Proceedings of the Institute of Acoustics

THE ROLE OF HARD-WIRING IN THE PERCEPTION OF SPEECH

S. J. Mashari (1) & M. J. Pont (2)

(1) University of Sheffield, Department of Computer Science, Mappin Street, Sheffield S1 4DU, UK.

(2) University of Leicester, Department of Engineering, Leicester, LE4 7RH, UK.

1. INTRODUCTION

Categorical Perception (CP) of speech stimuli has been studied extensively by means of behavioural experiments (see Repp, 1984). In particular, a large number of such studies have focused on the perception of voicing in initial stop-consonants (Lisker and Abramson, 1970). Liberman et al. (1958) demonstrated that adult listeners perceive such sounds in an almost categorical manner. This ability has since been shown by speakers of many different languages, and Voice Onset Time (VOT) has proved to be an important cue to this contrast (Lisker and Abramson, 1970).

An important question arising from the above experiments was whether the ability to categorise was innate or learned. This question remained largely unanswered until the discovery of new experimental techniques made it possible to use human infants as subjects (see Kuhl, 1987). The results showed that infants as young as one month old demonstrate quasi-categorical responses to speech sounds differing in VOT similar to those shown by adult English speakers (Eimas et al., 1971). These experiments have been widely taken to imply that categorical perception is an innate ability.

This paper is part of a series of studies (e.g. Pont and Damper, 1991; Mashari and Pont, in press) which aim to uncover the neural mechanisms underlying the observed categorical responses to VOT stimuli. Many studies conducted in this area have used stimuli drawn from a synthetic VOT series produced by Lisker and Abramson (1970), a fact that motivated us recently to examine these stimuli more closely (Mashari and Pont, in press). The Lisker and Abramson stimuli are intended to mimic the multi-dimensional variations real speech, and consequently the "VOT" variations are realised by varying the onset time of the first formant relative to the second and the third, and by exciting these higher formants when the first is not present. We have shown (Mashari and Pont, in press) that a simple analysis of these stimuli (specifically calculating Linear Prediction [LPC] coefficients for each stimulus and using Dynamic Time Warping [DTW] to compare the LPCs for the end-point stimuli with the intermediate values for each series) that

categorical effects are evident in the stimuli themselves, with category boundaries for each series similar to those seen in behavioural studies. These results suggested that certain aspects of the behavioural response may be artefacts of the stimuli used.

This paper is a report on the second part of this ongoing study, using tone-onset time stimuli to approximate VOT stimuli. We begin by describing the stimuli used, and confirming their "linear" nature. We then use a computer model of the auditory nervous system to investigate the possible neural basis for the CP of these stimuli.

2. TONE-ONSET-TIME STIMULI

Having confirmed the complex nature of VOT stimuli (Mashari and Pont, in press), it was necessary to use here a simpler, but related, set of stimuli to continue this investigation. Many previous behavioural studies have shown common features in the perception of synthetic speech sounds differing in VOT and the perception of tone-onset time (TOT) stimuli (e.g. Pisoni, 1977). TOT stimuli consists of two tones, with frequencies approximating the formant frequencies of natural vowels. The relative onset time of the tones is varied to generate stimuli with different TOT values in a manner roughly analogous to the movement of the first-formant in VOT stimuli (Pisoni, 1977). However - unlike a VOT series - a TOT stimulus series can be made inherently "linear" if the tone onset time is varied equal steps.

In the original experiments performed by Pisoni (1977), it was found that subjects generally divided the TOT continuum into two distinct classes, in a manner which resembled closely the VOT experimental findings. Based on his results, Pisoni suggested that because of the importance of these time differences, linguistic systems have incorporated this cue to distinguish between sets of stop consonants (Pisoni, 1977).

3. ANALYSIS OF "RAW" TOT STIMULI

We began the present study by reproducing a series of TOT stimuli first used by Pisoni (1977). The stimuli consist of two tones one at 500 Hz and one at 1500 Hz, with TOT differences produced as described above (Section 2). The result was a series of nine stimuli, from 0 to 80 ms TOT, in 10 ms steps. For completeness, we subjected these stimuli to the same analysis previously used with the VOT sounds, again mimicking the two analysis techniques used in behavioural studies - discrimination and labelling tests.

In labelling experiments, listeners are asked to classify sounds drawn from a continuum as voiced or voiceless: for example, in the case of labial VOT stimuli, subjects are asked to judge if a sound heard is a /ba/ or a /pa/ sound. We mimicked these experiments by first calculating LPC coefficients (Mashari and Pont, in press) for each TOT stimulus, producing nine different templates. We then calculated, using DTW, the eight cost ("labelling") function values when each templates was compared with the last template (80 ms TOT). Figure 1 shows the results of this comparison: note that the cost (zero) between the endpoint stimulus and itself is not shown. As expected, this "labelling" response is perfectly linear.

The second method of conducting the behavioural experiments is based on another important aspect of CP, the ability of subjects to discriminate between different sounds in the continuum. In so-called discrimination experiments, subjects are usually asked to perform an ABX test. Discrimination between sounds drawn from the "same" category is found to be almost at chance levels, while sounds drawn from the two sides of the (labelling) boundary are discriminated well.

We performed a simple discrimination test using LPC and DTW to calculate cost functions for adjacent pairs of TOT stimuli: thus, we compared the 0 ms and 10 ms stimuli, the 10 ms and 20 ms stimuli, etc. The bars in Figure 1 show the results: there is no discrimination peak and the comparison of all pairs of stimuli produced exactly the same result.

The results of these experiments were precisely as expected. We next went on to investigate how this linear series of sounds would be represented in our auditory model. The model is described briefly in the next section.

4. THE AUDITORY NERVE AND COCHLEAR NUCLEUS MODEL

We are interested in finding the lowest level in the auditory system at which categorical responses can be seen: at this stage, we are particularly interested in examining units in the auditory nerve and cochlear nucleus. In this study, the responses of these two levels of auditory processing were produced using a comprehensive computational model of afferent neural processing from the cochlea to the dorsal acoustic stria. Here, we very briefly describe this model; a comprehensive description is given elsewhere (Pont & Damper, 1991).

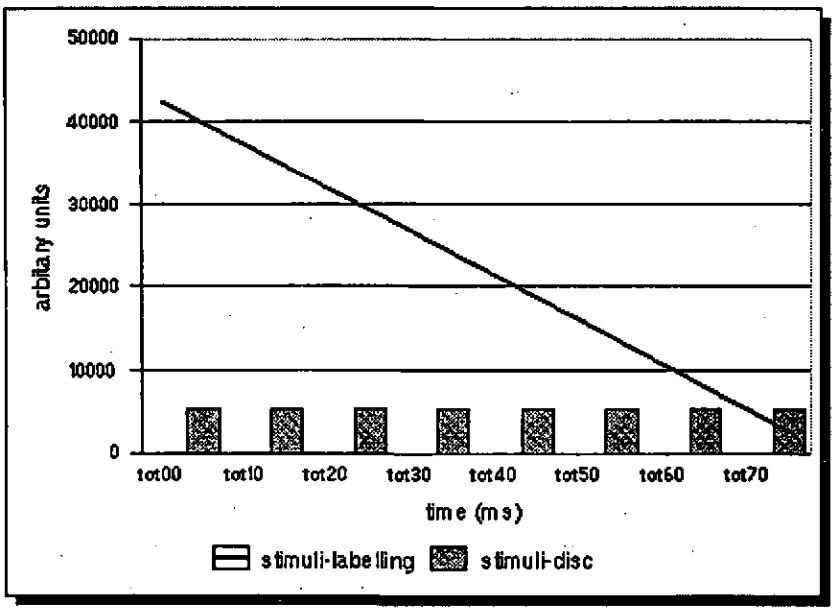

Figure 1. Labelling and discrimination responses to TOT stimuli

The computations are performed in two stages. The first stage simulates the cochlea and auditory nerve; and the second, the dorsal cochlear nucleus. The complete model consists of an array artificial neuron units whose properties and interconnections follows closely those reported in recent anatomical and physiological studies of the cat auditory nerve and dorsal cochlear nucleus. Input to the model is derived from a cochlear filter bank, spanning a frequency range from 100 to 5000 Hz. The output of the model consists of a binary stream recording, for each of the nerve cells modelled, the action potentials produced.

5. ANALYSIS OF THE MODEL RESPONSES

Here we used the model described in the last section to simulate responses at the AN and DCN level to the TOT stimuli. The responses were then (for consistency) analysed in a manner similar to that used for the stimuli themselves. Specifically, templates were produced for the AN

and DCN responses to each TOT value in turn, and "discrimination" and "labelling" responses were calculated using DTW.

The resultant AN responses curves are shown in Figure 2. The bars again represent the cost function values when the two adjacent TOT points are compared, while the curve shows the cost function values if all the templates are compared with one of the end-points (80 ms TOT).

Looking at the "discrimination" responses, there is a sharp peak at 20 to 30 ms TOT, very similar to that found by Pisoni (1977) in his behavioural study. The "labelling" responses also show a boundary at this point.

Figure 3 shows the results of the analysis at the DCN level. Again the bars and the curve represent "labelling" and "discrimination" responses respectively. From the figure, there is little evidence that the DCN plays an important part in Categorical perception of TOT stimuli: in contrast to the AN representation, there are no sharp discriminatory peaks and the labelling response is almost linear.

6. CONCLUSIONS

The studies described in this paper are part of a continuing investigation aimed at establishing which part of the auditory system might be responsible for the sharp categorical responses observed when synthetic stop-consonants differing in VOT are presented to subjects in behavioural experiments. In the experiments described here, we used a set of TOT stimuli which we analysed in a "raw" state, and at the AN and DCN levels in a computer model. The stimuli themselves were shown, as expected, to be completely "linear" in nature, while the responses of the computational model were shown to be clearly categorical, particularly at the level of the auditory nerve.

Our results suggest that the representation of speech at the level of the auditory nerve may be important in the generation of categorical responses in the perception of TOT, and to VOT stimuli. We are presently involved in further studies aimed at clarifying this issue.

7. ACKNOWLEDGEMENTS

The VOT stimuli used in the study described here were produced at Haskins laboratories, New

Haven, Connecticut, USA with the assistance from NICHD contract NO1-HD-5-2910.

8. REFERENCES

Eimas, P. D.; Siqueland, E. R.; Jusczyk, P. & Vigorito, J. (1971), "Speech perception in infants", Science, 171, 303 -306.

Kuhl, P.K., (1987), "The special-mechanisms debate in speech research: Categorisation tests on animals and infants", IN, Harnad, S.R, (ed), Categorical Perception: The Groundwork of Cognition, Cambridge.

Liberman, A. M., DeLattre, P. C., & Cooper, F. S.(1958), "Some cues for the distinction between voiced and voiceless stops in initial position", Lang. Speech 1, 153-167.

Lisker, L. & Abramson, A. S. (1970), "The voicing dimensions; some experiments in comparative phonetics", Proceedings of the sixth International congress on phonetic sciences, 1967, 563 - 567.

Mashari, S.J. & Pont, M.J. (in press), "Exploring the relationship between physiology and behaviour in the perception of VOT and TOT stimuli", British Journal of Audiology.

Pisoni, D. B. (1977), "Identification and discrimination of the relative onset time of two component tones: implications for perception of voicing in stops", Journal of the Acoustical Society of America, 61, 1352 - 1361.

Pont, M. J. & Damper, R. I. (1991), "A computational model of afferent neural processing from the cochlea to the dorsal acoustic stria", Journal of the Acoustical society of America, 89(3), 1213.

Repp, B.H., (1984), "Categorical perception: issues, methods, findings", IN, Lass, N.J. (ed), Speech and Language: Advances in Basic Research and Practice, Vol. 10, 243-335.

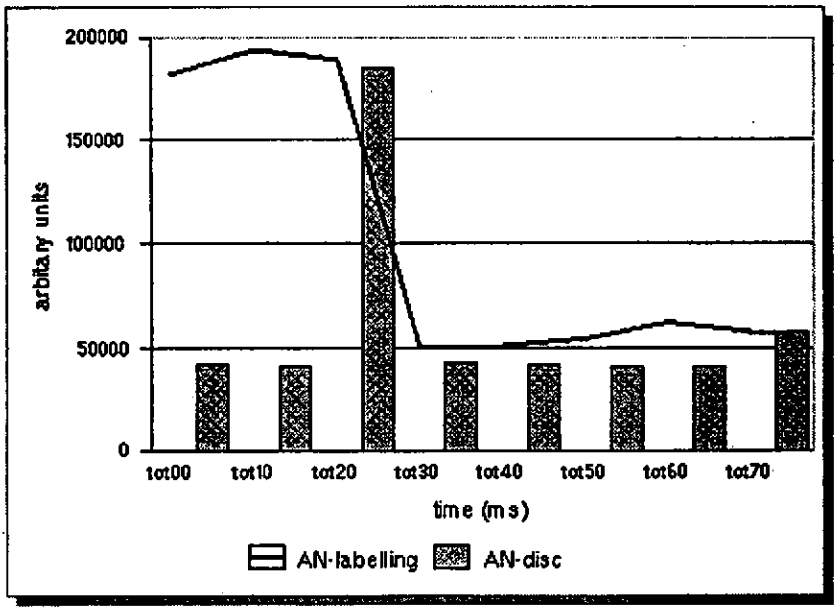

Figure 2. Labelling and discrimination responses to TOT stimuli at AN level

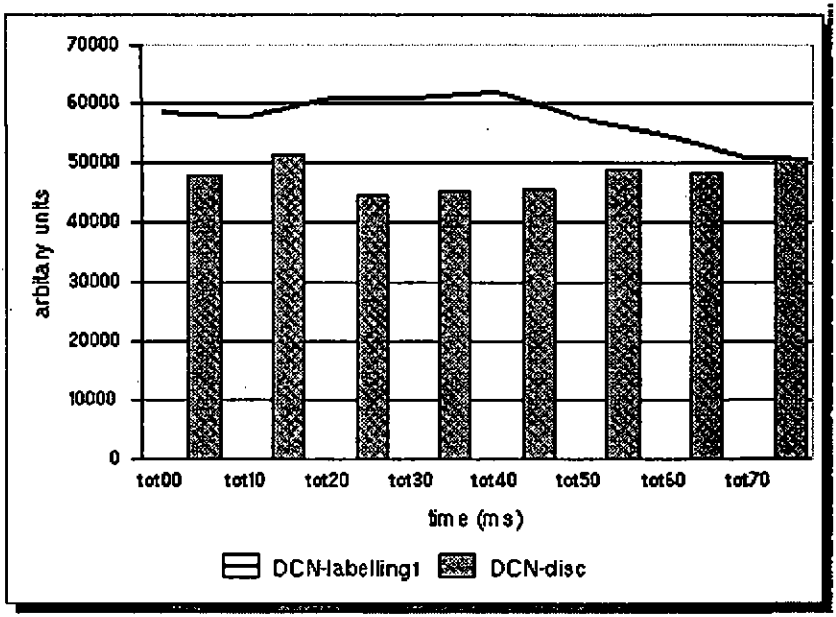

Figure 3. Labelling and discrimination responses to TOT stimuli at DCN level