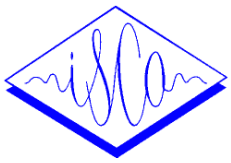

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

PHONETIC PROTOTYPES: MODELLING THE EFFECTS OF SPEAKING RATE ON THE INTERNAL STRUCTURE OF A VOICELESS CATEGORY USING RECURRENT NEURAL NETWORKS

*Mukhlis Abu-Bakar** *and Nick Chater*+*

*Departments of Linguistics & Psychology, University of Wales, Bangor, Gwynedd LL57 2DG, U.K.
+Department of Psychology, University of Edinburgh, 7 George Square, Edinburgh EH8 9JZ, U.K.

ABSTRACT

This is the last in a series of studies that attempts to account for rate effects in phonetic perception using a recurrent neural network. We present new findings that extend this research, emphasising particularly the effects of syllable duration on the internal structure of stimuli representing the voiceless /p/ category. In a series of simulations, network performance reveals a systematic rate effect that closely parallels that found in studies of human categorization. First, the internal structure of the voiceless category undergoes extensive alteration with changes in syllable duration. Second, in a selective adaptation type of experiment, the maximally effective adaptor shifts toward a longer VOT as syllable duration increases. The performance can be traced to the network's sensitivity to the training frequency of the individual stimulus, 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 are complex. For example, Miller & Liberman [1] examined the effect of speaking rate and syllable structure on the stop-semivowel distinction, specified by the formant transition duration. When syllable duration is increased by lengthening the vowel, 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 "articulatory" rate, defined in terms of syllable duration and the number of phonetic segments in the syllable. But an account that is dependent on listeners' sensitivity to variation in articulatory rate cannot explain human subjects' categorization of analogous nonspeech stimuli [2] nor nonhuman subjects' discrimination of speech stimuli [3]. 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 [4]. When speaking rate is varied, those changes that occur closest to the target segment will affect its perception most. In the case of /bla/ and /pla/, the auditory model predicts that varying the /l/ has 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 [5].

We have trained a recurrent neural network on rate-varying speech-like stimuli and compared its performance with these divergent predictions [6]. The results showed that the network behaved in line with the auditory account, suggesting that such an account could be instantiated computationally in a network. 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 a previous paper [7], we elaborated 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 showed how frequency of individual speech tokens as well as their range of variation played a role in determining phonetic judgements, as applied to rate varying stimuli. In this paper, we look at how a learning mechanism, such as our model, instantiates the effect of rate on the internal structure of phonetic categories.

DESCRIPTION OF THE MODEL

Recurrent neural networks [8] are very attractive for problems concerned with speech processing because they are suited to processing sequential material. Here, the network is trained to classify input sequences into a small number of categories corresponding to different syllables. For each sequence, one input pattern at a time was presented, 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 [9, 10, 11]. The network was trained by recurrent backpropagation [12] using conjugate gradient descent and implemented on the Xerion simulator [13]. The number of input, hidden and output units was 30, 60 and 33 respectively.

Training stimuli were based on a two-formant syllable with an initial period of formant transitions followed by a steady-state (vowel) (Fig. 2). A 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. 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 units of F1 and F2 and activating the F2 noise

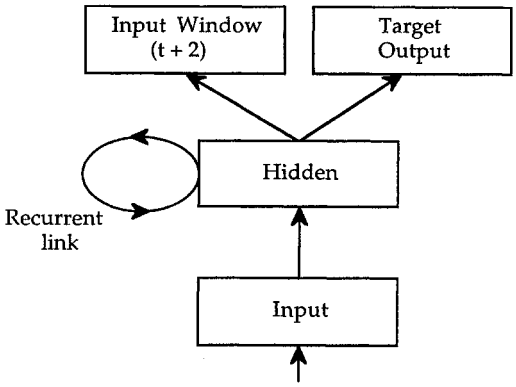

Fig.1 - The recurrent network used in the simulations. It is unfolded during training for as many time-steps as required to accomodate the longest training stimuli.

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.

The network's task is continuously to rate the probability of a stimulus belonging to a particular category as the stimulus unfolds. It is possible for the net to identify the consonant early in the sequence for some syllables, particularly those whose VOT values are unambiguous [6, 9]. But a thorough evaluation of category goodness is possible only when the net has scanned the entire 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.

INTERNAL STRUCTURE OF PHONETIC CATEGORIES

In an earlier study [7], we looked 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 duration, shown in Fig. 3, was adapted from the production patterns of /b/-/p/ tokens studied by Volaitis & Miller [14]. From the point of view of a network learning this distribution, however, the VOT scale specifies only two categories. While this has the desired effect of forcing the network to define the boundary between the categories, it leaves the net ignorant of the limits on the VOT values which the two categories may assume. In fact, the further away the stimuli are from the boundary the higher the ratings they get from the network. This is misleading but hardly surprising since

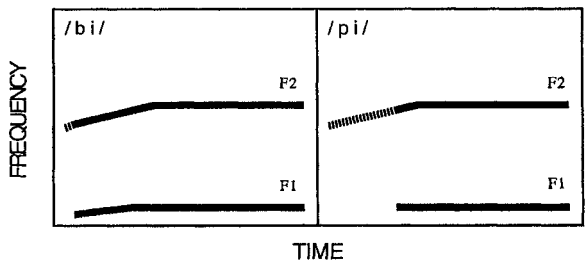

Fig.2 - Schematic representations of the formant motions of the endpoint stimuli corresponding to /bi/ (left) and /pi/ (right). Each representation consists of an interval of aspiration (striped line), followed by onset of voicing (dark lines).

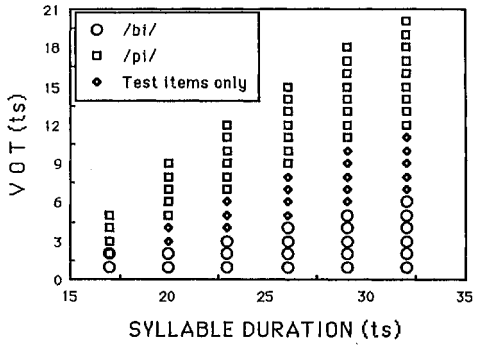

Fig.3 - Location of /bi/-/pi/ tokens on the VOT and syllable duration space, in time-step. Distribution of tokens used as tests only are also shown on each panel.

the net is trained to recognize nothing except /b/ or /p/. Thus we should expect the probability functions to be monotonic. On the other hand, if we explicitly teach the network to classify stimuli with extreme VOTs as something else, the /b/ and /p/ probability functions may become nonmonotonic, consistent with structures associated with phonetic categories. That is, as the stimulus moves away from the boundary region, it becomes a better exemplar of the category in question, but only to a certain limit. Beyond some point along the continuum, the stimuli should again sound like poorer members of the category [15, 16].

In this study, we focus on the /p/ category. Since this category is already constrained by the voiced category at the lower end of the VOT continuum, we needed only to introduce constraints at the other end by adding stimuli with extended VOT values which the net was taught to classify as neither /b/ nor /p/ but an exaggerated version of /p/ which we labeled as /p*/. Training proceeded in two stages:

Phase 1. The frequency scheme as represented by Fig.4(a) was used. The network was trained on 350 passes through the training stimuli, at which point the /bi/ and /pi/ stimuli with the minimum and maximum VOT, respectively, have gained activation levels of above 0.75.

Phase 2. The net was trained on the same /bi/ and /pi/ stimuli but with a reduction in the frequency distribution (see Fig.4(b)). The /pi*/ stimuli was also added to the training set with a frequency distribution represented in Fig.4(b). Training proceeded for a further 200 passes.

This staged learning strategy was developed in response to results of earlier pilot work where it was found that the network was unable to learn the task when the full set of stimuli from all three categories were presented from the beginning of training: The network confused /bi/ for /pi/. But when the network was permitted to focus on the /bi/ and /pi/ stimuli first, it was able to move on successfully to the "strange" stimuli, at the same time modifying its earlier responses to some of the /pi/ stimuli, particularly those near the higher end of the /p/ range. Thus, the earlier training constrains later training in a very useful way [17]. Additionally, phase 2 was initially unsuccessful when the same frequency distribution used in phase 1 was retained: The activation of /pi/ remained at a high peak despite being presented with the /pi*/ stimuli. Presumably, the earlier frequency distribution was too strongly in favour of /pi/ that the contribution of /pi*/ was lost.

We consider if the internal structure so constructed is altered by changes in syllable duration. That is, whether a change in syllable duration will modify the range of stimuli identified as members of a phonetic category, consistent with the distribution and frequency of the

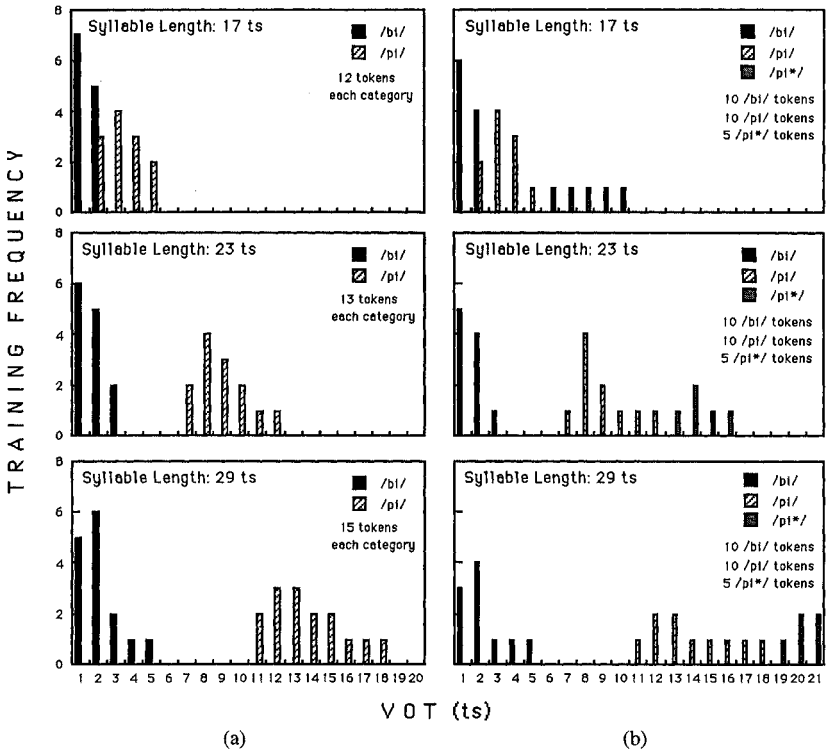

Fig.4 - Training frequency for (a) Phase 1, and (b) Phase 2 training. Only the 17-, 23- and 29-ts series are shown here.

Table I - Peak value, P, and lower, L, and upper, U, limits of the best-exemplar range, in ts VOT, for each rate (20 ts, 23 ts, 26 ts) series

	Best-exemplar limits		
	P	L	U
20 ts	9.869	7.418	12.326
23 ts	10.121	7.835	14.458
26 ts	10.991	8.392	17.011

training stimuli. The results for the 20-, 23-, and 26-ts series are displayed in Fig. 5. As expected, the 23-ts function is broader than the 20-ts function but narrower than the 26-ts function. Moreover, the function for the longer series was displaced toward longer VOT values, as was the location of the most representative member. This pattern is in accord with experimental findings [14, 16].

To quantify this change, we calculate the location of the peak and the best exemplar range for each function, using a procedure similar to that introduced by Miller and Volaitis [16]. We first locate the stimulus that has the highest probability (i.e., the peak of the function, obtained by linear interpolation) and then calculate the range of stimuli that have probabilities within 90% of the maximal probability. This is designated as the best-exemplar range. For instance, if the highest probability is 0.8, and this is the probability of a stimulus located at 10-ts VOT, then the best-exemplar range is between the two VOT values (also obtained by linear interpolation) that correspond to probabilities of 0.72 (90% of 0.8) on either side of 10-ts VOT. The VOT values delimiting the best-exemplar range on the lower (L) and upper (U) sides of the maximal score, and the peak values (P), are presented in Table I. These

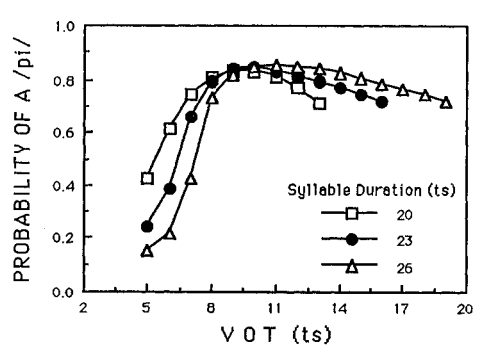

Fig.5 - Probability functions for /pi/ at the end of Phase 2 for three /bi/-/pi/-/pi/ series*

results show that the "prototypical" member and the range of best exemplars alter as a result of changes in syllable duration, in line with experimental data.

SELECTIVE ADAPTATION

Changes in internal phonetic structure as a function of speaking rate have also been reported in studies using a selective adaptation technique. In one such study [18], it was reported that increasing the syllable duration produced a displacement as well as expansion of the entire adaptation function, at the same time altering the location of the maximal adaptation effect. These results reveal differences in the effectiveness of adaptors as a function of their distance from the category boundary. In our earlier study [7], the model was shown to be sensitive to global frequency changes which resulted in a shift in the boundary that separates /b/ and /p/. However, this is only indirect evidence of the networks' capacity to simulate adaptation effects. A more direct method would be to alter the training frequency of individual members of the series one at a time and then measure the effect this has on the /b/-/p/ boundary.

In this study, we apply a selective adaptation type of technique to test the model's effectiveness in adjusting for changes in the adapting stimulus. The network weights from the preceding study served as the starting point for the simulations. The aim is to determine whether information about internal structure as distributed in the weights of the connections allows for graded adaptation effects that are also sensitive to changes in syllable duration. For simplicity, the study focussed only on the 20- and 26-ts series. Adapting stimuli were chosen from the two series such that they spanned a range of values from near the /b/-/p/ boundary to the most extreme stimulus created. The scheme represented in Fig. 4(a) was modified for each adapting stimulus so that the frequency of the adapting stimulus was increased by 5 exposures per cycle. On every cycle, the network was exposed to the adapting stimulus 5 times more than in the condition

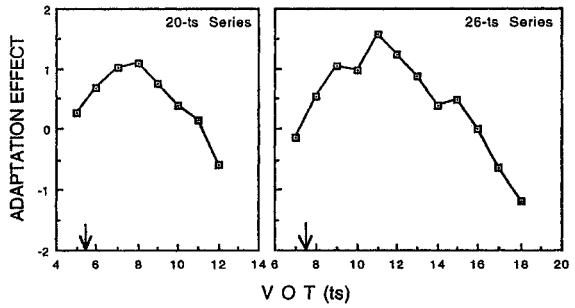

Fig.6 - Squared difference adaptation on two /bi/-/pi/-/pi/ series as a function of the VOT of the adapting stimulus for the 20-ts series (left panel), and the 26-ts series (right panel). The arrows in the two panels point to the location of the category boundary in conditions of no adaptation.*

where there is no adaptation. After 15 passes, the weights were frozen at their final values, the probability function of the /p/ category recorded, and the /b/-/p/ boundary calculated similarly as before. This boundary was taken as the measure of the adaptation performance. The procedure was repeated for all the stimulus adaptors. In the condition where there is no adaptation, the net was simply left to run for 15 iterations without making any modifications to the training scheme. To obtain a measure of the degree of adaptation, the non-adaptation boundary value was squared and the product subtracted from the square of the adaptation boundary value. Thus each simulation yielded a single measure of adaptation, a difference boundary value in squared units. A positive difference value indicates a shift in the boundary toward the /p/ category while a negative difference value signal a shift in the opposite direction. Of interest is the relative magnitude of these values.

In Fig. 6 are plotted the data from the two /bi-pi/ series. Each graph displays the adaptation effect (the difference in the squares of boundary values) as a function of the VOT value of the adapting stimulus. The location of the non-adaptation category boundary is shown by the arrow. Consider first the data from the 20-ts series. As the VOT value of the adaptor increases from 5 to 8 ts, there is an increase in the magnitude of adaptation. However, as the VOT value of the adaptor increases beyond 8 ts, there is a decline in the magnitude of adaptation, with no adaptation occurring for the most extreme adaptor. The general pattern for the 26-ts series is similar to that found for the 20-ts series in that the adaptation function shows a nonmonotonic curve. However, the increase in syllable duration has moved the location of the maximal adaptation effect from 8 ts VOT to 11 ts VOT. Moreover, the adaptation function for the 26-ts series is broader than that of the shorter series. These patterns are strikingly like those found for human listeners; that is, not only is the adaptation function nonmonotonic, increase in syllable duration has the effect of displacing and inflating the entire adaptation function towards longer VOTs. Importantly, this performance can be traced to the network's sensitivity to the training frequency of the individual stimulus.

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 suggesting that learning may play a more pervasive role in phonetic category formation than previously thought.

REFERENCES

[1] 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.

[2] 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.

[3] Stevens, E. B., Kuhl, P. K., & Padden, D. M. (1988). Macaques show context effects in speech perception. *Journal of the Acoustical Society of America*, 84 (Suppl. 1), S77.

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

[5] 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.

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

[7] Abu-Bakar, M., & Chater, N. (1994). Distribution and frequency: Modelling the effects of speaking rate on category boundaries using recurrent neural networks. *Proc. 16th Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Lawrence Erlbaum Associates.

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

[9] Abu-Bakar, M., & Chater, N. (1993). Processing time-warped sequences using recurrent neural networks: Modelling rate-dependent factors in speech perception. *Proc. 15th Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Lawrence Erlbaum Associates.

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

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

[12] 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 Microstructures of Cognition, Vol. 1*. Cambridge: MIT Press, .

[13] van Camp, D., & Plate, T. (1993). Xerion Neural Network Simulator. Department of Computer Science, University of Toronto.

[14] 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.

[15] Oden, G. C., & Massaro, D. W. (1978). Integration of featural information in speech perception. *Psychological Review*, 85, 172-191.

[16] Miller, J. L., & Volaitis, L. E. (1989). Effects of speaking rate on the perceived internal structure of phonetic categories. *Perception and Psychophysics*, 25, 457-465.

[17] Elman, J. L. (1991). Incremental learning or the importance of starting small. Tech. Report 9101, Center for Research on Language, University of California, San Diego.

[18] Miller, J. L., Connine, C. M., Schermer, T. M., & Kluender, K. R. (1983). A possible auditory basis for internal structure of phonetic categories. *Journal of the Acoustical Society of America*, 73, 2124-2133.