AUTOMATIC DISCOVERY OF FEATURES UNDERLYING THE PERCEPTION OF VOICING

R.I. Damper

Image, Speech and Intelligent Systems (ISIS) Research Group,
Department of Electronics and Computer Science,
University of Southampton,
Southampton SO17 1BJ, UK.

ABSTRACT

Responses of both human and animal listeners to synthetic stop-consonant/vowel stimuli in which voice-onset time (VOT) is uniformly varied are known to be 'categorical' but an explanation of this phenomenon remains elusive. A 'composite' model consisting of a physiologically-realistic auditory model feeding its patterns of neural firing to an artificial neural network is described. When trained by (supervised) error back-propagation on the extreme, endpoints of the VOT continuum, the composite model is capable of reproducing closely listeners' behaviour in classical categorical-perception (CP) studies. However, whether the model also reproduces the so-called boundary-shift phenomenon – whereby the phoneme boundary moves with place of articulation – apparently depends upon precise details of the auditory model and so, by implication, upon subtle aspects of peripheral auditory processing. A first attempt at unsupervised training has been unsuccessful: the likely reason for this is outlined. It is anticipated that future work comparing the model's responses for unsupervised versus supervised training will help to elucidate the mechanisms of categorical perception.

1. INTRODUCTION

Few topics in speech science have generated as much debate as the phenomenon of categorical perception (CP), whereby stimuli from an auditory 'continuum' are perceived as non-continuous. That is, stimuli are classified as belonging to one category or another, with a sharp boundary between them. Further, discrimination between classes is much better than discrimination within a class. To some, this categorisation into discrete classes lies at the very heart of linguistic decoding in speech perception – see Repp's authoritative review [1]. To others, however, CP studies are misguided and uninformative – see Crowder's withering attack [2].

Our approach is to build a 'composite' computational model of auditory processing and then to explore its ability to mimic the results of key perceptual studies in speech CP. We use detailed physiological and anatomical knowledge to produce a biologically-realistic model of the auditory periphery, the outputs from which are fed to an artificial neural network (ANN) which models higher-level auditory function in much less detail. Then, by manipulating the model, we try to identify those components that are essential to the observed behaviour. Further, by employing different learning paradigms, we can vary the function of the ANN from classification to feature detection – the discovery of regularities in the input patterns [3]. By this latter means, we can hope to discover automatically the features underlying the perception of voicing.

In the remainder of this paper, we briefly outline the key experimental results in the study of speech CP and the stimuli used in this work (Section 2). Section 3 describes the composite neural model, before detailing the results of processing the stimuli by the model (Section 4), discussing the results and concluding (Section 5).

2. CATEGORICAL PERCEPTION OF VOT

The voiced/unvoiced distinction is fundamental to speech communication, playing a major contrastive rôle in all languages. As such, it has received much attention in studies of speech perception. In early work, Liberman *et al* [4] studied the perception of voicing in syllable-initial stop consonants as voice-onset time (VOT) was varied and showed it to be 'categorical'. That is, perception changes abruptly from 'voiced' to 'unvoiced' as VOT is increased uniformly, and discrimination is far better between categories than within a category. Hence, labelling functions are non-uniform and discrimination functions are non-monotonic. Intriguingly, such categorical behaviour is also found for non-human listeners [5], indicating that the underlying mechanisms are not specific – or 'special' – to speech.

Figure 1 shows labelling curves obtained by Kuhl and Miller [5] for English speakers and for chinchillas in response to bilabial (/ba-pa/), alveolar (/da-ta/) and velar (/ga-ka/) stimuli in which VOT was varied. Taking the 50% points as the boundaries between voiced and unvoiced categories, there is a *phoneme boundary-shift effect* [6] with place of articulation. Also, the chinchillas exhibit boundary values not significantly different from the humans (although the curves are less steep).

Various explanations have been put forward for the observed categorisation including articulatory ('motor the-

0-7803-2431-5/95 $4.00

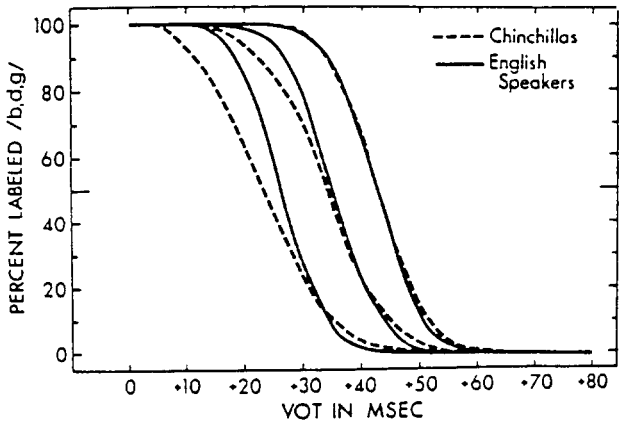

Figure 1: Mean labelling functions for (English speaking) human listeners and chinchillas obtained with (left to right) bilabial, alveolar and velar synthetic stimuli: from Kuhl and Miller [5].

ory'), auditory and learning hypotheses. As cogently argued by Rosen and Howell [7], however, none of them adequately explain all the data.

In this paper, we use connectionist modelling to gain insight into the issues involved in the controversy. The model's inputs are the stimuli initially synthesised by Abramson and Lisker [8] which have been widely employed in studies of speech CP. These are a series of bilabial, alveolar and velar /Ca/ syllables, with VOT varying from 0 ms to 80 ms in 10 ms steps. Note that in order to achieve perceptually acceptable tokens, some adjustments to parameters other than VOT were made during synthesis. Before commencing this work, labelling functions like those in Fig. 1 were obtained for 5 adult listeners to check the validity of the stimuli.

3. A COMPOSITE NEURAL MODEL

Ideally, any computational model of auditory processing should simulate all necessary details of neural function and anatomy at an appropriate level of abstraction. Unfortunately, we have neither the neurobiological knowledge nor the computer power to do so for the complete auditory system. Sufficient is known of peripheral function, however, to be able to construct models which mimic auditory nerve firing patterns well. Here, we use the 'P-D' model of Pont and Damper [9].

Figure 2 shows typical output ('neurogram') from this model to the alveolar stimulus (/da/) with 0 ms VOT. A dot represents the firing of an auditory nerve fibre (a 'spike'), with the horizontal axis indicating the time of firing and the vertical axis indicating the centre frequency (CF) of the fibre by an index in the range 1..128. All stimuli were applied at time $t = 0$ at a simulated level of 65 dB SPL. Activity before $t = 0$ is spontaneous, as is that in channels with CF index 1..8 (for reasons to do with the bandwidth of our auditory filters at low frequency). Fuller details are given in [9] and [10].

Artificial neural networks represent, in some sense, an

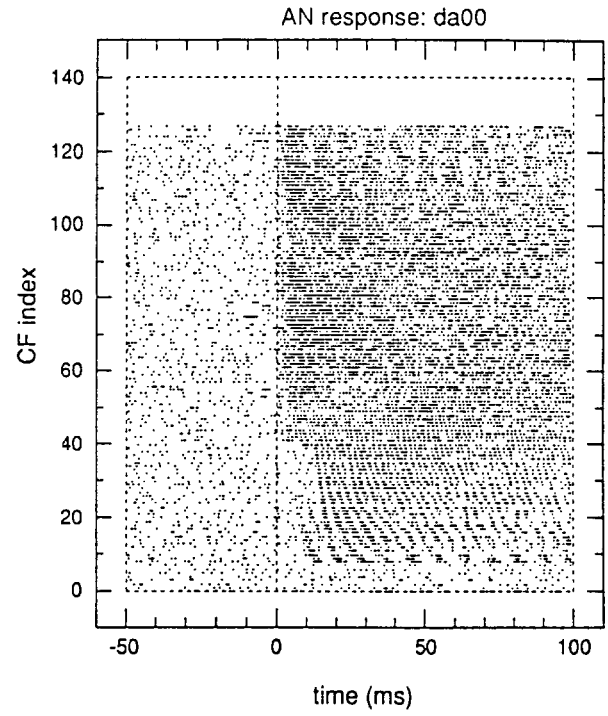

Figure 2: Neurogram response of the P-D auditory model to alveolar stimulus (/da/) with 0 ms VOT.

attempt to model neurobiological function at a high level of abstraction. This leads to the idea of a 'composite' neural model in which a physiologically-realistic model of the auditory periphery feeds its outputs to an ANN, acting as a 'synthetic listener'. Two varieties of ANN have been used:

1. a feed-forward net (perceptron) trained in supervised fashion by error reduction;
2. a competitive-learning (CL) net trained in unsupervised fashion.

4. RESULTS

4.1. Supervised training

Initially, a multilayer perceptron (MLP) trained by error back-propagation has been employed as the synthetic listener to label the patterns of neural firing activity from the P-D auditory model, using the **bp** software of McClelland and Rumelhart [11]. The neurogram data were presented to the net as follows. Spikes were counted in a (12×16)-cell analysis window stretching from -25 ms to 95 ms in 10 ms steps in the time dimension and from 1 to 128 in steps of 8 in the CF dimension. One MLP was constructed for each of the 3 stimulus series (bilabial, alveolar and velar). Each had $12 \times 16 = 192$ input units, a number of hidden units, and a single output unit to act as a voiced/unvoiced detector. Each net was trained 5 times starting from different, random, initial weights. The trained weight set selected for testing was that for which the smallest number of training epochs were necessary to reach the error criterion (0.005)

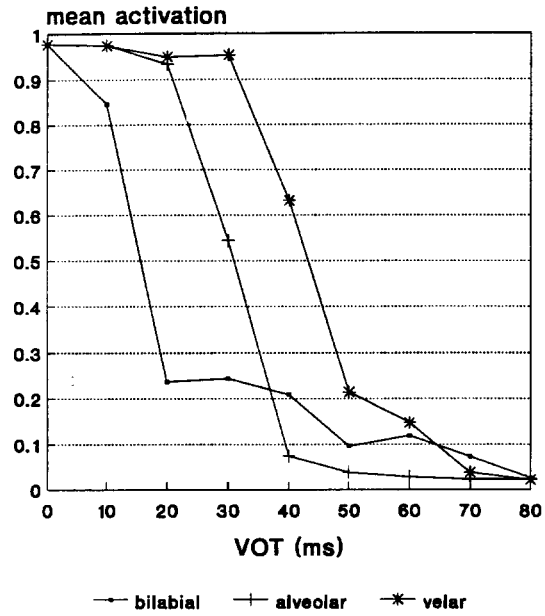

Figure 3: Mean output activation versus VOT for MLPs trained on neurograms from 0 ms and 80 ms endpoints.

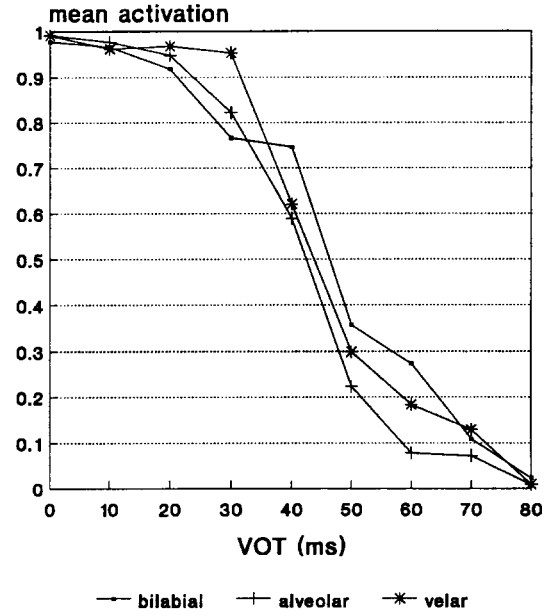

Figure 4: Mean output activation versus VOT for SLPs trained on filter-bank analysed stimuli.

on the presumption that this set is likely to have captured the most direct internal representation.

As in the Kuhl and Miller study with chinchillas (which had to be trained to respond appropriately to the stimuli), the MLP was trained on 50 repetitions of the endpoint stimuli (0 and 80 ms VOT) and tested on 50 repetitions of the intermediate values (10 ms to 70 ms in 10 ms steps). Because the P-D model simulates neural transduction at the hair cells (a stochastic process) it is probabilistic in nature. Hence, stimulus repetition produced non-identical neurograms. Target outputs were 1 for the voiced (0 ms VOT) stimuli and 0 for the unvoiced (80 ms VOT) stimuli.

Figure 3 shows the labelling function obtained by averaging output activations over the 50 stimulus presentations. In this case, there were 16 hidden units, but results were insensitive to this number: essentially the same curves were obtained with a single-layer perceptron.

Comparing with Fig. 1, the composite model's responses closely mimic those obtained from human and animal listeners, even to the extent of replicating the shift of category boundary with place of articulation seen in the original studies. Thus, the model is clearly capturing the essence of CP, but in some as yet unknown way.

To try to localise the site of the effects of categorisation and boundary shift, the P-D front-end with its hair-cell transduction model and filter-bank modelled on neural tuning data was substituted by a simple bark-spaced Fourier (FFT) analysis. The ANN was then trained on the spectral energy within the same time-frequency cells as previously. Given the insensitivity of the earlier results to presence or absence of a hidden layer, single-layer perceptrons (SLPs) were used here. This reduced training times markedly. Figure 4 depicts the results.

These are dramatic. While CP (i.e. non-uniform labelling and non-monotonic discrimination) are maintained, the boundary shift is totally abolished.

4.2. Unsupervised training

The motivation for using unsupervised learning was severalfold. First, CL nets [3] act as regularity detectors so, in principle, allowing the features underlying the perception of voicing to be explored. Second, the units are linear so removing one of the least interesting possible causes of non-uniformity in the model's responses (see Section 5 below). Further, it may be that is quite natural for ANNs trained in supervised mode on the endpoints of a continuum to partition the continuum in two about its centre. Such a 'label learning' (or 'anchor') effect is indeed one of the classical hypotheses of CP [12]. The CL paradigm offers the possibility of, at least, removing the labels. At this stage, however, training has (for simplicity) used only the endpoint patterns – those with the P-D model in place.

For consistency with the work using supervised (back-propagation) training, the cl software of McClelland and Rumelhart [11] has been used thus far. The ANN had 192 input units as before, fully connected via excitatory links to an output cluster of just 2 mutually-inhibitory ('winner-take-all') units (Fig. 5). In this configuration, the net should act as a binary feature detector with one of the output units detecting voicing and the other detecting its absence. Unfortunately, training this network has proved problematic: depending upon the initial, random weights, one unit or the other captures *all* the input patterns when learning stabilises.

We believe the most likely reason for this problem to be the inflexibility of McClelland and Rumelhart's software which implements a single, global learning rate such that

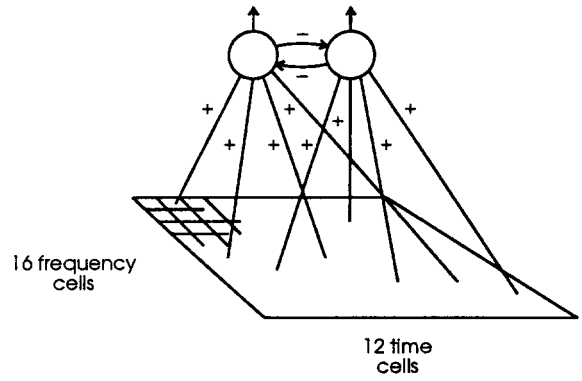

Figure 5: The competitive-learning net used had 192 input units and 2 output units so as to act as a binary feature detector. Connections are excitatory (+) between inputs and output units, and inhibitory (−) between output units.

only the winning unit learns. The problem of one unit capturing all input patterns is well documented in [3]. One solution is the so-called *leaky* learning model which adapts both the winning and losing units towards the input pattern. This model is currently being implemented in our own software, and results obtained with it will be presented at the conference.

5. DISCUSSION AND CONCLUSIONS

The work with supervised training clearly shows that a dissociation is possible between the 'basic' categorisation effect and the boundary shift with place of articulation. While the observed non-uniformity of labelling is insensitive to the details of peripheral auditory processing (i.e. to the presence or absence of the P-D front-end) in the model, this is definitely not true of the boundary shift. This suggests that either the hair-cell component or the precise auditory filter time-frequency characteristics are crucial to the latter effect. Interestingly, Darling *et al* [13] have repeated some of our earlier work whose results were summarised in Fig. 3. They confirmed the non-uniform categorisation but failed to find the boundary shift that we did. Our two implementations use identical stimuli and ANNs, but have a different front-end auditory model. Thus, by comparison, it should be possible to identify the exact component(s) of the P-D front-end which account(s) for the boundary shift effect. This comparison is proceeding.

The 'basic' CP effect (i.e. non-uniform labelling in this work) seems less critically related to peripheral processing. It could well be due to the inherent behaviour of the ANN itself. One possibility is that the non-uniformity reflects trivially the non-linear (sigmoidal) activation function of the output units. Harnad *et al* [14] have demonstrated that the sort of categorisation observed here is a more or less natural consequence of supervised training of non-linear networks. Continuing work with unsupervised, competitive learning should reveal the extent to which the model's responses reflect real aspects of auditory processing.

6. ACKNOWLEDGEMENT

The VOT stimuli used here were produced at Haskins Laboratory, New Haven, Connecticut, with assistance from NICHD Contract NO1-HD-5-2910.

7. REFERENCES

[1] B.H. Repp (1984) "Categorical perception: Issues, methods and findings", in *Speech and Language, Vol. 10, Advances in Basic Research and Practice*, N. Lass (ed.), Academic, Orlando, FL, 244–335.

[2] R.G. Crowder (1989) "Categorical perception of speech: A largely dead horse, surpassing well kicked", *Behavioral and Brain Sciences*, **12**: 760.

[3] D.E. Rumelhart and D. Zipser (1985) "Feature discovery by competitive learning", *Cognitive Science*, **9**: 75–112.

[4] A.M. Liberman, P.C. Delattre and F.S. Cooper (1958) "Some cues for the distinction between voiced and voiceless stops in initial position", *Language and Speech*, **1**: 153–167.

[5] P.K. Kuhl and J.D. Miller (1978) "Speech perception by the chinchilla: identification functions for synthetic VOT stimuli", *Journal of the Acoustical Society of America*, **63**: 905–917.

[6] C.C. Wood (1976) "Discriminability, response bias, and phoneme categories in discrimination of voice onset time", *Journal of the Acoustical Society of America*, **60**: 1381–1239.

[7] S. Rosen and P. Howell (1987) "Auditory, articulatory and learning explanations of categorical perception of speech", in *Categorical Perception: The Groundwork of Cognition*, S. Harnad (ed.), Cambridge University Press, 113–160.

[8] A. Abramson and L. Lisker (1970) "Discrimination along the voicing continuum: cross-language tests", *Proceedings of 6th International Congress of Phonetic Sciences, Prague, 1967*, Academia, Prague, 569–573.

[9] M.J. Pont and R.I. Damper (1991) "A computational model of afferent neural activity from the cochlea to the dorsal acoustic stria", *Journal of the Acoustical Society of America*, **89**: 1213–1228.

[10] R.I. Damper, M.J. Pont and K. Elenius (1990) "Representation of initial stop consonants in a computational model of the dorsal cochlear nucleus", *Speech Transmission Laboratory Quarterly Progress and Status Report, Royal Institute of Technology (KTH), Stockholm*, **STL-QPSR 4/1990**: 7–41.

[11] J.L. McClelland and D.E. Rumelhart (1988) *Explorations in Parallel Distributed Processing: A Handbook of Models, Programs, and Exercises*, MIT Press/Bradford Books, Cambridge, MA.

[12] H. Lane (1965) "Motor theory of speech perception: a critical review", *Psychological Review*, **72**: 275–309.

[13] A.M. Darling, M.A. Huckvale, S. Rosen and A. Faulkner (1992) "Phonetic classification of the plosive voicing contrast using computational modelling", *Proceedings of the Institute of Acoustics*, **14**(6): 289–292.

[14] S. Harnad, S.J. Hanson and J. Lubin (1991) "Categorical perception and the evolution of supervised learning in neural nets", in *Working Papers of the AAAI Spring Symposium on Machine Learning of Natural Language and Ontology*, D.W. Powers and L. Reeker (eds.), 65–74.