Distribution and frequency: Modelling the effects of speaking rate on category boundaries using a recurrent neural network

Mukhlis Abu-Bakar
Department of Psychology
University of Wales, Bangor
Gwynedd LL57 2DG
U.K.
mukhlis@cogsci.ed.ac.uk

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

Abstract

We describe a recurrent neural network model of rate effects on the syllable-initial voicing distinction, specified by voice-onset-time (VOT). The stimuli were stylized /bi/ and /pi/ syllables covarying in VOT and syllable duration. Network performance revealed a systematic rate effect: as syllable duration increases, the category boundary moves toward longer VOT values, mirroring human performance. Two factors underlie this effect: the range of training stimuli with each VOT and syllable duration, and their frequency of occurrence. The latter influence was particularly strong, consistent with exemplar-based accounts of human category formation.

Introduction

Speaking rate is a well-established source of contextual variation in the speech signal for which listeners must compensate. The effects of rate variation are complex. For example, Miller & Liberman (1979) examined the effect of speaking rate and syllable structure on the stop-semivowel distinction, specified by a change in the formant transition duration. When syllable duration is increased by lengthening the vowel, they found that the stop semivowel boundary moves toward transitions of longer duration. But when syllable duration is lengthened by adding a final transition corresponding to a third phonetic segment, this boundary moves in the opposite direction.

Miller & Liberman argued that speakers compensate by normalizing for speaking or "articulatory" rate, defined in terms of syllable duration and the number of phonetic segments in the syllable. It has been noted that an account that is dependent on listeners' sensitivity to variation in articulatory rate cannot explain human subjects' categorization of analogous nonspeech stimuli (Diehl & Walsh, 1989) nor nonhuman subjects' discrimination of speech stimuli (Stevens *et al.*, 1983).

An alternative account is that some rate effects on phonetic perception are derived from the general auditory principle of durational contrast that applies to speech and nonspeech signals alike (Diehl & Kluender, 1989). According to this account, when speaking rate is varied, those changes that occur closest to the target segment will most affect its perception. Consider the syllables /bla/ and /pla/. The auditory model predicts that varying the /l/ should have a greater effect on the voicing distinction than varying the more distant /a/, whereas the articulatory model predicts that the effect is just as strong irrespective of which segment is varied, as long as the overall syllable duration is varied (*cf.* Newman & Sawusch, 1992).

We have trained a recurrent neural network on rate-varying speech-like stimuli (Abu-Bakar & Chater, 1993b), and compared its performance with these divergent predictions. The network was trained to classify stimuli as /ba/, /wa/, /bad/ or /wad/ and trained to decide if an initial /b/ or a /w/ had been encountered. From the results, it was evident that the duration of the syllable's CV component provided the network with reliable and sufficient information to distinguish the initial consonants. That is, irrespective of syllable structure (CV or CVC), the identity of the syllable-initial consonant was distinguishable on the basis of a durational contrast between the transition duration and the adjacent vowel. The network therefore appeared to behave in line with the auditory account.

This might suggest that the network could simply be viewed as a computational instantiation of the auditory account. However, the fact that the network *learns* to apply durational contrast suggests a possible modification to the standard auditory view, in which it is assumed that the contrast strategy is wired into the structure of the auditory system. In this paper, we elaborate the suggestion that effects normally viewed as falling out of the structure of the auditory system might also be learned from experience of language.

We begin by noting Diehl & Kluender's (1989) distinction between the space in which speech and nonspeech sounds are represented and the partitioning of that space into categories. For speech, this partitioning is properly called "phonetic". But humans' categorization of speech and nonspeech signals (Diehl & Walsh, 1989), and animals' discrimination of speech sounds (Kuhl, 1988) are so strikingly similar - they all correspond to regions of relatively high auditory discriminability - that pre-existing auditory boundaries are often assumed to be natural locations for phonetic partitioning. But experimental studies that show discrimination peaks at identification boundaries are equally consistent with a learning interpretation, as it appears that discrimination can be nearly the same across entire continua when a reasonable amount of experience and training is implemented.

The partitioning of auditory space has a prototype structure so that some stimuli are perceived as better category members than others (Samuel, 1982). Transitory prototypes may also be created in conditions such as selective adaption where selected stimulus items are

repeatedly exposed. In non-experimental settings, one can find, say, a vowel, not always ending up quite where the auditory "hot spot" is, even if the whole vowel system is subject to auditory-based selection pressures. What is significant in these cases is the apparent correspondence between the prototypical location and the perceived category boundary. Repp & Liberman (1987) sum up this observation by proposing that category prototypes are responsible for determining boundaries. That is, phonetic boundaries may not conform to boundaries set by discontinuities in the auditory system, but are instead flexibly determined by the acoustic consequences of the articulatory gestures specifying the category prototype. The category prototypes, and hence category boundaries, may therefore be learned from exposure to language rather than determined by auditory discontinuities (Kluender *et al.*, 1987). Hence, they may be learnable by mechanisms, such as a connectionist network, which do not embody auditory constraints. Here, we show that a learning account, embodied in a recurrent neural network, can capture the results outlined above, focussing on the /bi/-/pi/ distinction, specified by the VOT, in the context of changing speech rate.

Description of the Model

Recurrent neural networks (e.g. Elman, 1990) are very attractive for problems concerned with speech processing because they are suited to processing sequential material. 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 simulations, the network is trained to classify input sequences into a small number of categories corresponding to different syllables. The network is fed with the relevant sequences one input pattern at a time, with the target output pattern kept 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.

In addition to identifying the syllable presented, a set of output nodes was trained, at time t, to attempt to predict the input pattern at time t+2 (Fig. 1). This forces the network to encode the input sequence more deeply leading to better network performance (*cf.* Abu-Bakar & Chater, 1993a; Maskara & Noetzel, 1992; Shillcock *et al.,* 1992). The network was trained by recurrent backpropagation (Rumelhart *et al.*, 1986) which computes gradient descent by "unfolding" the recurrent network into a sequence of serially connected feedforward networks, and then trains the resulting network using standard backpropagation. 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

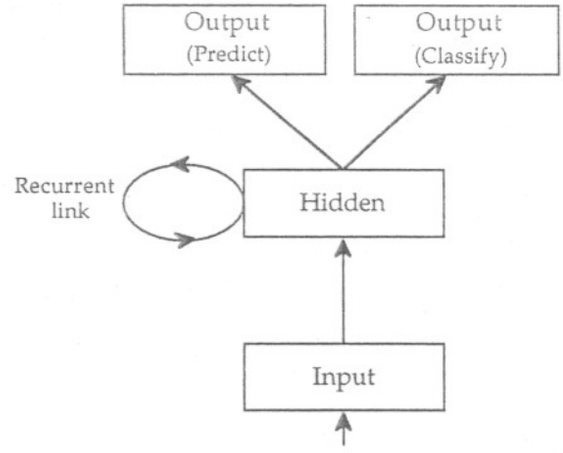

Figure 1: The recurrent network used in the simulations. It is unfolded during training for as many time-steps as required to accomodate the longest training stimuli.

off after some point, because very deep feedforward networks are very slow to train (see Chater and Conkey, 1994). Training used conjugate gradient descent and implemented on the Xerion simulator (van Camp & Plate, 1993). The number of input, hidden and output units was 30, 60 and 32 respectively.

Stimuli

Training stimuli were based on a two-formant syllable with an initial period of formant transitions followed by a steady-state (vowel) (Fig. 2). Frequency values were represented binarily using several input units. Each unit represents a particular range of frequency (at intervals of 20 Hz (F1) and 40 Hz (F2)). If a formant has frequency F, then all and only the units which represent frequency values F and less will be active.[1] One group of units, which consisted of two further sub-groups (corresponding to F1 and F2 units), represents formants with a periodic source, while another group represents formants (namely, F2) excited with a noise source. Beginning with the end-point /bi/, we built a pool of /bi/ and /pi/ syllables by varying VOT. This is effected by simultaneously switching off the activation of the periodic

[1] This may not seem the most ideal method of representing frequency information nor a realistic description of the neural encoding of speech sounds in the auditory pathways (see Greenberg, 1988, for a review). However, given the specific focus of this paper, a more complex representation is not considered essential. A further simplification is that time is measured in time-steps (henceforth "ts") rather than milliseconds to allow flexibility in apportioning segmental lengths with a view to lower computational demands. Apart from ensuring that the relative VOT distribution of the /b/ and /p/ category is observed, there is no pressing need for the syllable length to be kept similarly natural. Thus for each "rate", vowel segment is cut back such that its proportion to the VOT is smaller than that found in natural speech. Since the number of unfoldings is set to be dependent on the length of the longest stimuli, shorter stimuli help to trim the number of unfoldings, leading to faster training time.

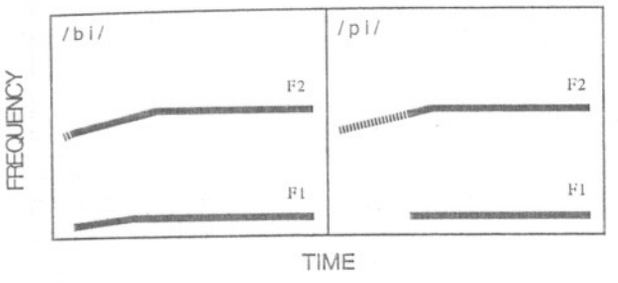

Figure 2: Schematic representations of the formant motions of the endpoint stimuli corresponding to /bi/ (left) and /pi/ (right). Each representation consists of an interval of aspiration (striped line), followed by onset of voicing (dark lines).

units of F1 and F2 and activating the F2 noise units for the appropriate duration. This can be interpreted as eliminating all energy in the region of F1 and replacing the higher formants (only F2, in this instance) with noise.

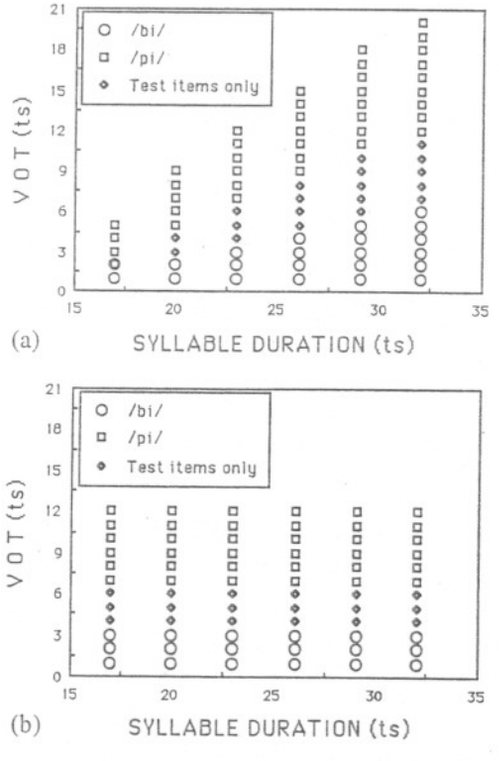

Figure 3: Location of /bi/-/pi/ tokens on the VOT and syllable duration space, in ts, for (a) the "natural" training set and (b) the "artificial" training set. Distribution of tokens used as tests only are also shown on each panel.

Procedure

The network's task is continuously to rate the probability of a stimulus belonging to a particular category as the stimulus unfolds. The nature of the voicing distribution makes it possible for the net to identify the consonant early in the sequence for some syllables, particularly those whose VOT values are unambiguous (see Abu-Bakar & Chater, 1993a, 1993b). But a thorough evaluation of category goodness is possible only when the net has scanned the entire length of the syllable and the proportion of VOT to syllable duration has been calculated. The activation values of the output units at the offset of each syllable was therefore taken as an accurate measure of the probability that a stimulus belongs to the category which the unit represents.

Distribution Effects on Category Boundary

In this study, we look at how VOT distribution, in relation to varying speaking rate, plays a role in influencing the boundary locations of /b/ and /p/. The range of VOT/syllable durations, shown in Fig. 3(a), is based on the production patterns of the voiced and voiceless tokens studied by Volaitis & Miller (1992). Notice that as syllable duration increases, the VOT specifying a category also increases but this increase was greater for /p/ than for /b/. In addition, the width of the range of VOT of each category also increases.

We also trained another network on an "artificial" training set whose members were artificially distributed in the VOT/syllable duration space (Fig 3(b)). Here, the range of VOT for each voicing category is constant regardless of

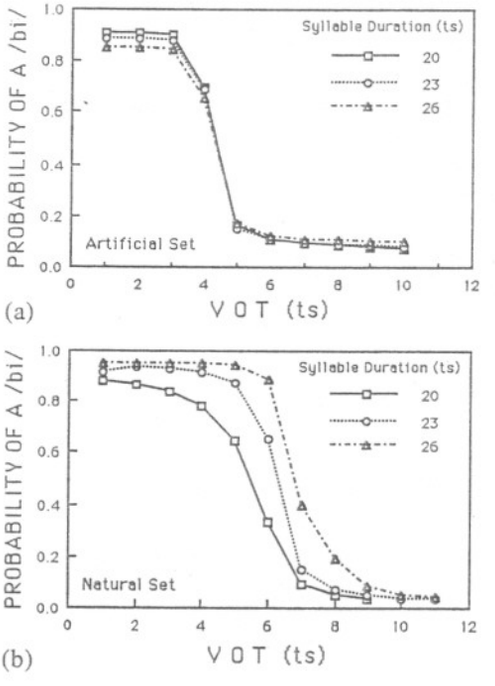

Figure 4: Probability functions for the /bi/-/pi/ stimuli used in the (a) natural set, and (b) artificial set. The results of only the 20-, 23- and 26-ts series are displayed here.

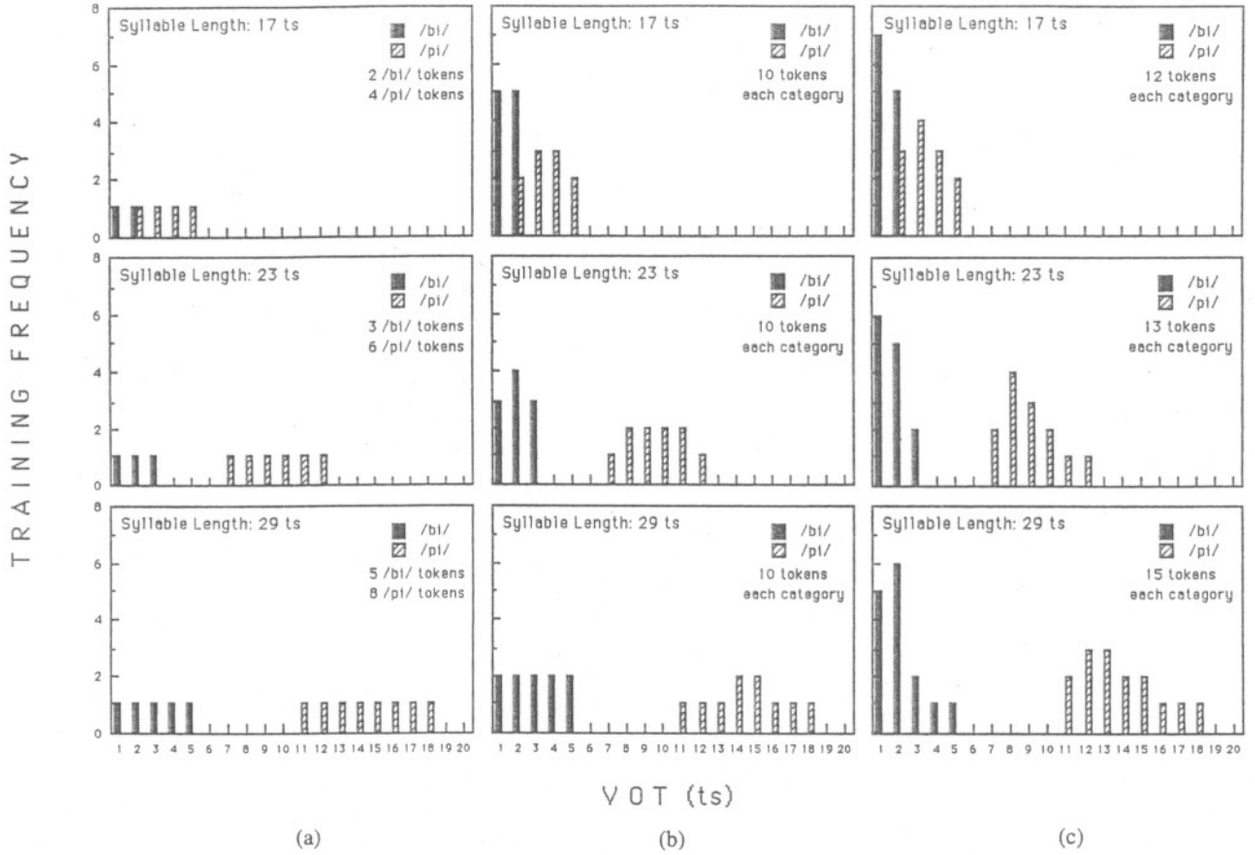

Figure 5: Training frequency for /bi/ and /pi/ tokens in (a) Scheme O, (b) Scheme A, and (c) Scheme B. Only the 17-, 23- and 29-ts series are shown here.

syllable duration. We predicted that the network would show a criterion shift only for the "natural" set. This would support our suggestion that the criterion shift found experimentally partly has its roots in the particular distribution of the two stop consonants.

During training, the network encountered each permissible VOT/syllable duration token once for every pass through the training set. During the test phase, the net was presented with the original training tokens as well as novel tokens whose VOT values straddle the category boundary (see Fig. 3).

Fig. 4 shows the activation values of the output unit representing the /b/ category obtained at syllable offset. These may represent the probability of a /bi/ as a function of the VOT value of the stimulus. Fig. 4(a) shows that no syllable duration effect on the voicing distinction was observed for the control set. Fig. 4(b) shows, by contrast, a systematic effect of syllable duration on the identification of the voiced and voiceless consonant for the "natural" training set. To quantify the boundary shift, we calculated the location of the /b-p/ phonetic boundary for each of the three syllable durations by fitting a regression line to the data and taking the boundary to be the stimulus value corresponding to the 0.5 probability. The boundary locations for the 20-, 23-, and 26-ts series were at VOT values of 5.359, 6.263, and 6.886 respectively, representing a substantial shift in the boundary location.

Effect of Training Frequency on Category Boundary

According to the learning account we are advocating, the frequency of individual speech tokens, not just their range of variation, may be expected to play an important role in determining phonetic judgements. In the previous simulation, each stimulus is presented equally often. Coupled with the fact that the voicing distribution locates the voiceless syllables over a wider range on the VOT continuum than the voiced syllables, this produces an unbalanced exposure between voiced and voiceless tokens in favour of the latter. Consider the 23-ts series (see Fig. 3(a) in conjunction with Fig. 5(a)). The voiced tokens are distributed over three points along the VOT scale while the voiceless tokens are fixed at six points on the same scale. If every such point is encountered once, this means the network encounters voiceless tokens twice as often as voiced tokens. This imbalance may be expected to affect categorization performance.

We therefore ran new simulations in which the exposure of voiced and voiceless tokens in each syllable duration

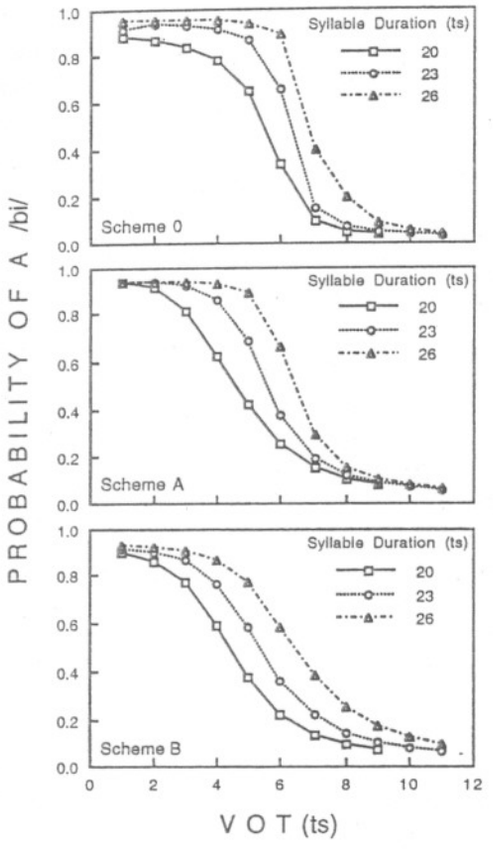

Figure 6: Probability functions for /bi/ showing the rate effect for three /bi/-/pi/ series.

series was held equal (Fig. 5(b)). Here, each category in each /bi/-/pi/ series is allowed 10 exposures in a single training cycle. The ten exposures are first divided equally between all the tokens within the permitted VOT range. Where the number 10 is not divisible by the number of VOT points on which the tokens are distributed, the remainder of the exposure is simply divided between each of the middle tokens of the category.

While this modified frequency scheme maintains an equal number of exposures of each category, it ignores the fact that some variants of a category occur more frequently. It seems likely that this will affect category prototype formation, and consequently the category boundary (Repp & Liberman, 1987).[2] We therefore ran a further set of simulations using an improved frequency scheme (Fig. 5(c)) which, in the absence of a more realistic appraisals of the frequency information delimiting the occurrence of members of the /b/-/p/ series, only loosely models the production data of Volaitis & Miller (1992). Notice that while there is no frequency bias between categories, within a category, some tokens are more frequent than others.

[2] We consider only category boundaries here, leaving prototypes to another paper.

Table 1: Voiced-voiceless boundary locations, in ts VOT, for each training scheme (O, A, B) and rate (20-, 23-, 26-ts)

	O	A	B
20 ts	5.359	4.681	4.432
23 ts	6.263	5.609	5.376
26 ts	6.886	6.498	6.497

After training, the network was tested on the same set of test stimuli as used previously. Results from the two training frequencies (the 'modified' type referred to as Scheme A and the 'improved' type as Scheme B) were compared in conjunction with the results obtained earlier in the preceding section (referred to as Scheme O). The number of iterations during training for the two schemes varies between 200 to 300.

Fig. 6 shows the activation of the unit representing a /bi/ as a function of VOT. To quantify the effects of syllable duration and frequency schemes, six new boundary locations arising from schemes A and B were computed using the same procedure as before. These boundary values, together with the three already calculated (due to scheme O), are shown in Table 1.

Consider first the boundary shift as a result of changes in syllable duration (Fig. 6; Table 1)). All the frequency schemes show a shift toward longer VOT as syllable duration increases. With the exception of the boundary shift between the 23- and 26-ts series under the O scheme, all other shifts are roughly equal (between 0.9 to 1.1 ts). Next consider the effect of varying the frequency scheme on the boundary location for each individual series (Table 1). Now the boundary is shifted in the opposite direction. Changing from scheme O to scheme A results in a larger boundary shift (mean 0.57 time-steps) than the change between A and B schemes (mean 0.16 ts). Also, the change from O to B for the longer series produces small shifts as compared to the shorter series.

While the shift as a consequence of changing rate is expected, the corresponding movement in the opposite direction due to training frequency is novel. To get an intuition for why the latter may have occurred, consider first the change in frequency schemes from O to A (Fig. 5). This transition is accompanied by a greater increase in exposure for the voiced tokens as compared to the voiceless tokens. Take, for example, the 23-ts series. In scheme O, the voiced tokens are exposed three times, but in A, this is increased to seven. In contrast, the voiceless tokens gain an increase of only four exposures over the same transition. There is thus a differential of three exposures in favour of the voiced category. A possible explanation for the observed shift in the boundary location therefore is that increase in exposure in favour of one category pulls the boundary towards that category.

The changes accompanying the transition from Scheme A to Scheme B are more difficult to quantify. The number of exposures have been increased in Scheme B but between-category differential is still maintained at zero (see Fig. 5(c)). The differential in exposure between members of the same category may be crucial. Table 1 suggests that biasing

the frequencies of members of the voiced and voiceless categories in the way we did (Fig. 5(c)) made the voiced category optimally more effective as a "boundary puller" than the voiceless category, particularly for the shorter series.

Conclusion

The results from this work have implications for spoken language processing and models of perception and categorization of human speech. A connectionist network can learn to show a systematic rate effect that can be traced to the network's sensitivity to the type and frequency of training stimuli. The factors that matter to the network may also matter to humans in fundamental ways, which raises the possibility that boundary locations need not be determined solely by listeners' capacity to discriminate auditorily but could be learnt from experience. More generally, this suggests that learning may play a more pervasive role in phonetic category formation than previously thought.

Acknowledgements

We are thankful to Randy Diehl, Keith Kluender and Joanne Miller for providing fruitful e-mail discussion which helped us clarify our ideas. 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. (1993a). Processing time-warped sequences using recurrent neural networks: Modelling rate-dependent factors in speech perception. In *Proceedings of the 15th Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Abu-Bakar, M., & Chater, N. (1993b). Studying the effects of speaking rate and syllable structure on phonetic perception using recurrent neural networks. *Irish Journal of Psychology*, 14, 410-425.

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

Diehl, R. L. (1981). Feature detectors for speech: A critical reappraisal. *Psychological Bulletin*, 89, 1-18.

Diehl, R. L., & Kluender, K. R. (1989). On the objects of speech perception. *Ecological Psychology*, 1, 121-144.

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*, 57, 462-469.

Elman, J. L. (1990). Representation and structure in connectionist models. In G. T. M. Altmann (Ed.), *Cognitive Models of Speech Processing: Psycholinguistic and Computational Perspectives.* Cambridge: MIT Press.

Greenberg, S. R. (1988). The ear as a speech analyzer. *Journal of Phonetics,* 16, 139-150.

Kluender, K. R., Diehl, R. L., & Killeen, P. R. (1987). Japanese quail can learn phonetic categories. *Science*, 237, 1195-1197.

Kuhl, P. K. (1988). Auditory perception and the evolution of speech. *Human Evolution,* 3, 19-43.

Maskara, A., & Noetzel, A. (1992). Forced simple recurrent neural networks and grammatical inference. In *Proceedings of the 14th Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Miller, J. L., & Liberman, A. M. (1979). Some effects of later-occurring information on the perception of stop consonant and semivowel. *Perception and 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 (Suppl. 2), SP11.

Repp, B. R., & Liberman, A. M. (1987). Phonetic category boundaries are flexible. In S. Harnad (Ed.), *Categorical Perception: The Groundwork of Cognition*. New York: Cambridge University Press.

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

Samuel, A. G. (1982). Phonetic prototypes. *Perception and Psychophysics*, 31, 307-314.

Shillcock, R., Lindsey, G., Levy, J., & Chater, N. (1992). A phonologically motivated input representation for the modelling of auditory word perception in continuous speech. In *Proceedings of the 14th Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Lawrence Erlbaum Associates

Stevens, E. B., Kuhl, P. K., & Padden, D. M. (1983). Macaques show context effects in speech perception. *Journal of the Acoustical Society of America,* 84 (Suppl. 1), S77.

van Camp, D., & Plate, T. (1993). Xerion Neural Network Simulator. Department 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 interrnal structure of voicing categories. *Journal of the Acoustical Society of America*, 84, 723-735.