6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

LEARNING PHONETIC DISTINCTIONS FROM SPEECH SIGNALS

Robert I. Damper and Steve R. Gunn
Image, Speech and Intelligent Systems (ISIS) Research Group
Department of Electronics and Computer Science
University of Southampton
Southampton SO17 1BJ, UK.

ABSTRACT

Previous work has shown that connectionist learning systems can simulate important aspects of the categorization of speech sounds by human and animal listeners. Training is on representations of synthetic, exemplar voiced and unvoiced stop consonants passed through a computational model of the auditory periphery. In this work, we use the modern inductive inference technique of support vector machines (SVMs) as the learning system. Visualization of the SVM's weight vector reveals what has been learned about the voiced/unvoiced distinction.

INTRODUCTION

For several years, we have worked on computational models of sound-to-symbol transformation, with a view to understanding the possible acoustic and auditory bases of the categorical perception (CP) of voicing in synthetic syllable initial stop consonants [1, 2, 3]. This has revealed that any reasonably general learning system is able to categorize the patterns of simulated auditory nerve activation in a way which mimics the psychophysical behaviour of real listeners. The question which then arises, and which we address here, is: what phonetic knowledge has been captured by the network?

Previous work used single-layer perceptrons (and other neural networks) as the learning system. However, from the point of view of modern statistical learning theory, the perceptron approach has several shortcomings. Our purpose in this paper is to overcome these using the *support vector machine* (SVM) method [4], in the specific context of the extraction of phonetic knowledge pertaining to the voiced/unvoiced distinction.

CP OF INITIAL STOPS

The voiced/unvoiced distinction is fundamental to speech communication, and has received much attention in studies of speech perception. In early work, Liberman et al. [5] 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. Thus, bilabial stimuli with small VOTs are perceived and labeled as /ba/ while those with large VOTs are perceived and labeled as /pa/. As a consequence, labeling (identification) functions are 'warped', 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. As the place of articulation moves back in the vocal tract 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.

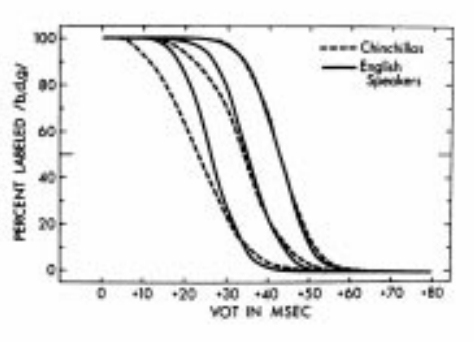

Figure 1: Labeling curves for syllable-initial stop consonants varying in voice-onset time (VOT) for human and chinchilla listeners from [6].

An intriguing fact is that this categorical behaviour is also observed in non-human listeners. This was first shown for chinchillas by Kuhl and Miller [6] but has since been confirmed for other animal species. Figure 1 shows labeling curves illustrating the warping around the category boundary and the boundary movement with place of articulation. Observed behaviours are remarkably close for the two kinds of listener: the chinchillas exhibit boundary values not significantly different from the humans (although the curves are less steep). This striking similarity has usually been taken to indicate that categorization is basic to the operation of animal auditory systems.

The topic of CP has generated much study, debate and controversy: see [3] for full discussion and original references. According to Rosen and Howell [7], three of the most influential theories of CP have been: articulatory ('motor' theory), auditory and learned explanations. Overall, their opinion is that none of the extant theories can, by itself, explain all the experimental data. Thus, they concur with Soli's view [8, p. 2150]: "Although the existence of discrimination peaks at the voicing boundary is a robust experimental phenomenon, a satisfactory theoretical explanation of their occurrence has yet to be given."

AUDITORY PREPROCESSING

The synthetic consonant-vowel syllables used in this study were supplied by Haskins Laboratories. They are digitally-sampled versions (sampling rate 10 kHz) of those developed by Abramson and Lisker [9], consisting of three series in which VOT varies in 10 ms steps from 0 to 80 ms, simulating a series of English, pre-stressed, bilabial (/ba–pa/), alveolar

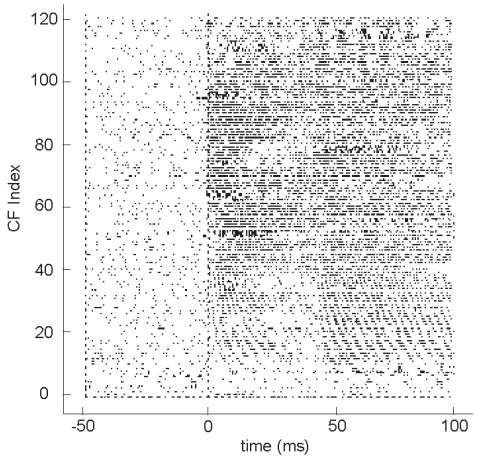

Figure 2: Response of the P-D model to the bilabial stimulus with 40 ms VOT in the form of a neurogram. Each dot depicts the firing of a neuron at the indicated CF and time.

(/da–ta/) and velar (/ga–ka/) syllables. The use of these synthetic tokens, rather than real speech, allows comparison with the seminal studies (such as [6]) which have also used them.

We have used the model of Pont and Damper [10] (hereafter the P-D model) as an auditory front-end pre-processor. Input stimuli are passed through a filterbank designed to mimic the physiological tuning curves of cat auditory nerve data, with appropriate basilar-membrane delay characteristics and frequency rescaling reflecting the range of human hearing. The filters are equally spaced in terms uniformly in terms of basilar membrane place, rescaled to take account of the different ranges of human and cat hearing. After filtering, mechanical-to-neural transduction, amplitude compression and two-tone suppression are modeled phenomenologically. Output is 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 so as to fit observed gross responses where relevant physiological data are not available.

The outputs from the P-D model form the inputs to the learning system. Conveniently, the mechanical-to-neural transduction component of the model reflects the stochastic nature of this process in the (real) auditory system. This allows us to produce a data set for training, even though we only have one example of each stimulus for each VOT and place of articulation, by using each stimulus repetitively as input to the P-D model. However, the model is computationally expensive so 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. Activity before $t = 0$ is spontaneous. The model output is visualized as a neural time-frequency spectrogram ('neurogram') as in Figure 2. A dot indicates the firing of a neuron: the x and y coordinates are the time of firing and the neuron's center frequency (CF) respectively. To avoid loss of detail, only some 25,000 spikes are shown, corresponding to approximately the first 30 or 40 of the 50 repetitions. Damper et al. [1] confirm that the P-D model's responses are an excellent fit to available physiological data.

Neurograms in the form of Figure 2 are not suitable for input to the learning system. Retaining such detailed information implies a very high data rate and, consequently, a learning system with too many parameters to be estimated from sparse training data. Hence, spikes were counted in a $(12 \times 16) = 192$-cell 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. The time limits were chosen to exclude too much irrelevant detail pre-onset or during the steady-state portion of the /a/ vowel. 192 cells represents a reasonable compromise between the need for data reduction and the retention of important information. Figure 3 shows a typical such 'reduced' neurogram (in gray-scale form). Comparing with Fig. 2, the extent of the data reduction is obvious.

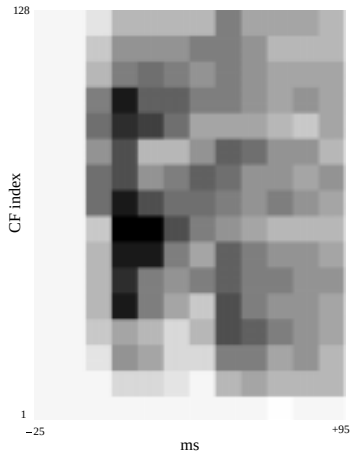

Figure 3: Typical 'reduced' neurogram as presented to the learning system: bilabial stimulus, 40 ms VOT (i.e. reduced version of the neurogram in Fig. 2.

MODELING CP WITH SVM'S

To address some of the shortcomings of perceptrons (see below), we use support vector machines (SVMs) [4]. SVMs incorporate the structural risk minimization principle, derived from the theory of small sample sizes. As well as requiring correct classification, a further constraint is added which maximizes the *margin*, i.e. the distance between the separating hyperplane and the nearest data point of each class. This leads to the notion of an *optimal* separating hyperplane (OSH) which – being more robust than a perceptron solution – typically provides better generalization.

The distance of point $\mathbf{x}$ from 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 are equivalent to the weight vector and bias of a formal neuron. For a two-class $(A,\ B)$ problem, as here, the margin is given as:

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

Maximizing ρ with respect to $\mathbf{w}$ and b produces good control of the generalization ability of the learning machine and guarantees a unique solution to the problem, unlike perceptron or back-propagation learning.

Quadratic programming optimization used Mészáros' BPMPD package [11], and was performed in the input space, i.e. assuming linear separability of the patterns. Three SVMs

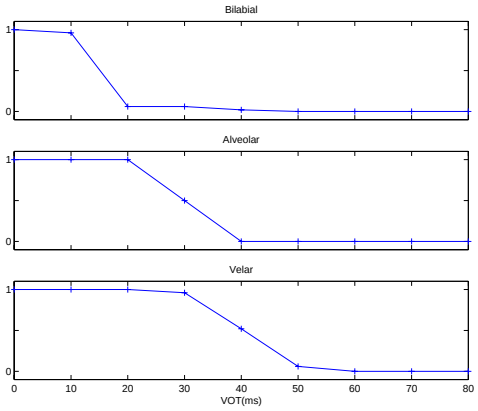

Figure 4: Labeling curves for the SVM classifier showing correct boundary placement as a function of place of articulation.

were constructed: one for each place of articulation. Training using 100 patterns: 50 repetitions of responses to the 0 ms VOT stimuli and 50 repetitions of responses to the 80 ms VOT stimuli. SVMs were used with an architecture equivalent to a perceptron [4], with a hard-limiting (signum function) threshold unit on the output. There was no additional capacity control.

Results are shown in Figure 4. Each of the three curves depicts the average classification over each of the 50 repetitions for the relevant series. Taking the mid-point (0.5) to represent the category boundary gives 16 ms, 30 ms and 40 ms for the bilabial, alveolar and velar series respectively – comparing very well to the values for real listeners.

The SVM implicitly realizes a form of data selection. Only input patterns with non-zero Lagrange multipliers – the so-called *support vectors*, (SVs) – will contribute to the model. Thus, the SVs are the patterns that convey the vital information about the category boundary. They lie on the boundary of the maximized margin. The two margin boundaries (one for each class) are parallel, and the OSH is parallel and equidistant to both. The margin boundaries fully characterize the separation of the classes, and so provide a convenient and powerful basis for knowledge extraction. We use the normal vector to the OSH (i.e. the weight vector $\mathbf{w}$) for this purpose. The percentage of support vectors for each SVM was 41%, 37% and 45% for bilabial, alveolar and velar stimuli respectively, divided roughly half and half between the voiced and unvoiced categories.

To visualize the knowledge captured, the 192-dimensional weight vectors $\mathbf{w}^2$ (formed by simply squaring each individual component) are depicted for the three stimuli series in Figure 5. Squaring emphasizes the magnitude of the information differentiating voiced and unvoiced categories. Dark areas in these figures correspond to areas of rich information in the input space. It can be seen that the crucial information lies in the low frequency (first formant transition) region just after acoustic stimulus onset. The precise location of this region shifts in the three cases (bilabial, alveolar, velar) in the same way as the boundary point for real listeners. The importance of the highly-localized $F1$ transition region close to the category boundary is also seen. Further, the spectrum at stimulus onset appears important in distinguishing voiced from unvoiced bilabial tokens. This is less the case for the velar stimuli and even less so for the alveolar stimuli. The $F1$ region at vowel onset (approximately 95 ms) is also distinctive

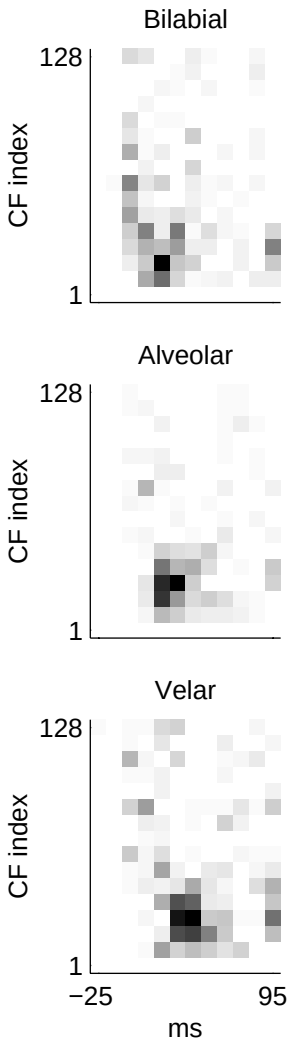

Figure 5: Gray-scale depiction of the squared weight vectors of the SVM, $\mathbf{w}^2$, for the three stimuli series.

in all three cases.

ROLE OF THE AUDITORY MODEL

The P-D front-end is capable of essentially perfect reproduction of detailed neural firing patterns. But how much of this sophistication is actually necessary? We have already shown that the back-end can be quite simple – a linear support vector machine (with hard, signum classification rule) gives good results. In this section, we attempt to simplify the front-end processing as far as possible, using short-time Fourier analysis.

The power spectral densities of each of the 3×9 stimuli were computed using a 256-point fast Fourier transform (FFT) with 25.6 ms frames centered on the 10 ms cell widths previously employed. (The overlap between consecutive frames was $(25.6 - 10)/2 = 7.8$ ms.) Spectral energy was summed in a (12×16) analysis window intended to parallel the reduced auditory representation used earlier as input to the back-end. Thus, the time dimension stretched from -25 ms to 95 ms in 10 ms steps and frequency from 0 to 5 kHz in steps of 312.5 Hz. So, here, the frequency dimension is divided up linearly (in Hz) rather than according to CF as earlier.

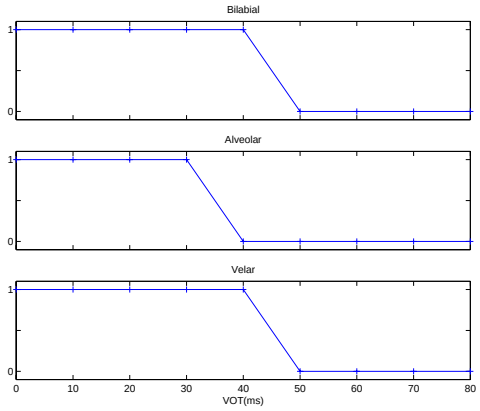

Figure 6: Labeling curves obtained from an SVM classifier with decision boundary constructed as the bisector (in 192-dimensional space) between FFT-analyzed endpoints. Correct movement of the boundary with place of articulation is not maintained.

There is now only a single token for each acoustic stimulus. This is unavoidable since we are no longer simulating the stochastic process of mechanical-to-neural transduction by the cochlear hair cells. Labeling curves for each of the three series were obtained by constructing a decision boundary as the bisector in 192-dimensional space between the two FFT-analyzed endpoints. A hard classification rule was used. Figure 6 shows the results. Correct movement of the boundary with place of articulation is not maintained, indicating that at least some aspect of the auditory transformation is essential to realistic simulation of CP. It is easy to see that this *must* be so. Since the SVM is linear, the non-linear segregation of the stimuli by place of articulation can only result from processing by the auditory front-end.

CONCLUSIONS

SVMs feature implicit data selection (support vectors) and, consequently, data reduction which equates to a powerful form of knowledge extraction. In this work, we have used normal vectors (192-dimensions) to the optimal separating hyperplane to represent the essential information about the voiced/unvoiced distinction. In all three cases (bilabial, alveolar, velar), the region of high information content is localized to the low-frequency (first formant transition) region shortly after stimulus onset. The precise location of this region shifts in the three analyses in the same way as the phoneme boundary point. Analysis of the SVMs allowed extraction of consistent knowledge to support the notion that the phonetic percept of voicing is easily and directly recoverable from a highly-localized region of the auditory representations. The place of articulation distinction is explained on the basis of systematic movement of the position of this highly-localized region. This offers fairly direct support for an auditory discontinuities explanation of CP. Thus, there is no need to posit any speech-specific mechanisms as in articulatory ('motor') theory. The fact that the auditory front-end was essential to correct modeling also points to an auditory rather than acoustic explanation (cf. [8]).

Replacement of the auditory front-end by a simpler Fourier analysis abolished correct movement of the boundary with place of articulation. Hence, some aspect(s) of peripheral auditory function is/are essential to correct simulation of categorization behavior. In future work, we intend to explore this further using a variety of simplified auditory front-ends. At this stage, however, we believe that frequency warping on a psychophysical scale and onset enhancement by the action of the hair cells are most likely to be essential. There are several other potentially fruitful avenues for future research. The Haskins stimuli were studied here because of their prior use in key studies of CP, and because synthesis allows easy production of 'ambiguous' tokens. However, the relation of these stimuli to real speech is tenuous. Hence, it is a priority to study real speech in the near future. We also intend to train SVMs on pairs of complete stimuli series, e.g. bilabial and alveolar, alveolar and velar etc., so that the trained network can be analyzed to uncover the feature(s) signaling the place of articulation distinction.

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, Royal Institute of Technology (KTH), Stockholm, 1990.

[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 and S. Harnad. Neural network models of categorical perception. *Perception and Psychophysics*, in press, 1999.

[4] V. N. Vapnik. *Statistical Learning Theory*. Wiley, New York, NY, 1998.

[5] 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.

[6] 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.

[7] S. Rosen and P. Howell. Auditory, articulatory and learning explanations of categorical perception of speech. In S. Harnad, editor, *Categorical Perception: the Groundwork of Cognition*, pages 113–160. Cambridge University Press, Cambridge, UK, 1987.

[8] S. D. Soli. The role of spectral cues in the discrimination of voice-onset time differences. *Journal of the Acoustical Society of America*, 73:2150 – 2165, 1983.

[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] C. Mészáros. The BPMPD interior point solver for convex quadratic problems. Technical Report WP 98-8, Computer and Automation Research Institute, Hungarian Academy of Sciences, Budapest, 1998.