ON THE LEARNABILITY OF THE VOICING CONTRAST FOR INITIAL STOPS

R.I. Damper *S.R. Gunn*

Department of Electronics and Computer Science,
University of Southampton, Southampton SO17 1BJ, UK

ABSTRACT

The categorical perception (CP) of syllable-initial stop consonants has been intensively studied using psychophysical procedures over many decades. However, computational models consisting of an auditory 'front end' and a learning system as a 'back end' convincingly mimic the essentials of CP. Unlike real listeners, such models can be systematically manipulated to uncover the basis of their categorisations. In this paper, we explore the use of modern inductive learning techniques in simulating CP.

1. INTRODUCTION

A vital part of understanding speech perception is understanding the transformations which relate physically-continuous acoustic stimulation to the discrete code of phonetic percepts. It is immeasurably easier to observe the restructuring of information in a software model of auditory processing than in experiments using human or animal listeners. We have worked for several years on such models, trying to understand the mechanisms of the categorical perception (CP) of voicing in syllable initial stop consonants [1, 2, 3, 4]. Our models have a physiologically-realistic 'front end' producing simulated firing patterns in response to synthetic speech sounds at the level of the auditory nerve. A trainable neural network 'back end' then learns to categorise these patterns. We find that subtle aspects of the psychophysical behaviour of real listeners are mimicked by these models.

However, the neural network paradigm has fundamental shortcomings in terms of learning theory, which mean that our results could be artefactual. Most important, there is only one instance of each specific synthesised speech token, so that networks are inevitably undertrained. Here we use the modern inductive inference technique for small sample sizes of support vector machines [5] to simulate CP.

The voiced/unvoiced distinction is fundamental to speech communication, playing a major contrastive role in all languages. As such, it has received much attention in studies of speech perception. In early work, Liberman and his colleagues [6] investigated the perception of voicing in syllable-initial stop consonants by English listeners 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. As a consequence, labelling (identification) functions are non-linear, having a steep region around the category boundary, and discrimination functions are non-monotonic, peaking at the boundary. There is also a phoneme-boundary shift with place of articulation. Taking the 50% points on the labelling functions as the voiced/unvoiced boundaries, then as the place of articulation moves from bilabial (/ba–pa/ VOT series) through alveolar (/da–ta/) to velar (/ga–ka/), so the boundary moves from about 25 ms through about 35 ms to approximately 42 ms (e.g. [7]).

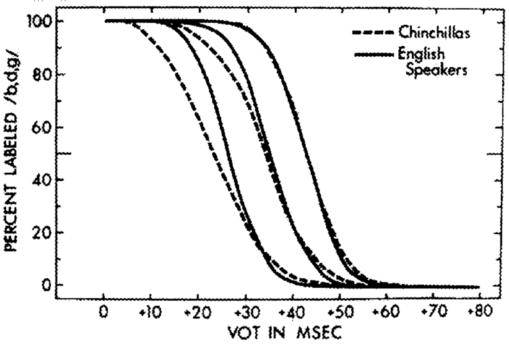

Figure 1: Labelling curves for syllable-initial stop consonants varying in voice-onset time (VOT) for human and chinchilla listeners, from Kuhl and Miller [8].

An intriguing finding is that such CP is also observed in non-human listeners. This was first shown for chinchillas by Kuhl and Miller [8] but has since been confirmed for a number of animal species. Figure 1 shows labelling curves from humans and chinchillas from [8] illustrating the steep slope around the category boundary and the movement of the boundary with place of articulation. Observed behaviours are remarkably close for the two different species: chinchillas exhibit boundary values not significantly different from humans (although the curves are less steep). This convergence of behaviours has usually been taken to indicate that categorisation is basic to the operation of animal auditory systems, rather than relying on the existence of a 'phonetic' sub-system specialised for speech perception.

2. AUDITORY PREPROCESSING

The synthetic consonant-vowel syllables used in this study were supplied by Haskins Laboratories. They are digitally-sampled (rate 10 kHz) versions of those developed by Abramson and Lisker [9] and employed extensively in the psychophysical experimentation reviewed above. They consist of three series in which VOT varies in 10 ms steps from 0 to 80 ms, simulating a series of English, prestressed, bilabial (/ba–pa/), alveolar (/da–ta/) and velar (/ga–ka/) syllables.

We have used Pont and Damper's [10] model (hereafter the 'P-D' model) extensively as an auditory front-end. Input stimuli are passed through a filterbank designed to mimic the physiological tuning curves of cat AN data, with appropriate basilar membrane (BM) delay characteristics and frequency rescaling reflecting the range of human hearing. The filters are uniformly spaced in terms of BM place. Mechanical-to-neural transduction, amplitude compression and two-tone suppression are modelled phenomenologically. The original P-D model includes simulations of cochlear nucleus processing but here outputs are taken from the auditory-nerve level, in the form of time of firing of 128 simulated auditory nerve fibres spanning the frequency range 50 Hz to 5 kHz. The parameters of the model are fixed according to physiological measurements (or other direct evidence) where available and to fit observed gross responses where relevant parametric, physiological knowledge is not available.

The P-D model outputs form the inputs to one of a variety of learning systems. The mechanical-to-neural transduction component of the P-D model reflects the stochastic nature of this process in the (real) auditory system. This allows us to produce a data set for training the learning system, even though we only have one example of each stimulus for each VOT and place of articulation, simply by reusing each stimulus repeatedly as input. However, the P-D model is computationally expensive so, for practical reasons, we have limited this to 50 repetitions. Thus we face (unavoidably) a small-sample size problem.

The stimuli were applied at time $t = 0$ at a simulated level of 65 dB sound pressure level (SPL). Activity before $t = 0$ is spontaneous. Damper *et al.* [1] confirm that the responses ('neurograms') are an excellent fit to the available physiological data. However, neurograms are not suitable for input to the neural network to be trained to categorise the auditory patterns. Retaining detailed information on the time of firing of each (simulated) spike implies a very high data rate and, consequently, a learning system with too many parameters to be estimated given the paucity of the data. To effect data reduction, spikes were counted in a $(12 \times 16) = 192$-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.

3. PERCEPTRONS

We have previously employed a variety of neural network architectures as the back end, including associative networks [4]; competitive-learning networks [3]; multilayer perceptrons [1, 4] and single-layer perceptrons (SLPs) [3]. All are capable of mimicking the behaviour of real listeners with more or less fidelity. However, because of their simplicity, we focus here on SLPs.

Three SLPs were constructed: one for each of the bilabial, alveolar and velar series. Each had 192 inputs and a single output node (sigmoidal activation function). Networks were trained by back-propagation to produce a '1' output on the 50 repetitions of the 0 ms VOT responses (auditory patterns from the P-D model) and a '0' output on the 50 repetitions of the 80 ms VOT responses. The output activation for unseen patterns can thus be construed as signifying the degree of voicing. Training on the endpoints in this way mimics the training of Kuhl and Miller's chinchillas [8]. As in the latter study, generalisation was then tested on the full range (0 ms to 80 ms in 8 steps) of responses.

Figure 2 shows typical labelling functions obtained by averaging output activations over the 50 stimulus presentations for each of the three places of articulation. Labelling functions like these were consistently obtained over many repetitions of the training. They closely mimic results from human and animal listeners, even replicating the shift of category boundary with place of articulation. Boundaries are at approximately 17 ms, 30 ms and 42 ms for the bilabial, alveolar and velar series respectively – very close to those for real listeners. Thus, the neural model is capturing the 'essence' of CP. The behaviour is *emergent* – it is not explicitly programmed into the simulation – which strengthens the feeling that the effects are quite basic to the way these stimuli are perceived. It is surely suggestive that very similar results are obtained from very different human, animal and machine listeners.

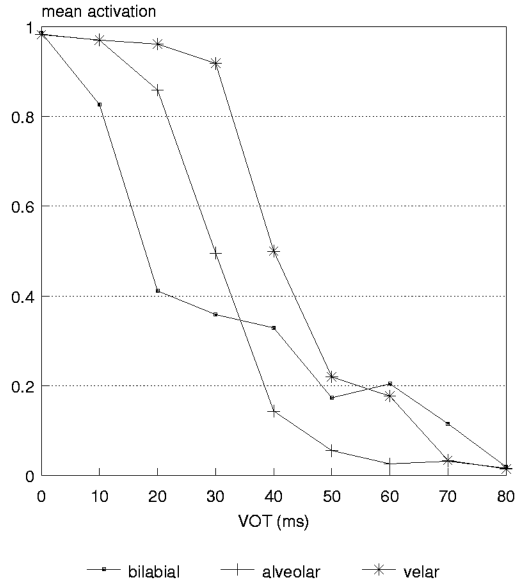

Figure 2: Mean output activation versus VOT for SLPs trained on neurograms from 0 ms and 80 ms endpoints.

Unlike real listeners, a computational model can be analysed to determine the basis of its behaviour. A major attraction of the SLP is that we can straightforwardly identify the areas of the neurogram which contribute to the categorisation behaviour: all connections are direct from the neurogram to the SLP's output, without intervening hidden units. In [3], we present the results of such an analysis. We find that categorisation can be explained by mechanism in which higher levels of the auditory system focus on a particular region of auditory nerve time-frequency activity and aggregate spike activity in this region. But the SLP has some shortcomings as a learning system which mean that these findings need to be treated with caution.

First, we have very sparse training data so are they sufficient? Several authors (notably [11]) have considered the bounds on the required number of training examples to produce valid generalization for nets of a given size, but (because of the assumptions made) these are generally "too loose, leading to impractical results" [12, p. 238]. A well-known rule of thumb [13] is that there should be 10 times as many training examples as adjustable parameters, implying a need here for about 1930 training instances whereas we have only 2 (endpoints) × 50 (repetitions) = 100 auditory patterns. Hence, we confront a problem for statistical learning of small sample size. Second, the sampling statistics of the training data reflect the (Poisson) statistics of the mechanical-to-neural transduction taking place in the hair cells of the cochlea, which will tend to produce training data clustered around the expected (mean) value. However, from the learning theory perspective, we would prefer data close to the category boundary to be well represented. Third, perceptron learning has no explicit control of generalisation. Further, the technique of supervised training on the VOT endpoints predisposes the nets to produce something close to the 'correct' 1/0 voiced/unvoiced values at extreme VOTs. In particular, it means that each net may be doing no more than simply placing the boundary at the midpoint of VOT between the (average) endpoint exemplars, albeit in a 192-dimensional space.

Mitigating these objections, however, the SLPs do produce very realistic behaviours and do so consistently indicating that the issues detailed above are not fatal to the modelling enterprise. Nonetheless, it is prudent to overcome as many of these shortcomings as possible. Thus, we treat the results outlined above as preliminary, and seek to confirm them using the modern inductive inference technique of support vector machines (SVMs) [5, 14].

4. SUPPORT VECTOR MACHINES

To address some of the shortcomings of SLPs (e.g. lack of capacity control), we use SVMs [5, 14]. These incorporate the structural risk minimisation principle, derived from the theory of small sample sizes. In addition to enforcing correct classification, a further constraint maximises the *margin*, i.e. the distance between the separating hyperplane and the nearest data point of each class.

The distance of a point $\mathbf{x}$ from a hyperplane $(\mathbf{w}, b)$ is:

$$d((\mathbf{w}, b), \mathbf{x}) = \frac{|\mathbf{w} \cdot \mathbf{x} + b|}{||\mathbf{w}||}$$

where $\mathbf{w}$ and b can be interpreted as the weight vector and bias of a formal neuron. For a two-class problem, as here, the margin is given as:

$$\rho((\mathbf{w}, b), \mathbf{x}) = \min_i\{(\mathbf{w}, b), \mathbf{x}_i\} + \min_j\{(\mathbf{w}, b), \mathbf{x}_j\}$$
$$x_i \in A, \ x_j \in B$$

Maximising $\rho()$ produces good control of generalisation ability and guarantees a unique solution to the problem, unlike SLPs.

Three SVMs (bilabial, alveolar and velar) were constructed. They were trained using the same 100 patterns (50 repetitions of responses to the 0 ms endpoint and 50 repetitions of responses to the 80 ms endpoint) as the SLPs. A straightforward SVM was used with an architecture equivalent to an SLP [14] with a hard-limiting (signum function) threshold unit on the output:

$$f(\mathbf{x}) = \mathrm{sgn}(\mathbf{w} \cdot \mathbf{x} + b)$$

and no additional capacity control [5].

The results with this hard classifier are shown in Figure 3. Each of the three curves depicts the average classification over each of the 50 repetitions for the relevant series. Taking the 50% midpoint between voiced and unvoiced to represent the category boundary gives values of 16 ms, 30 ms and 40 ms for the bilabial, alveolar and velar series respectively, comparing well to the values for the SLP and for real listeners.

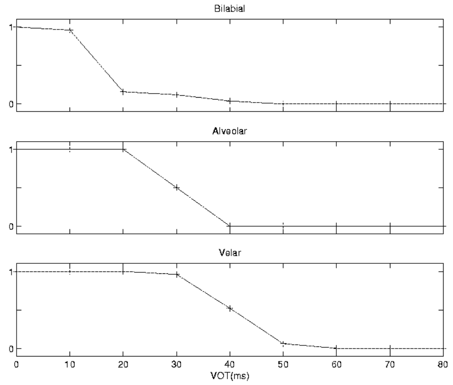

Figure 3: Labelling curves for the hard SVM classifier. Category boundaries are essentially identical to those for the SLP.

The SVM implicitly realises a form of data selection. Only input patterns with non-zero Lagrange multipliers – the *support vectors*, (SVs) – will contribute to the model. Thus, the SVs are the subset of patterns conveying the vital information about the category boundary. These will lie on the boundary of the maximised margin. The two margin boundaries (one for each class) are parallel, and the optimal separating hyperplane (OSH) is parallel and equidistant to both. These three hyperplanes fully characterise the separation of the classes, and so provide a convenient method for knowledge extraction. We use components of the normal vector to the hyperplane(s) – actually the weight vector $\mathbf{w}$ – for this purpose. This 192-dimensional vector uniquely characterises the knowledge extracted by the model – which differentiates voiced from unvoiced categories. The percentage of support vectors for each SVM was 41%, 37% and 45% for bilabial, alveolar and velar respectively, divided roughly half and half between the voiced and unvoiced categories.

To visualise the information extracted by the model, squared components of the normal-to-the-OSH vectors are plotted in grey-scale form in Figure 4. The crucial information (dark regions) lies in the low frequency (first formant transition) region just after acoustic stimulus onset. This parallels the finding with the SLPs [3]. Again, the precise location of this region shifts in the three cases (bilabial, alveolar, velar) in the same way as the boundary point for the SLPs and for real listeners.

5. CONCLUSIONS

The categorisation of syllable-initial stops into voiced and unvoiced categories has been intensively studied in human and animal listeners. More recently, attention has turned to the perception of such synthetic speech sounds by machine. A variety of computational learning systems is capable of mimicking the categorisation behaviour of real listeners, including the systematic shift of the phoneme boundary with place of articulation. This behaviour is emergent: it is not programmed into the simulation but arises as a consequence of aggregating time-frequency information in auditory-nerve firing patterns. The key property of software models of audition is that they can be analysed – to extract the learned phonetic knowledge which underpins their behavior – in a way which is not possible with real listeners.

We have described the use of SLPs and SVMs to simulate the categorisation behaviour of real listeners. Analysis of the averaged activity of each SLP's weighted connections and of the normal-to-the-margin vector of the SVMs

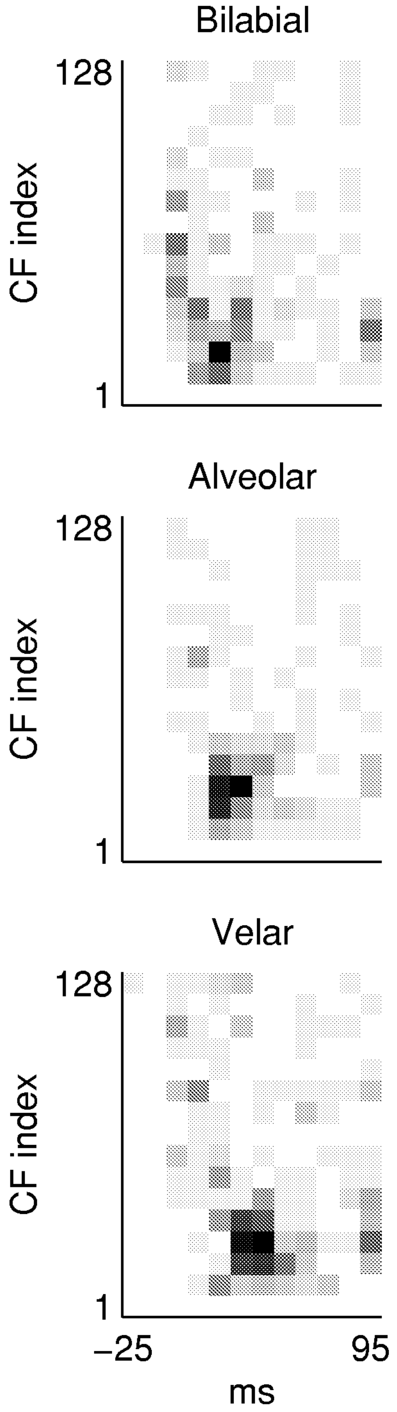

Figure 4: Squared normal-to-the-OSH vectors in greyscale form for the three stimulus series.

reveals that highly-localised low-frequency information in the time period shortly after stimulus onset is sufficient to predict the category boundary. Thus, a compact explanation of the basic phoneme-boundary effect is available. We stress that the full range of perceptual phenomena associated with the categorisation of these speech sounds is rather more complex than it has been portrayed here. To keep our treatment concise and focused, we have limited consideration to the most fundamental aspects of CP.

6. REFERENCES

1. R. I. Damper, M. J. Pont, and K. Elenius. Representation of initial stop consonants in a computational model of the dorsal cochlear nucleus. Technical Report STL-QPSR 4/90, Speech Transmission Laboratory Quarterly Progress and Status Report, Royal Institute of Technology (KTH), Stockholm, 1990. Also published in W. A. Ainsworth (Ed.), *Advances in Speech, Hearing and Language Processing, Vol. 3 (Part B)* (pp. 497–546). Greenwich, CT: JAI Press, 1996.
2. R. I. Damper. Connectionist models of categorical perception of speech. In *Proceedings of IEEE International Symposium on Speech, Image Processing and Neural Networks*, volume 1, pages 101–104, Hong Kong, 1994.
3. R. I. Damper, S. Harnad, and M. O. Gore. A computational model of the perception of voicing in initial stop consonants. Submitted to *Journal of Phonetics.*
4. R. I. Damper and S. Harnad. The psychophysics of synthetic categorical perception. Submitted to *Perception and Psychophysics.*
5. V. N. Vapnik. *The Nature of Statistical Learning Theory.* Springer, New York, NY, 1995.
6. A. M. Liberman, P. C. Delattre, and F. S. Cooper. Some cues for the distinction between voiced and voiceless stops in initial position. *Language and Speech*, 1:153–167, 1958.
7. L. Lisker and A. Abramson. The voicing dimension: Some experiments in comparative phonetics. In *Proceedings of 6th International Congress of Phonetic Sciences, Prague, 1967*, pages 563–567. Academia, Prague, 1970.
8. P. K. Kuhl and J. D. Miller. Speech perception by the chinchilla: Identification functions for synthetic VOT stimuli. *Journal of the Acoustical Society of America*, 63:905–917, 1978.
9. A. Abramson and L. Lisker. Discrimination along the voicing continuum: Cross-language tests. In *Proceedings of 6th International Congress of Phonetic Sciences, Prague, 1967*, pages 569–573. Academia, Prague, 1970.
10. M. J. Pont and R. I. Damper. 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, 1991.
11. E. B. Baum and D. Haussler. What size net gives valid generalization? *Neural Computation*, 1:151–160, 1989.
12. N. K. Bose and P. Liang. *Neural Network Fundamentals with Graphs, Algorithms and Applications.* McGraw-Hill, New York, NY, 1996.
13. B. Widrow. Adaline and madaline - 1963: Plenary speech. In *Proceedings of 1st IEEE International Conference on Neural Networks*, volume 1, pages 143–158, San Diego, CA, 1987.
14. C. Cortes and V. N. Vapnik. Support vector networks. *Machine Learning*, 20:1–25, 1995.