be conceived of as divided into two banks. The most important bank of units simply consists of two units (six in the second experiment) which represent the network's classification of which syllable is being heard. In addition, there is also an additional bank of units which is trained to predict the next but one pattern in the input sequence. This additional prediction task seems to improve the network's performance on the categorization task by requiring the network to extract more of the structure of the input sequence (cf. Maskara & Noetzel 1992; Shillcock, et al., 1992).

Neither the classification nor prediction tasks can be solved perfectly by the network for it is not determinate which phonetic category is being heard until a considerable amount of the sequence has been encountered. The optimal performance is to withhold a firm decision about which phonetic item is encountered just until the sequence can be classified unambiguously. This optimal pattern of behaviour will therefore be very sensitive to the precise degree of global overlap between the different categories. We find, in the simulations reported below, that network performance is indeed, qualitatively in line with this pattern of behaviour.

EXPERIMENT 1

In this experiment, we want to establish for the /b/-/w/ distinction the conditions in which later-occurring information in the syllable is used to perceive an earlier-occurring cue.

Stimuli

The stimuli for the two experiments we report in this paper are loosely patterned after the synthesised speech used by Miller and Liberman (1979). For the present experiment, we constructed nine pairs of /ba/-/wa/ syllables, each corresponding to a different speech rate. The syllables were two-formant patterns, consisting of a brief prevoicing (first formant only), a variable duration of formant transition appropriate for /b/ or /w/, and a subsequent period of steady-state formants. The formant movements were approximated from the following frequency specifications: a rise linearly for both the first (F1) and second (F2) formants from their starting frequencies of 234 Hz and 616 Hz to 769 Hz and 1232 Hz respectively. We built the /wa/ syllables with longer TDs and over a broader range than the /ba/ syllables which, in contrast, were concentrated over a shorter range at the lower end of the TD scale (Figure 3). Two of these pairs were set aside as test items.

: important
‹periment)
le is being
h is trained
additional
ce on the
e structure
al., 1992).
olved per-
:ategory is
:n encoun-
›out which
classified
re be very
: different
t network
›ehaviour.

nction the
is used to

re loosely
an (1979).
syllables,
were two-
it only), a
w/, and a
ients were
e linearly
r starting
ively. We
e than the
r range at
: set aside

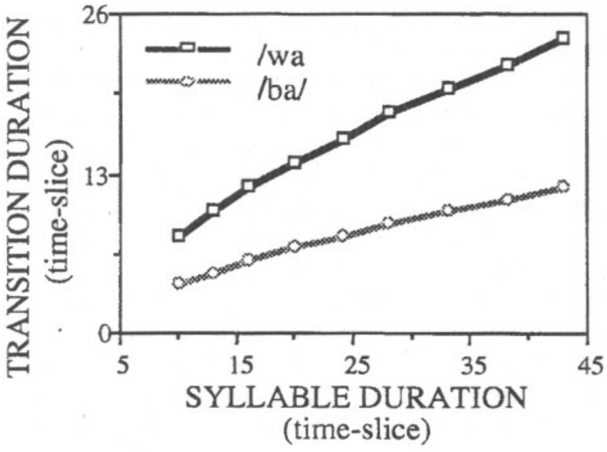


Figure 3. Distribution of /ba/-/wa/ syllables in the TD-syllable duration space.

Results

After 2000 sweeps through the training corpus, the network correctly recognised all the training stimuli. Our concern, however, is with how well the network generalizes - that is, how well it recognizes syllable-initial /b/ -/w/ spoken at a new speech rate, especially for two of the four test tokens which were ambiguous from the start due to their identical TDs. In general, the network performed as expected, with its response similar to that obtained in the previous study. Importantly, the network paid attention to the change in the formant frequency values. More specifically, information about the formant transition was not sufficient to disambiguate items in the overlap region (Figure 4). For these items, activation values reach optimal point either towards the end of the sequence, as is the case with the /ba/ syllable, or at exactly the point of syllable offset, as in the case of the / wa/ syllable which is the shorter of the two in this range. On the other hand, for tokens outside the overlap region, the rate of the formant movements alone was adequate to trigger a correct and early response, with activation values for the appropriate nodes rising high immediately after the first few time slices through the transition (Figure 5).

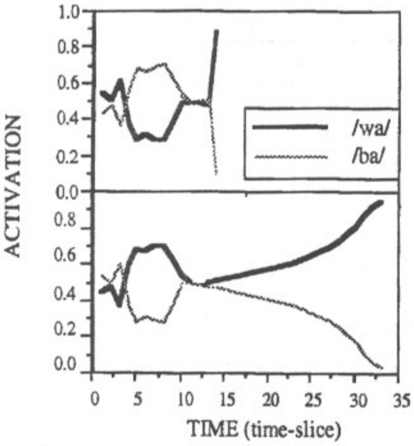


FIGURE 4. Activation curves for (a) a /wa/ syllable which shares the same TD value (10 t/s) as (b) a /ba/ syllable of a longer syllable duration [Test tokens].

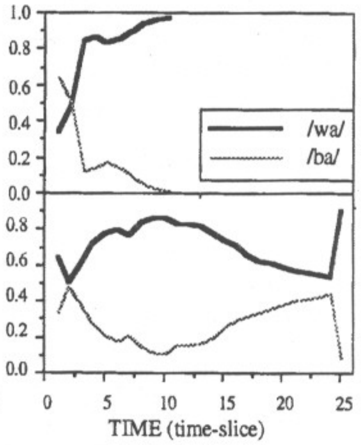


FIGURE 5. Activation curves for (a) a very short /ba/ syllable with a characteristically /b/ TD value (4 t/s), and (b) a moderately long /wa/ syllable with a characteristically /w/ TD value (16 t/s) [Training tokens].

It must be acknowledged that some aspects of the network's behaviour were unexpected. We did not anticipate the gradual decline in activation for the long /wa/ syllables, especially after the boost cued earlier by the TD (see Figure 5b). In fact, we noticed a general trend toward a declining activation during the steady-state portion of all the /wa/ syllables; the effect being greater with increasing syllable duration. We were hard pressed for an answer as to why this occurs at all, and especially why this is so apparent for the long /wa/ syllables. Any explanation, however, will only be conjectural at this stage until a thorough analysis of the hidden units are carried out.[1] For the present, we were satisfied with the general pattern of the network's behaviour.

[1] Although the reason for this is computationally not clear at present, a striking parallel can be found with empirical data. Miller (1987) reported a reaction time asymmetry with respect to the identification of the /ba/-/wa/ syllables. She observed that for stimuli with long transition durations that are consistently identified as /w/ , there is a reliable increase in reaction time with increasing syllable durations. This effect, however, does not happen with stimuli typically recognised as /b/.

EXPERIMENT 2

In this experiment, we have two main aims: firstly, we want to establish whether the conditions which constrain the use of syllable duration information for the /b/-/w/ contrast also apply to the use of information about syllable structure; secondly, we want to determine the ways in which the combined effects of these two variables affect the recognition process.

Stimuli

An additional two pairs of /ba/-/wa/ syllables were added to the original stimuli. From these 11 pairs, /bad/-/wad/ syllables were constructed by compressing the /ba/-/wa/ syllables and inserting a final transition appropriate for the stop consonant /d/. The effect is that, for any given syllable duration, the TD value is always shorter for syllables with a CVC structure than for those with a CV structure (see Miller & Liberman, 1979; Summerfield, 1981; cf. Volaitis & Miller, 1991). Furthermore, every increase in syllable duration, either by adding steady-state or transition segments, effectively corresponds to the duration of the final transition, which in turn does not remain constant across speech rate. During the final transition, the first formant fell linearly from its steady-state value of 769 to 234 Hz, while the second formant rose linearly from its steady-state level of 1,232 Hz to 1,541 Hz (see Figure 1). At the output layer, these changes are reflected by the number of units added; one each for /bad/ and /wad/, and another for /b/ and /w/, making it a total of six output units. 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 itself.

The distribution of the 11 pairs across TD and syllable duration is shown in Figure 6. 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 (three each across category and syllable structure) of varying syllable duration were reserved as test items.

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 7 shows the activity curves of two of the four test tokens with identical TD value. Since the TD is ambiguous between /b/ and /w/, the activation of the respective phoneme detectors for both syllables remain very low for much of the duration of the syllables. In fact, for the /wad/ syllable, the activation of the /w/ detector shoots up only at the offset

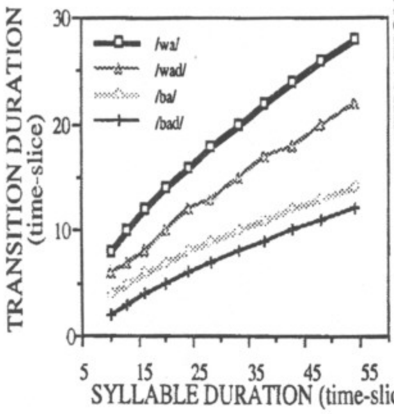


FIGURE 6. Distribution of /ba/- /bad/- /wa/- /wad/ syllables in the TD-syllable duration space.

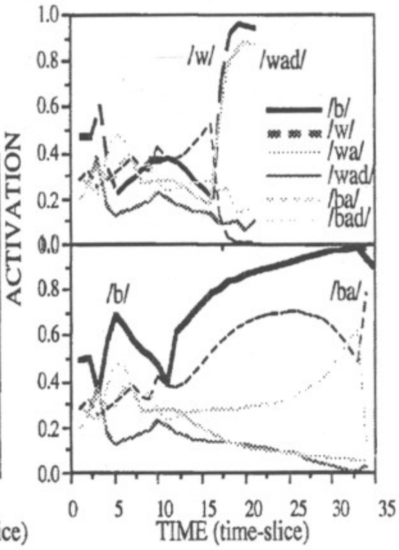
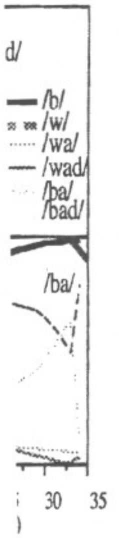


FIGURE 7. Activity curves of the syllables (a) /wad/ and (b) /ba/. They have identical TDs (10 t/s) but differ in syllable duration. Only the activation of the crucial phoneme detector is shown in each diagram [Test tokens].

of the steady-state section of the syllable (Figure 7a). It is also at this point that the syllable is distinguished from the other syllables by way of an abrupt increase in its activation. it should be noted that the activation pattern for the recognition of the /wa/ syllable of the same TD (not shown here) also follows a similar description, with the activation for both the /w/ and /wa/ detectors rising abruptly at the end of the vowel.

The recognition of the /ba/ syllable, on the other hand, proceeds in a slightly different fashion. After processing a number of time frames through the steady state portion of the syllable, the network realizes that there can be no /wa/ or /wad/ syllables with a steady-state value beyond that point. It therefore pushes the activation of the /b/ detector up at the expense of the /w/ detector from that point onward until the activation reaches an optimum (Figure 7b).



syllables
ientical
iration.
ioneme
Test

iis point
i abrupt
tern for
re) also
nd /wa/

eeds in
frames
:es that
nd that
xpense
:hes an

The effect of syllable structure can be seen at the end of the /ba/ syllable when the activation curves for the /ba/ and /bad/ detectors switch direction. The default seems to be that the /bad/ detector increases in activation until such time when there can be no /bad/ syllable but a /ba/ syllable with a steady-state value beyond that point. Its activation thus falls rapidly to give way to the /ba/ syllabl.e.

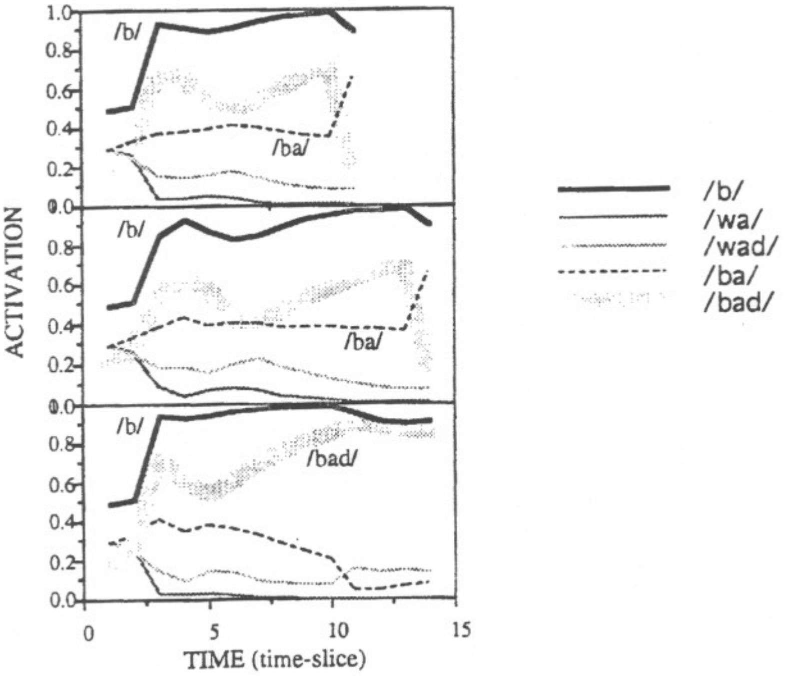


Figure 8. Activity curves for (a) a /ba/ syllable of 10 t/s syllable duration, (b) a /ba/ syllable of 13 t/s syllable duration, and (c) a /bad/ syllable of 13 t/s syllable duration. All three are outside the category overlap [Test and training tokens].

Earlier we asked if rate information is still crucial for the phonetic contrast between /b/ and /w/ if the TD value is outside the category overlap. Here we have chosen to examine a group of fast /ba/ and /bad/ syllables. As Figure 8 shows, the identity of the syllable-initial /b/ is obvious right from the start since the TD values are exclusive to the /b/ category. This applies to all /ba/ and /bad/ syllables outside the category overlap.

In contrast to the syllable-initial segment identification, syllable recognition occurs late, with the classification of the /bad/ syllables occuring just before the offset of the vowel segments, while for the /ba/ syllables, they are distinguished from /bad/ only at the end of the vowel (which also coincides with the offset of the syllable). It should be pointed out that three syllables similarly taken from within the category overlap (not shown here) would have produced a different picture, that is, the phonetic identity of the syllable-initial consonants would have been established much later in the sequence and at the same time as the classification of the syllables.

The above seems to indicate that phonetic and syllable identity are both processed in relation to the syllable's CV component, and not the entire syllable. Specifically, the durational contrast between the critical TD and the adjacent vowel is a crucial factor affecting the perception of the syllable-initial segments and the syllables.

DISCUSSION

There were two main results of the present series of experiments. First, the present findings are similar to those obtained from our previous study (Abu-Bakar & Chater 1993) in that rate information, specified by the overall syllable duration, is important for the phonetic distinction between the syllable-initial segments, but only for items in the category overlap. Second, and more importantly, our network picked up the constraints imposed by syllable structure on the duration of the temporal characteristics that provide the phonetic contrast. That is, 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 findings show that this was crucial to the network in two circumstances; firstly, 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-intial phoneme for all tokens along the TD continuum.

This work highlights the issue of whether the information provided by syllable structure is also used by listeners during perception. Miller and Liberman (1979) have shown that this was indeed the case for their subjects. 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 inital consonant does not seem to converge with the present findings. 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. What is noteworthy about syllable

ecogni-
ing just
they are
incides
yllables
) would
of the
r in the
s.
tity are
e entire
TD and
of the

irst, the
(Abu-
overall
een the
verlap.
straints
eristics
t from
pecific
ute the
in two
ast for
uishing
al pho-

ovided
ler and
bjects.
onetic
se this
es not
twork
itional
jacent
llable

structure is that it reconfigures the duration of the vowel and the 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 vowel and initial transition duration, 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 and 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 - that is, 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 distal segment that does not contribute to the contrast effect is the final transition. The same explanation can also be used to interprete 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, a finding similar to ours.

The present model thus embodies a possible auditory basis for speech perception. This model is not built with any special rate parameter that gets adjusted on-line to cope with different presentation speeds, yet it was able to perform the categorization task successfully via a simple technique that contrasts durations of adjacent segments. That this model invokes an auditory-based mechanism does not, however, diminish the importance of speech-specific knowledge in phonetic categorization. There is little doubt that listeners make extensive use of speech-specific tacit knowledge in identifying phonetic segments. But, as pointed out by Diehl and Walsh (1989), the problem of speech perception can be made more theoretically tractable if general auditory mechanisms are exhausted first where possible with speech-specific knowledge being appealed to only as a last resort. Auditory explanations are simpler, more general, and more likely to follow from known principles (Diehl, Kluender & Walsh, 1990).

ACKNOWLEDGEMENTS

We would like to thank Joanne L. Miller for helpful comments on an earlier version of the manuscript. We are also grateful to the Centre for Cognitive Science, University of Edinburgh for making the computing resources available.

REFERENCES

Abu-Bakar, M. & Chater, N. (1993). Processing time-warped sequences using recurrent neural networks: Modelling rate dependent factors in speech perception *Proceedings of the 15th Annual Meeting of the Cognitive Science Society,* Hillsdale, NJ: LEA.

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

Chater, N. & Conkey, P. (1993). Sequence processing with recurrent neural networks. In M. Oaksford & G. D. A. Brown (Eds.), *Neurodynamics and Psychology,* London: Academic Press.

Dalston, R. M. (1975). Acoustic characteristics of English /w, r, l/ spoken correctly by young children and adults. *Journal of the Acoustical Society of America,* **57**, 462-469.

Diehl, R. L. & Walsh, M. A. (1989). An auditory basis for the stimulus-length effect in the perception of stops and glides. *Journal of the Acoustical Society of America*, **85**, 2154-2164.

Diehl, R. L., Kluender, K. R. & Walsh, M. A. (1990). Some auditory bases of speech perception and production. In *Advances in Speech, Hearing and Language Processing,* Vol. 1. JAI Press.

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

Green, K. P. & Miller, J. L. (1985). On the role of visual rate information in phonetic perception. *Perception & Psychophysics*, **38**, 269-276.

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

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 A. W. Ellis (Ed.) *Progress in the Psychology of Language Vol. 3,* Hillsdale, NJ: LEA

Miller, J. L. & Baer, T. (1983). Some effects of speaking rate on the production of /b/ and /w/. *Journal of the Acoustical Society of America,* **73**, 1751-1755.

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.

ı earlier
gnitive
;ources

:s using
:ch per-
Science

Report
, RSRE

neural
:ics and

orrectly
merica,

h effect
ciety of

'speech
nguage

9-211.
honetic

erences
urnal of

ı initial

'orks to
ı Artifi-

V. Ellis
LEA
ction of
1755.
A look
voicing

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

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

Newman, R. S. & Sawusch, J. R. (1992). Assimilative and contrast effects of speaking rate on speech perception. *Journal of the Acoustical Society of America*, **92**, 2300.

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, Mass: MIT Press.

Pisoni, D. B., Carrell, T. D. & Gans, S. J. (1983). Perception of the duration of rapid spectrum changes in speech and nonspeech signals. *Perception & Psychophysics*, **34**, 314-322.

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.

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 14th 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. (1991). Influence of a syllable's form on the perceived internal structure of voicing categories. *Journal of the Acoustical Society of America*, **89**, 1998.

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.