Manuscript for Special Issue of *Applied Intelligence* on “Neural Networks and Structured Knowledge”, final version: April 8,1999

Extracting Phonetic Knowledge from Learning Systems: Perceptrons, Support Vector Machines and Linear Discriminants

Robert I. Damper, Steve R. Gunn and Mathew O. Gore

Image, Speech and Intelligent Systems (ISIS) Research Group

Department of Electronics and Computer Science

University of Southampton

Southampton SO17 1BJ, UK.

Acknowledgement: This work was partially funded by the UK Engineering and Physical Sciences Research Council via grant GR/K55110 “Neurofuzzy Model Building Algorithms” and by a studentship to Mat Gore. The synthetic speech stimuli were produced at Haskins Laboratories, New Haven, Connecticut, with assistance from NICHD Contract NO1-HD-5-2910. Thanks to Doug Whalen for his time and patience.

Address for correspondence:

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

Tel: +44 1703 594577
FAX: +44 1703 594498
Email: rid@ecs.soton.ac.uk

Abstract

Speech perception relies on the human ability to decode continuous, analog sound pressure waves into discrete, symbolic labels ('phonemes') with linguistic meaning. Aspects of this signal-to-symbol transformation have been intensively studied over many decades, using psychophysical procedures. The perception of (synthetic) syllable-initial stop consonants has been especially well studied, since these sounds display a marked categorization effect: they are typically dichotomised into 'voiced' and 'unvoiced' classes according to their voice onset time (VOT). In this case, the category boundary is found to have a systematic relation to the (simulated) place of articulation, but there is no currently-accepted explanation of this phenomenon. Categorization effects have now been demonstrated in a variety of animal species as well as humans, indicating that their origins lie in general auditory and/or learning mechanisms, rather than in some 'phonetic module' specialized to human speech processing.

In recent work, we have demonstrated that appropriately-trained computational learning systems ('neural networks') also display the same systematic behaviour as human and animal listeners. Networks are trained on simulated patterns of auditory-nerve firings in response to synthetic 'continuua' of stop-consonant/vowel syllables varying in place of articulation and VOT. Unlike real listeners, such a software model is amenable to analysis aimed at extracting the phonetic knowledge acquired in training, so providing a putative explanation of the categorization phenomenon. Here, we study three learning systems: single-layer perceptrons, support vector machines and Fisher linear discriminants. We highlight similarities and differences between these approaches. We find that the modern inductive inference technique for small sample sizes of support vector machines gives the most convincing results. Knowledge extracted from the trained machine indicated that the phonetic percept of voicing is easily and directly recoverable from auditory (but not acoustic) representations.

Keywords: speech perception, auditory processing, perceptrons, support vector machines, linear discriminant analysis

1 Introduction

Speech sound pressure waves impinging on the ear are subjected to a series of mechanical and then neural transformations resulting in the ultimate percept of a linguistic message. Hence, a vital part of understanding speech perception is understanding the staged transformations which relate the physically-continuous acoustic stimulation to the discrete code of phonetic percepts. In particular, in some as yet unknown way, the continuous-to-discrete transformation effects a variance reduction such that a variety of physical realizations map to the same speech-sound category, with obvious importance for effective communication between individuals with different speech production apparatus. Understanding how this is achieved is an aspect of the celebrated 'speech invariance problem' [1]. It is clear that the way speech sounds are categorized by a listener's auditory system is a matter of considerable scientific interest.

In the words of Summerfield [2, p. 51], however:

> "... the relationship between acoustical structure and perceived phonetic structure is complex and not obviously explained by known properties of the mammalian auditory system."

Further, as Kuhl and Miller [3, p. 906] write:

> "Ideally, *[one would like]* experimental methods that somehow allow one to intervene at various stages of the processing of sound to observe the restructuring of information that has occurred at each stage."

This intervention is obviously difficult or impossible to achieve in experiments using human or animal listeners. It is, however, immeasurably easier "to observe the restructuring of information" in a software model of auditory processing. For this reason, we have worked for several years on such models, with a view to understanding the possible acoustic and auditory bases of the categorical perception (CP) of voicing in syllable initial stop consonants [4]–[9].

We have focused on simulated representations of synthetic speech sounds at the level of the auditory nerve – the only neural pathway from ear to brain. Thus, the restructuring of information implicit in the acoustic input by the peripheral auditory system has been an integral part of the model. This work has revealed very clearly that any reasonably general learning system (i.e. 'neural network') – acting as a 'synthetic listener' [10] – 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? Given the similarity of behaviour between real and synthetic listeners, the operational principles employed by the latter offer a credible explanatory model of CP in humans and animals. Such an explanation would not only be useful for theorizing about speech perception, but might also assist in building technological devices such as improved speech recognition systems and speech preprocessors for cochlear implants for deaf people.

In initial work, we have used single-layer perceptrons (among other neural networks) as the learning system. However, from the point of view of modern statistical learning theory, the perceptron approach (dating as it does back to 1958 [11]) has several shortcomings. Vapnik [12, p. viii] has characterized modern theory as comprising four components which distinguish it from earlier 'neural network' learning:

1. Necessary and sufficient conditions for consistency of inference.
2. Bounds describing generalization ability.
3. Inductive inference for small samples based on these bounds.
4. Methods of implementation for the new type of inference.

Accordingly, a primary purpose in this paper is to exploit advantages 1–3, using the *support vector machine* (SVM) method of implementation [13], in the specific context of the extraction of phonetic knowledge pertaining to the voiced/unvoiced distinction from simulated auditory-nerve representations of syllable-initial stop consonants. Because both perceptrons and SVMs share similarities with the classical Fisher linear discriminant, we also include the latter among the techniques studied.

The remainder of this paper is organized as follows. Since this Special Issue is centrally concerned with representation and processing of structured knowledge, we first outline our modeling philosophy of extracting knowledge from trained neural networks. Then, in Section 3, we briefly review the very large literature on the categorical perception of initial stop consonants. In Section 4, we describe the computational auditory model employed here, including the synthetic speech stimuli which act as its inputs, and the pre-processing of its outputs which (among other things) effects a data reduction to make the learning problem tractable. We then describe, in Section 5, the use of perceptron learning to infer the auditory basis of the voiced/unvoiced distinction. In Section 6, we present corresponding results when perceptrons are replaced by support vector machines. We use a hard SVM classifier but, also, an arguably more realistic soft SVM classifier. In Section 7, the role of the auditory transformation in phonetic categorization is explored, and in Section 8 we ask if a Fisher linear discriminant can simulate categorization as well as the SVM classifier. Finally, Section 9 concludes.

2 Neural System Identification

Starting from auditory representations of important speech sounds, we aim to train neural networks to mimic the phonetic categorization behaviour of real listeners. Once a trained network displays the appropriate behaviour, we then need somehow to make its generalization abilities transparent. That is, the trained networks must be analysed to discover the features that have been learned, so extracting knowledge about the structure of phoneme categories. One objection to this kind of endeavor has been voiced by Crick, who writes [14]:

> "A possible criticism of … neural networks is that since they use such a grossly unrealistic learning algorithm, they really don't reveal much about the brain."

As pointed out by Crick himself, however, an answer to this criticism has been provided by Zipser [15] who treats "supervised learning models of the brain as a special case of system identification" [p. 853]. Zipser refers to this as *neural system identification*: an approach which "provides a systematic way to generate realistic models starting with a high-level description of a hypothesized computation and some architectural and physiological constraints about the area being modeled". The approach is based on the observation that learned solutions to perceptual and cognitive tasks typically, in their general characteristics, resemble the real biological ones.

This is very much the philosophy adopted here. Thus, we seek (1) to see if a learning system is capable of mimicking the behaviour of real listeners [1] and then (2) to analyse the system to see what it has learned. This contrasts with the ('Popperian') philosophy that science is based not on induction but on *ab initio* explanatory conjectures (hypotheses) which are open to empirical falsification and, thereby, refinement if not refutation [16]. It is, however, consistent with the perspective of Clark [17], who writes:

> "For the connectionist effectively inverts the usual temporal and methodological order of explanation … in connectionist theorizing, the high-level understanding will be made to revolve around a working program which has learnt how to negotiate some cognitive terrain."

Here, our programs learn to negotiate the cognitive terrain of the phonetic discrimination of voiced/unvoiced initial stops.

We do not, of course, deny the importance of explanatory hypotheses in general. Indeed, this work will culminate in at least one such hypothesis, whose falsification should be the basis of future work.

3 Categorical Perception of Initial Stop Consonants

In this section, we briefly review the background to this work in terms of current knowledge of the categorical perception of initial stop consonants. There is an enormous literature on this topic, so we can do no more here than give the flavor of the field. The reader is referred to [9] for fuller details.

3.1 Phoneme Boundary Effect

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 [18] investigated the perception of voicing in syllable-initial stop consonants by English listeners as voice-onset time (VOT) [2] 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, leading to the notion of a *phoneme-boundary effect* [19]. 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 [3]. Taking the 50% points on the labeling functions as the boundaries between voiced and unvoiced categories, then 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 (e.g. [20]). Why does this happen? To quote Kuhl [21, p. 365]:

> "We simply do not know why the boundary 'moves'."

Although this quote dates back to 1987, its sentiment is still relevant today (Kuhl, personal communication, 1998).

3.2 Animal Studies

An intriguing finding is that the categorical behaviour described above is also observed in non-human listeners. This was first shown for chinchillas by Kuhl and Miller [3] but has since been confirmed for a number of animal species. Figure 1 shows labeling curves [4] from [3] illustrating the warping around the category boundary and the movement of the boundary with place of articulation. Observed behaviours are remarkably close for the two different species of listener: the chinchillas exhibit boundary values not significantly different from the humans (although the curves are less steep). The fact that non-speaking animals produce results so similar to those from humans has usually been taken to indicate that categorization is basic to the operation of animal auditory systems, rather than relying on the existence of a 'phonetic' sub-system specialized for speech perception – see immediately below.

*** FIGURE 1 ABOUT HERE ***

3.3 Theories of CP

Few topics in speech science have generated as much study, debate and controversy as the phenomenon of CP. It would be out of place to say too much about this here: the reader unfamiliar with the background is referred to [8], [9] and [22] for full discussion and original references. It is, however, necessary to outline the bare bones of the debate.

Many theories partially explain CP. According to Rosen and Howell [23] writing in 1987, three of the most influential have been:

ARTICULATORY: in this theory, categorization is held to reflect the operation of a system which is specific to speech, in which there is early conversion of the continuous physical signal to a discrete code. For reasons of parsimony, this code is held to subserve both production (articulation) and perception (audition) – hence the alternative name of "motor theory".

AUDITORY: the notion here is that non-linearities in auditory sensitivity are exploited by human speech processes to introduce perceptual boundaries into a continuum of sounds. Auditory theory takes its main support from the findings (detailed above) of behavioral similarities between human and animal listeners. Clearly, the latter are unlikely to possess sub-systems for speech-specific processing but they do have similar peripheral auditory systems to humans.

LEARNED: this theory holds that phoneme categories are learned, rather than innate as in the two previous cases.

Of course, there are still other theories, given less prominence in their review (or dating from after 1987). We mention just one here – the acoustic discontinuity explanation advanced by Soli [24] in 1983. This author challenges the auditory non-linearities explanation, which is predicated on the assumption that the acoustic stimuli form a continuum. He writes [p. 2163]:

> "In its most basic form, this view asserts that psychophysical thresholds coincide with phonetic boundaries, so that discrimination peaks are in effect due to discontinuities in the auditory dimensions which comprise speech ... The present research demonstrates clearly that VOT series are not continuous auditory dimensions."

As pointed out by Rosen and Howell, however [p. 151]:

> "Although Soli's claims may seem to buttress the auditory-sensitivities point of view, they ... presume no special auditory sensitivities at all. The non-uniform discriminability across the VOT continuum is seen to be a function of the *acoustic* structure of the stimuli and would occur with a wide variety of *auditory* systems, as long as they did some kind of frequency analysis." *[our emphasis]*

Overall, Rosen and Howell's opinion is that none of the extant theories can, by itself, explain all the experimental data. Thus, they concur with Soli's view [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."

A prime motivation for this work is to see what insights neural network modeling can offer into the theoretical controversy surrounding CP. We wish to emphasize that – at the outset – we are essentially agnostic in respect of the various theoretical positions, except that we believe that some component of learning must play a part in any credible theory of CP.

4 Auditory Preprocessing

In this section, we detail the acoustic (synthetic speech) stimuli and outline the way that they were processed to produce simulated auditory representations. We also describe how these auditory representations were further processed to form the inputs to the learning system.

4.1 Stimuli

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 [25] 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, pre-stressed, bilabial (/ ba–pa /), alveolar (/da–ta /) and velar (/ga–ka /) syllables.

There were (at least) two reasons for using synthetic speech stimuli in this study, rather than real speech. The first, and probably the most important, is to allow comparison with the seminal studies (such as that of Kuhl and Miller) which have used these stimuli. Indeed, the Abramson and Lisker stimuli have become something of a 'gold standard' in this area. The second reason is that the VOTs of human productions of stop/vowel syllables tend to cluster around values away from the phoneme boundary [20]. Thus, we were concerned that a corpus of real speech (unless very large) would not have enough productions in the 'ambiguous' region close to the boundary for our purposes.

4.2 Computational Modeling of the Auditory Periphery

Nossair and Zahorian [26] used automatic speech recognition techniques in an attempt to discover the important acoustic features which distinguish initial stops – a similar goal to ours. In conclusion, they say [p. 2990]: "Such features might be more readily identifiable if the front-end spectral processing more closely approximated that performed by the human auditory system". It could be argued that, ideally, any such front-end computational model should faithfully simulate details of neural function and anatomy. Obviously, we have neither the neurobiological knowledge nor the computer power to do so for the complete auditory system. Enough is known of peripheral function, however, to be able to construct models which mimic auditory nerve (AN) firing patterns well.

We have used the model of Pont and Damper [27] (hereafter the P-D model) extensively as an auditory front-end. Briefly, input stimuli are passed through a filterbank designed to mimic the physiological tuning curves of cat AN data, with appropriate basilar-membrane delay characteristics and frequency rescaling reflecting the range of human hearing. The filters are equally spaced in terms of the Greenwood equation [28] (i.e. uniformly in terms of basilar membrane place) rescaled to take account of the different ranges of human and cat hearing [29]. After filtering, mechanical-to-neural transduction, amplitude compression and two-tone suppression are modeled phenomenologically [5]. Output is in the form of time of firing of 128 simulated auditory nerve fibres spanning the frequency range 50 Hz to 5 kHz. Again, the center frequencies (CFs) of the simulated fibres reflect the spacing of the filters according to the Greenwood equation. The parameters of the model are not adjustable (during some training process) but are fixed according to physiological measurements (or other direct evidence) where available and so as to fit observed gross responses where relevant parametric, physiological knowledge is not available.

The outputs from the P-D model form the inputs to one of a variety of learning systems – from which we aim to extract phonetic knowledge relevant to the voiced/unvoiced distinction. Conveniently for our purposes, the mechanical-to-neural transduction component of the model reflects the stochastic nature of this process in the (real) auditory system. This gives us a way of producing 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. We simply use each stimulus repeatedly as input to the P-D model. (Naturally, the probability density function of the resulting data set will reflect that of the (Poisson) stochastic transduction process taking place in the hair cells of the cochlea.) However, the model is computationally expensive so, for practical reasons, we have limited this to 50 repetitions. Thus we face (unavoidably) a small-sample size problem.

*** FIGURE 2 ABOUT HERE ***

The stimuli were applied to the P-D model at time $t = 0$ at a simulated level of 65 dB sound pressure level. Activity before $t = 0$ is spontaneous [6]. The model output is best visualized in the form of a neural time-frequency spectrogram or 'neurogram', as depicted in

Figure 2 for the bilabial (/ba–pa/) stimulus with VOT of 40 ms. In this figure, a dot indicates the firing of a neuron: the dot's x and y coordinates are the time of firing and the neuron's CF respectively. To avoid loss of detail in this plot, only some 25,000 spikes are shown, corresponding to approximately the first 30 or 40 of the 50 repetitions. It is important to emphasize that, since the auditory nerve is the only neural pathway from ear to brain, auditory percepts can *only* be encoded in some such form as Figure 2. Damper et al. [4] confirm that the P-D model's responses are an excellent fit to the available physiological data.

4.3 Data Reduction

Neurograms in the form of Figure 2 are not suitable for input to the neural network which is to be trained to categorize 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. Accordingly, to effect the necessary 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. The time limits were chosen to exclude too much irrelevant detail pre-onset or during the steady-state portion of the /a/ vowel. Figure 3 shows a typical such 'reduced' neurogram (in gray-scale form) in response to the bilabial stimulus with 40 ms VOT, from which the extent of the data reduction is obvious.

*** FIGURE 3 ABOUT HERE ***

This data reduction step clearly needs some justification. From a psychoacoustic point of view, a time bin width of 10 ms was chosen as it corresponds approximately to one pitch period (the tokens were synthesized with a 114 Hz fundamental). The frequency bin width was chosen on the following basis. Since the critical bandwidth varies from about 100 Hz at low frequencies to about 1000 Hz at 5000 Hz [30], taking 8 contiguous filters with CFs equispaced on the Greenwood scale corresponds to a frequency bin width which is approximately a constant fraction (about 0.7) of the critical bandwidth. From a statistical pattern processing point of view, by counting spikes, we are trying to estimate a continuous density function from a discrete one. Obviously the more spikes we have the better resolution we can obtain. The (12×16) window represents a trade-off between a very coarse representation with very few cells (where we have a very good estimate of the density) versus a very fine resolution with many cells (but a poor estimate of the density). Hence 192 cells represents a reasonable compromise.

Even with the above justifications, the data reduction might still be criticized as over-crude; in the worst case, effectively negating the use of the detailed auditory model. As well as being essential to make the learning problem tractable, however, it also represents a deliberate attempt to test the notion that gross features are by themselves sufficient to explain the perception of the plosive voicing contrast. Further, in Section 7 below, we show that – even with this severe data reduction – processing by the auditory model is still essential to correct simulation of listening behavior. This strongly implies that its function has not been negated.

5 Modeling CP with Perceptrons

In previous work, we have employed a variety of neural network architectures as synthetic listeners, including brain-state-in-a-box associative networks [8]; competitive-learning networks [9]; multilayer perceptrons (MLPs) [4, 8] and single-layer perceptrons (SLPs) [9]. All of these are capable of mimicking the behaviour of real listeners with more or less fidelity although, overall, MLPs have generally given the best results in terms of modeling human and animal data [8]. However, because of their simplicity, and the fact that non-linear separability is not an issue (see below), we focus here on the single-layer perceptron type of network.

5.1 Network Construction and Training

Three SLPs were constructed: one for each of the three places of articulation (bilabial, alveolar and velar). Each had 192 inputs (as above) and a single output node with signum activation function:

$$f(\mathbf{x}) = \mathrm{sgn}(\mathbf{w}\cdot\mathbf{x}+b) = \begin{cases} 1 & (\mathbf{w}\cdot\mathbf{x}+b) > 0 \\ 0 & (\mathbf{w}\cdot\mathbf{x}+b) \leq 0 \end{cases} \tag{1}$$

where $\mathbf{x}$ is an input vector, $\mathbf{w}$ is the weight vector and b is the bias of the perceptron. Thus, each SLP is a 'hard' classifier. The networks were trained on Rosenblatt's rule [31] to produce a '1' output on the 50 repetitions of the 0 ms VOT responses (auditory patterns from the P-D model), and to produce a '0' output on the 50 repetitions of the 80 ms VOT responses. Hence, the activation value of the output node for unseen patterns can be construed as signifying a hard voiced/unvoiced decision.

We refer to the 0 ms and 80 ms responses as *endpoints* of the VOT continuua. Training on the endpoints in this way – although it could be criticized (see below) – parallels the training of Kuhl and Miller's chinchillas [3]. As in Kuhl and Miller's study, generalization was then tested on the full range (0 ms to 80 ms in 8 steps) of stimuli (50 of each).

5.2 Results

Figure 4 shows typical labeling functions obtained by averaging output activations over the 50 stimulus presentations for each of the three places of articulation. The form of the obtained labeling functions was insensitive to the initial random weight settings for the training. That is, labeling functions like these – with the correct order of boundary shift – were consistently obtained over many repetitions of the training.

*** FIGURE 4 ABOUT HERE ***

The results of Figure 4 closely mimic the labeling functions 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. Taking a mean activation of 0.5 to correspond to the category boundary gives approximate values of 16 ms, 27 ms and 46 ms for the bilabial, alveolar and velar stimuli series respectively, which are reasonably close to those for real listeners. (These are not the best results: as stated above, they are typical.)

*** FIGURE 5 ABOUT HERE ***

To help visualize the agreement, Figure 5 shows composite labeling functions for the alveolar series for humans, chinchillas and a typical single-layer perceptron. The human and animal data are taken from [3]. The root-mean-square (rms) deviations between the SLP's labeling curve and the human and chinchilla curves are 20.69 and 13.05, respectively. This compares to 10.47 for the rms deviation between Kuhl and Miller's human and animal data.

Clearly, the neural model is capturing the 'essence' of categorical perception. 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 obviously very different human, animal and machine listeners.

5.3 Linear Separability of the Patterns

The success of a simple SLP in simulating CP indicates that the auditory patterns are linearly separable. Given the high dimensionality of the input space (relative to the small number of training patterns), this is more or less a direct result of Cover's theorem [32]. (See also pp. 257–8 of [33].) This linear separability has important consequences for the interpretation of the component parts of the model. Since the SLP 'back-end' is a linear system [7], the non-linear segregation of the patterns according to place of articulation can only be a result of transformation by the auditory front-end. This implies that the auditory transformation is essential to the realistic simulation of listening behavior (if the back-end classifier is linear) – as confirmed in Section 7.

5.4 Analysing the Perceptrons

Unlike real (human and animal) listeners, a computational model can be analysed to determine the basis of its behaviour. Early on in their development, neural network models were often criticized unfairly for a lack of transparency, but there are many powerful techniques available for analyzing and making explicit the structured knowledge captured by networks trained to generalize appropriately. In particular, classic work by Oja [34, 35] has shown that a single linear neuron with weights trained by Hebbian learning extracts the first principal component of its input distribution. That is, its weight vector approximates the first eigenvector of the underlying (random) process. This result easily extends to a single layer of neurons allowing principal component analysis of arbitrary degree [36].

*** FIGURE 6 ABOUT HERE ***

In our initial perceptron work, however, we have not used a linear output neuron nor is our training unsupervised (Hebbian). Nonetheless, it is still worthwhile examining the weight vectors, $\mathbf{w}$, to gain insight into what each perceptron has learned. Figure 6 shows a gray-scale depiction of these vectors in elementwise squared form, defined here as:

$$\mathbf{w}^2 \stackrel{\text{def}}{=} (w_1^2, w_2^2, \ldots, w_{192}^2) \tag{2}$$

This form emphasizes the large magnitude components of the weight vector, irrespective of whether they are positive or negative. From Figure 6, where dark cells indicate large magnitudes, we see that the largest-magnitude weights are located around the low-frequency region (the four frequency channels covering CF indices 8 to 48 in the auditory model, corresponding to 73 to 675 Hz) just after acoustic stimulus onset where voicing activity varies maximally as VOT varies. Further, the precise location of this region shifts in the three nets (bilabial, alveolar, velar) in the same way as the boundary point.

To gain further insight into these results, a simplified contribution analysis [37] was conducted by identifying the connections [8] associated with the highest absolute *product* of input and weight, averaged across all 50 presentations of the endpoint patterns and all VOTs. Basing the analysis on this product, rather than just the weight values as above, was felt intuitively to be worthwhile since it is conceivable (if unlikely) that a connection could have a large magnitude weight as a result of training, yet for the corresponding input always to be small in the unseen, test data. The analysis considered positive and negative weight values separately. (Note that the input values are spike counts and are always positive.) Again, highest products of input and *positive* weight were located around the four frequency channels covering CF indices 8 to 48 just after acoustic stimulus onset, and the precise location of this region shifts in the three nets (bilabial, alveolar, velar) in the same way as the boundary point. This, then, confirms the findings when considering weights values only.

In a sense, it is not surprising that the frequency range between 100 and 600 Hz turns out to be key, as this is the region of the perceptually-important first formant ($F1$) transition. To quote Soli [24, p. 2150]:

> "$F1$ onset frequency differences ... exert a substantial influence on discrimination, and, along with other spectral cues provided by source differences at stimulus onset, can account for the discontinuities often reported in research with VOT continua."

Also, the $F1$ transition varies characteristically with place of articulation (see Fig. 3, p. 908, of [3] for a specification of the formant center frequencies of the Abramson and Lisker stimuli as a function of time). Possibly more surprising is that the region of the $F2$ transition (above about 1 kHz for the alveolar and velar stimuli) appears relatively unimportant. However, this transition is more characteristic of the place of articulation. It seems likely that the use of three distinct nets for the three different places of articulation (so that none of the nets has to learn this distinction) is the reason for its relative unimportance in our results.

Averaging the inputs to the five SLP nodes with the largest positive product across all 50 stimuli at each value of VOT produces the pattern in Figure 7 (depicted for the alveolar series). This curve is noticeably similar in shape to the curve for the unmodified net, with its characteristic steep labeling function. In this case, however, there is no thresholding or compression by an activation function as there is in the neural model, so indicating that categorization behaviour is not merely a trivial consequence of specific details of the network architecture. Similar findings obtain for the bilabial and velar series.

*** FIGURE 7 ABOUT HERE ***

The corresponding plot focusing on the large magnitude negative weights does not reproduce this pattern of variation with VOT: it is essentially flat. We conjecture that the role of the negative weights is simply to provide an 'offset', reducing the labeling function to 0 as necessary in spite of all the inputs (spike counts) being positive. These negatively-weighted lines can connect to any region of the neurogram where there is significant activity which remains more or less constant as VOT changes. Generally, this is the region of high CF and the period some time after stimulus onset.

The implication of these results – taken at face value – is that categorization can be explained in terms of a mechanism by which higher levels of the auditory system focus on a particular region of auditory nerve time-frequency activity and, in essence, aggregate spike activity in this region. But the perceptron has some shortcomings as a learning system which mean that these findings need to be treated with caution.

5.5 Shortcomings of the Perceptron

Caution is required in interpreting the above results for several reasons:

1. We have very sparse training data so the question arises: are they sufficient? Several authors (notably Baum and Haussler [38]) have considered the bounds on the required number of training examples to produce valid generalization for nets of a given size [9], but (because of the unrealistic theoretical assumptions made) these are generally "too loose, leading to impractical results" [39, p. 238]. A much-cited rule of thumb (due to Widrow [40]) is that there should be about 10 times as many training examples as adjustable parameters, implying a need in this work for about 1930 training instances [10]. However, we have only 100 (2 endpoints $\times$ 50 repetitions) auditory patterns. Hence, we confront a problem for statistical learning of small sample size. In particular, it is difficult or impossible to use techniques for controlling overlearning (e.g. cross-validation) which rely on partitioning the available data into disjoint sets.

2. The sampling statistics of the training data necessarily reflect the (Poisson) statistics of the mechanical-to-neural transduction process taking place in the hair cells of the cochlea. This will tend to produce training data clustered around the expected (mean) value of the auditory responses to the endpoint stimuli. Thus, many of the training data points may be very similar. From the point of view of learning theory, we would prefer to give most credence to that subset of the data points which best discriminates the voiced and unvoiced categories. These are unlikely to be data points close to the class means.

3. There is no explicit control of generalization in perceptron learning. Thus, the solution obtained is not unique. By contrast, as we shall see, support vector machines solve a constrained optimization problem leading to a unique solution, which implies better generalization.

4. The technique of supervised training on the endpoints inevitably predisposes the nets to produce something very close to the 'correct' 1/0 voiced/unvoiced values at extreme VOTs. Although we have addressed this concern elsewhere – showing that appropriate categorization behaviour is maintained when a single-layer, two-output network is trained without supervision using competitive learning [41] to dichotomise the endpoint data [9] – we remain conscious of the problem. In particular, it means that each net may be doing no more than simply placing the boundary at the mid-point of VOT between endpoint exemplars, albeit in a 192-dimensional space.

However, mitigating all these objections, we emphasize that the perceptrons do actually produce very realistic behaviours and do so very consistently. This implies that the issues detailed above are far from fatal to the aim of extracting phonetic knowledge from the trained perceptrons. Nonetheless, it seems worthwhile and prudent to overcome as many of these shortcomings as possible. Here we address concerns 1 to 3. Thus, we treat the perceptron results outlined above as essentially preliminary, and seek to confirm them using the modern inductive inference technique of support vector machines [12, 13, 42].

6 Modeling CP with SVMs

To address some of the shortcomings of perceptrons, notably the lack of a unique solution, we apply the technique of support vector machines (SVMs) [12, 13, 42]. Readers unfamiliar with the background are referred to Chapter 6 of [33] for an excellent tutorial introduction. SVMs incorporate the structural risk minimization (SRM) principle, which is derived from the theory of small sample sizes. Earlier work in pattern recognition was based on the principle of *empirical risk minimization*, according to which one tries to reduce the occurrence (or 'risk') of classification errors when learning from empirical data. When the amount of empirical data is small, however, the learning problem becomes increasingly ill-posed. In structural risk minimization, another term is added to the function to be minimized which measures the complexity of the machine implementing the learned solution, so as to control against overfitting to the data. SRM is thus closely related to regularization [43, 44]. According to Haykin [33, p. 101]: "The challenge in solving a supervised learning problem is ... to realize the best generalization performance by matching the machine capacity to the available training data. SRM does this by making some measure of machine capacity a control variable". Thus, in addition to the requirement of correct classification, a further constraint is added which maximizes the *margin*, i.e. maximizes 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 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 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:

$$\rho(\mathbf{w}, b) = \min_i\{d(\mathbf{w}, b; \mathbf{x}_i)\} + \min_j\{d(\mathbf{w}, b; \mathbf{x}_j)\} \qquad \mathbf{x}_i \in A,\ \mathbf{x}_j \in B \tag{3}$$

Maximizing ρ with respect to $\mathbf{w}$ and b produces good control of the generalization ability of the learning machine that is absent from the perceptron. Furthermore, it guarantees a unique solution to the problem, in contrast to perceptron learning. Rosenblatt's algorithm [31] merely aims to find a separating hyperplane – *any* separating hyperplane.

The quadratic programming optimization was done using Mészáros' BPMPD package [45]. Optimization was performed in the (192-dimensional) input space, i.e. assuming linear separability of the patterns as in subsection 5.3. If we had a lot more data, we could try to model the decision function more accurately and test the hypothesis that a non-linear discriminant would be more appropriate. This is easily achieved within the SVM framework by replacing the inner product in the input space [11] with a kernel inner-product function induced in the feature space [33, pp.330–1].

6.1 Hard Classification with SVMs

Three SVMs were constructed: one for each of the three places of articulation (bilabial, alveolar and velar). As for the perceptron experiments, the networks were trained using the same 100 patterns: 50 repetitions of responses to the 0 ms VOT stimuli and 50 repetitions of responses to the 80 ms VOT stimuli. A straightforward SVM was used with an architecture equivalent to a perceptron [13, 46, 47], with a hard-limiting (signum function) threshold unit on the output (equation 1). There was no additional capacity control. The auditory data are easily separated with a linear hyperplane in the 192-dimensional feature space; as such, no additional capacity control to tolerate errors in overlapping classes was necessary.

6.1.1 Results

The results with this hard classifier are shown in Figure 8. Each of the three curves depicts the average classification [12] over each of the 50 repetitions for the relevant series. Taking the 50% mid-point 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. These compare very well to the values obtained for perceptron training (and for real listeners too).

Referring back to Figure 5, the labeling function for the alveolar SVM has been plotted against Kuhl and Miller's data for humans and chinchilla's, as well the corresponding result for a typical SLP. Again, the SVM convincingly mimics the labeling curves for real listeners – better, if anything, than the SLP does. The rms deviations between the SVM labeling curve and the human and animal data are 12.53 and 9.58, respectively (cf. 20.69 and 13.05

for the SLP).

*** FIGURE 8 ABOUT HERE ***

6.1.2 Analysing the Hard SVM Classifiers

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 are really conveying the vital information about the category boundary. They will lie on the boundary of the maximized margin (Figure 9). The two margin boundaries (one for each class) are parallel, and the optimal separating hyperplane 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 hyperplane(s) (i.e. the weight vector $\mathbf{w}$) for this purpose. This 192-dimensional vector uniquely characterizes 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 stimuli respectively, divided roughly half and half between the voiced and unvoiced categories.

*** FIGURE 9 ABOUT HERE ***

To visualize the information extracted by the model, the 192-dimensional weight vectors $\mathbf{w}^2$ (see equation 2) are depicted for the three stimuli series in Figure 10. As before, the squaring operation emphasizes the magnitude of the information which differentiates 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 which parallels the finding with the perceptrons. 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 perceptrons and for real listeners.

*** FIGURE 10 ABOUT HERE ***

In our opinion, the results in Figure 10 for the SVMs are more plausible than those of Figure 6 for the perceptrons because they depict a much smoother function in the time-frequency space [13]. Figure 10 clearly confirms the importance of the highly-localized $F1$ transition region close to the category boundary, as found previously for the perceptrons and the spike-counting model. Additionally, it now becomes clear that the spectrum at stimulus onset is 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 in all three cases.

*** FIGURE 11 ABOUT HERE ***

In an attempt to gain an appreciation of the time-frequency features distinguishing the place of articulation, we have also examined the differences between the weight vectors for the three SVMs. Figure 11 shows $(\mathbf{w}_{bil} - \mathbf{w}_{alv})^2$, $(\mathbf{w}_{bil} - \mathbf{w}_{vel})^2$ and $(\mathbf{w}_{alv} - \mathbf{w}_{vel})^2$, where the squaring operation is elementwise as in equation 2 and the subscripts denote the particular classifier (one for each stimuli series). Unfortunately, this does not appear to be very revealing. A clearer insight into the features underlying the place of articulation distinction awaits the analysis of a net trained to make this distinction (or of a single net trained on all three stimuli series).

6.2 Soft Classification with SVMs

The support vector classifier is a hard classifier, and it could be argued that this is inappropriate for modeling the VOT transition region [14]. To address this objection, we consider a modified output activation function:

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

where:

$$h(v) = \begin{cases} -1 & v < -1 \\ v & -1 \leq v \leq 1 \\ 1 & v > 1 \end{cases}$$

which corresponds to a linear interpolation within the margin. Hence, outside of the margin we maintain a hard classifier but inside, where we have no training data, we employ a soft classification scheme.

*** FIGURE 12 ABOUT HERE ***

The results corresponding to Figure 8 are reproduced in Figure 12 using this modified activation function. Again, category boundaries are essentially identical to those for the perceptron and the hard classifier. However, the slope is much too shallow to be realistic.

7 Role of the Auditory Model

Thus far, we have used a sophisticated front-end capable of essentially perfect reproduction of detailed neural firing patterns. This raises an important question: how much of this sophistication is actually necessary? Just how simple can our composite (auditory front-end and back-end learning system) model be yet still produce a convincing simulation of human and animal listening behavior? 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 – consistent with performing "some kind of frequency analysis" (see the quote from Rosen and Howell above). Thus, we use short-time Fourier analysis. This is admittedly a worst-case scenario: the ear is far from being a linear Fourier analyzer.

Each of the 3×9 stimuli were pre-processed by fast Fourier transform (FFT). That is, the power spectral densities of the stimuli were computed using a 256-point FFT over 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 in the time dimension but from 0 to 5 kHz in steps of 312.5 Hz in the frequency dimension. So, here, the frequency dimension is divided up linearly (in Hz) rather than approximately logarithmically according to CF as earlier.

Note that in this case there is 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. As the intention here is to see if a greatly simplified model can still mimic categorization behavior convincingly, we feel this is justified. Further justification comes from training perceptrons on single, *averaged* neurograms. Appropriate categorization behavior is maintained [7], indicating that information about the statistical distribution of training data it is not essential to the simulation, and that the extreme sparsity of the training data need not be fatal.

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-analysed endpoints. A hard classification rule was used. Figure 13 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 this effect. It is easy to see that this *must* be so. Since the SLP and the SVM are both linear (but see note 7), the non-linear segregation of the stimuli by place of articulation can only result from processing by the auditory front-end.

*** FIGURE 13 ABOUT HERE ***

We speculate that the auditory model is essential to the simulation for two main reasons. First, the frequency warping in accordance with the Greenwood scale makes the important $F1$ region more prominent in the auditory representation than in the acoustic representation. Second, the onset enhancement by the hair cell model makes the onset of voicing more prominent. Without the enhancement effected by the auditory transformation, the learning system cannot learn the voicing distinction. The stochastic component due to mechanical-to-neural transduction by the cochlear hair cells seems not to be involved as argued above. We do, however, need to confirm these speculations in future work.

8 Fisher Linear Discriminant Analysis

Both perceptrons and support vector machines have similarities to Fisher linear discriminant analysis (FLDA) – see Chapters 4 and 5 of [48] and pp. 201–2 of [33]. Our initial feeling was that the differences between the SVM approach and FLDA are important. Both solve a constrained optimization problem. However, the Fisher discriminant tries to find a hyperplane such that data from different classes are maximally separated (in terms of their common variance) when projected onto it, while the SVM approach tries to maximize the margin. As such, the latter is concerned with the points which lie on the boundary, and does not make assumptions about the distribution of the data within the input space. FLDA assumes that classes can be approximated with Gaussians (or that the distribution can be appropriately parameterized) and so uses all the data in constructing the discriminant function. Contrast this with the SVM which in our case only uses some 40% of the data in constructing the discriminant function.

*** FIGURE 14 ABOUT HERE ***

Figure 14 illustrates the differences between FLDA and the SVM solution for a simple two-feature example. Also shown is any one of the possible infinity of perceptron solutions. The SVM hyperplane is formed with reference to just three support vectors whereas the Fisher discriminant reflects the statistical distribution of the data.

*** FIGURE 15 ABOUT HERE ***

Figure 15 shows labeling curves obtained from FLDA for the three stimuli series. That is, the discriminant ($\mathbf{w}$, b) was computed as:

$$\begin{aligned} \mathbf{w} &= \Sigma^{-1}(\mathbf{m}_{voiced} - \mathbf{m}_{unvoiced}) \\ b &= -\frac{1}{2}(\mathbf{m}_{voiced} + \mathbf{m}_{unvoiced}) \cdot \mathbf{w} \end{aligned}$$

where Σ is the sum of the individual covariance matrices (i.e. computed for each class) and $\mathbf{m}_{voiced}$ and $\mathbf{m}_{unvoiced}$ are the class means. As each class is assumed equally likely, the bias (b) was computed to pass through the mean of the class means. Then, using a hard classification rule (equation 1), the complete range of VOT values was tested, just as for the SVMs. Results are noticeably poorer than for the SVM; they are more like the perceptron results.

*** FIGURE 16 ABOUT HERE ***

Figure 16 shows a gray-scale depiction of the squared weight vectors of the Fisher linear discriminant for the three stimuli series. It is abundantly plain that FLDA fails to capture knowledge about the features distinguishing voiced from unvoiced tokens in anything like the vivid way that support vector machines do.

9 Conclusions

The categorization of syllable-initial stops into voiced and unvoiced categories has been well studied in human, and even animal, listeners over many years. More recently, attention has turned to the perception of such speech (or synthetic speech) sounds by machine. We have previously shown that a variety of computational learning systems is capable of mimicking the categorization 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 just not possible with real listeners. This is a specific instance of Zipser's "neural system identification". We stress that the full range of perceptual phenomena associated with the categorization 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 manifestations of CP.

In this paper, we have described the use of single-layer perceptrons (SLPs) to simulate the categorization behaviour of real listeners. Visualization of the learned weight vectors of each perceptron and analysis of the averaged activity in their input-output connections reveals that highly-localized 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.

The perceptron, however, was the earliest computational learning device, dating back more than 40 years. In the intervening spell, several shortcomings of the now classical 'neural network' paradigm have become apparent. Given these shortcomings, the phonetic knowledge extracted from the perceptron has to be treated with caution. In particular, we have very little training data (100 instances) from which to estimate the 193 adjustable parameters of the perceptron model. A major goal of this work was to confirm (or otherwise) the findings with the SLPs using the modern inductive inference methodology of support vector machines (SVMs), based on the principle of structural risk minimization, which is founded on small-sample statistics and overcomes many of the objections of the earlier paradigm.

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 (OSH) to represent the essential information about the voiced/unvoiced distinction. This parallels the way that the knowledge extracted by the perceptron was visualized, but the results are much more vivid. 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.

The SVM approach of maximizing the margin leads to a unique solution, since there is only one hyperplane which maximizes the minimum distance from the hyperplane to marginal data points from each of the two categories. This implies a degree of consistency which is unobtainable in the classical neural network paradigm where, typically, training has to be repeated with a large number of different, random starting configurations in order to achieve any kind of statistical reliability. In this work, the perceptron results were actually very consistent over different trials (in terms of order of boundary shift if not the actual boundary value), but it was nonetheless necessary to perform these multiple trials to confirm this fact. By contrast, the SVM methodology requires no such trial-and-error procedure. Categorical perception has been modeled using both hard and soft SVM classifiers. While the latter seemed initially to be more appropriate, the hard scheme actually produces better results (i.e. closer to labeling curves from real listeners).

The auditory front-end used here is rather complex. In an attempt to see to what extent it could be simplified, the front-end was replaced by a more prosaic short-time Fourier analysis. This abolished the correct movement of the boundary with place of articulation, indicating that some aspect or aspects of peripheral auditory function 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. We also used a Fisher linear discriminant analysis as the back-end learning system. This produced results consistent with – but rather poorer than – the perceptrons.

How do our results support or deny extant theories of CP? Analysis of both perceptrons and 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. In addition, 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. Indeed, Figure 7 depicts just such a discontinuity in mean spike-count from a localized time-frequency area of the neurogram as VOT is varied. Thus, there is no need to posit any speech-specific mechanisms. The fact that the auditory front-end was essential to correct modeling also points to an auditory rather than acoustic explanation (cf. [24]). Since our simulations are based on learning systems as the back-end, we are also inclined to believe that learning plays an important part.

The work reported in this paper suggests several potentially fruitful avenues for future research. The Lisker and Abramson stimuli were studied here because of their prior use in key studies of CP, and because synthesis allows easy production of ‘ambiguous’ tokens (close to the relevant category boundary). A valid criticism is that the relation of these stimuli to real speech is tenuous. Hence, it is a priority to study real speech in the near future. Further, Lisker and Abramson showed that Thai listeners make a three-way distinction along the VOT continuum, as opposed to the two-way distinction in English. Accordingly, it would be interesting to model the perception of native speakers/listeners of languages other than English. 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. Finally, we will employ a variety of simplified front-end analyses to determine those aspects of the peripheral auditory transformation which are essential to simulating boundary movement with place of articulation.

The key finding of this paper is that the crucial information which distinguishes voiced from unvoiced tokens is restricted to a highly-localized region in time-frequency space. This has been inferred from analysis of a variety of (bottom-up) learning systems, of which the support vector machine produced the model most consistent with human and animal listeners. Suppose, instead, that the importance of this region was a prior assumption. We might then construct a (top-down) parametric model to explain categorization. However, this would leave out of account the other features that the SVM discovered, namely the spectrum at stimulus onset and the $F1$ region at vowel onset. This vividly illustrates the benefits of extracting knowledge from learning systems.

Notes

[1]... and we know from previous work that many learning systems are.

[2]Voice-onset time describes the point in time at which vocal cord vibration starts relative to the release of vocal tract closure [49].

[3]According to Crystal [49], *place of articulation* is “one of the main parameters in the phonetic classification of speech, referring to where in the vocal apparatus a sound is produced”. Thus, a bilabial sound is produced by using both lips to stop the flow of air from the vocal tract before releasing the obstruction; a velar sound is produced by placing the tongue against the velum to obstruct air flow.

[4]Note that smooth curves have been fitted to the raw data by probit analysis. This should be borne in mind when interpreting the labeling curves to be presented shortly, where such smoothing has not been done.

[5]The original P-D model includes simulations of cochlear nucleus processing but, for the purposes of this work, these are ignored and outputs are taken from the auditory-nerve level of the model.

[6]... as is that in the first 8 frequency channels, because the first 8 filters of the auditory filterbank are very broadly-tuned, down to and including dc in fact. For this reason, the output from these filters was ignored in determining the driven responses.

[7]... apart from the non-linear activation function of the output node. This, however, cannot produce segregation of the labeling functions if they are not segregated already.

[8]Sanger's original work was concerned with the contribution to network behaviour of hidden nodes, but the same concepts are directly transferable to the contributions of weighted connections.

[9]Usually feedforward nets with hidden layers are considered.

[10]There are 192 weighted connections from the inputs and a variable threshold on the output unit making a total of 193 adjustable parameters.

[11]Since weights are computed from training data, $(\mathbf{w} \cdot \mathbf{x})$ is seen to be an inner product in input space.

[12]The SVM formulation requires the classes to be labeled $(-1, +1)$. For comparability with other results, however, these were transformed to (0, 1) labels prior to graphing.

[13]This was verified by examining the (unsquared) weight vectors.

[14]In very early work on neural network models of CP, Anderson and his colleagues [50, 51] used a hard classifier: the brain-state-in-a-box. It was necessary to add noise to the input to model a finite transition between categories with any sort of realism. See [8] for criticism of this artifice.

References

[1] J. S. Perkell and D. H. Klatt, editors. *Invariance and Variability in Speech Processes*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1986.

[2] A. Q. Summerfield. Differences between spectral dependencies in auditory and phonetic temporal processing: Relevance to the perception of voicing in initial stops. *Journal of the Acoustical Society of America*, 72:51–61, 1982.

[3] 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.

[4] 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.

[5] 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.

[6] R. I. Damper. A biocybernetic simulation of speech perception by humans and animals. In *Proceedings of IEEE International Conference on Systems, Man and Cybernetics*, volume 2, pages 1638–1643, Orlando, FL, 1997.

[7] R. I. Damper. Auditory representations of speech sounds in a neural model: The role of peripheral processing. In *Proceedings of International Joint Conference on Neural Networks (IJCNN '98)*, pages 2196–2201, Anchorage, AL, 1998.

[8] R. I. Damper and S. Harnad. Neural network models of categorical perception. *Perception and Psychophysics*, in press.

[9] 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*.

[10] A. M. Darling, M. A. Huckvale, S. Rosen, and A. Faulkner. Phonetic classification of the plosive voicing contrast using computational modelling. *Proceedings of the Institute of Acoustics*, 14(6):289–295, 1992.

[11] F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65:386–408, 1958.

[12] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, NY, 1995.

[13] C. Cortes and V. N. Vapnik. Support vector networks. *Machine Learning*, 20:1–25, 1995.

[14] F. Crick. *The Astonishing Hypothesis*. Simon and Schuster, London, UK, 1994.

[15] D. Zipser. Identification models of the nervous system. *Neuroscience*, 47(4):853–862, 1992.

[16] K. Popper. *The Logic of Scientific Discovery*. Basic Books, New York, NY, 1965.

[17] A. Clark. Connectionism, competence and explanation. *British Journal for the Philosophy of Science*, 41:195–222, 1990.

[18] 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.

[19] C. C. Wood. Discriminability, response bias, and phoneme categories in discrimination of voice onset time. *Journal of the Acoustical Society of America*, 60:1381–1389, 1976.

[20] 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.

[21] P. K. Kuhl. The special-mechanisms debate in speech research: Categorization tests on animals and infants. In S. Harnad, editor, *Categorical Perception: the Groundwork of Cognition*, pages 355–386. Cambridge University Press, Cambridge, UK, 1987.

[22] S. Harnad, editor. *Categorical Perception: the Groundwork of Cognition*. Cambridge University Press, Cambridge, UK, 1987.

[23] 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.

[24] 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.

[25] 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.

[26] Z. B. Nossair and S. A. Zahorian. Dynamic spectral shape features as acoustic correlates for initial stop consonants. *Journal of the Acoustical Society of America*, 89:2978–2991, 1991.

[27] 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.

[28] D. D. Greenwood. Critical bandwidth and the frequency coordinates on the basilar membrane. *Journal of the Acoustical Society of America*, 33:780–801, 1961.

[29] D. D. Greenwood. A cochlear frequency-position function for several species – 29 years later. *Journal of the Acoustical Society of America*, 87(6):2592–2605, 1990.

[30] E. Zwicker and E. Terhardt. Analytical expressions for critical-band rate and critical bandwidth as a function of frequency. *Journal of the Acoustical Society of America*, 68(5):1523–1525, 1980.

[31] F. Rosenblatt. On the convergence of reinforcement procedures in simple perceptrons. Technical Report VG-1196-G-4, Cornell Aeronautical Laboratory Report, Buffalo, NY, 1960.

[32] T. M. Cover. Statistical and geometrical properties of systems of inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, EC-14:326–334, 1965.

[33] S. Haykin. *Neural Networks: A Comprehensive Foundation*. Prentice-Hall, Upper Saddle River, NJ, 1999.

[34] E. Oja. A simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology*, 15:267–273, 1982.

[35] E. Oja. Neural networks, principal components, and subspaces. *International Journal of Neural Systems*, 1:61–68, 1991.

[36] T. D. Sanger. An optimality principle for unsupervised learning. *Neural Networks*, 12:459–473, 1989.

[37] D. Sanger. Contribution analysis: A technique for assigning responsibilities to hidden units in connectionist networks. *Connection Science*, 1:115–138, 1989.

[38] E. B. Baum and D. Haussler. What size net gives valid generalization? *Neural Computation*, 1:151–160, 1989.

[39] N. K. Bose and P. Liang. *Neural Network Fundamentals with Graphs, Algorithms and Applications*. McGraw-Hill, New York, NY, 1996.

[40] 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.

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

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

[43] A. N. Tikhonov. On regularization of ill-posed problems. *Doklady Akademii Nauk USSR*, 153:49–52, 1973.

[44] A. N. Tikhonov and V. Y. Arsenin. *Solutions of Ill-Posed Problems*. W. H. Winston, Washington, DC, 1977.

[45] 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.

[46] V. N. Vapnik and A. J. Chervonenkis. A note on the class of perceptrons. *Automation and Remote Control*, 25:103–109, 1964.

[47] V. N. Vapnik and A. J. Chervonenkis. *Theory of Pattern Recognition*. Nauka, Moscow, USSR, 1974. (In Russian).

[48] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. Wiley, New York, 1973.

[49] D. Crystal. *A First Dictionary of Linguistics and Phonetics*. André Deutsch, London, 1980.

[50] J. A. Anderson, J. W. Silverstein, S. A. Ritz, and R. S. Jones. Distinctive features, categorical perception, and probability learning: Some applications for a neural model. *Psychological Review*, 84:413–451, 1977.

[51] J. A. Anderson. Cognitive and psychological computation with neural models. *IEEE Transactions on Systems, Man and Cybernetics*, 13:799–815, 1983.

Author Biographies

Bob Damper was born in Tunbridge Wells, England, in 1948. He obtained his MSc in biophysics in 1973 and PhD in electrical engineering in 1979, both from the University of London. He also holds the Diploma of Imperial College, London, in electrical engineering. He was appointed Lecturer in electronics at the University of Southampton in 1980, Senior Lecturer in 1989, and Reader in 1998. His research interests include speech recognition and synthesis, neural computing, pattern recognition and intelligent systems engineering. Dr. Damper heads the Image, Speech and Intelligent Engineering research group in the Department of Electronics and Computer Science at Southampton. He has published approximately 190 articles and authored the undergraduate text *Introduction to Discrete-Time Signals and Systems*. He is currently Chairman of the Signal Processing Chapter of the UK and Republic of Ireland Section of the IEEE, and is a Chartered Engineer and a Chartered Physicist, a Fellow of the Institution of Electrical Engineers, a Fellow of the Institute of Physics, a member of the Acoustical Society of America, a member of the European Speech Communication Association and a member of the Association for Computing Machinery's special interest group on computational phonology.

Steve Gunn was born in Bristol, England, in 1970. He received his BEng in electronic engineering (1992) and PhD in computer vision (1996) from the University of Southampton. His post-doctoral work investigated intelligent modelling algorithms. He is currently employed as a Research Lecturer within the Image, Speech and Intelligent Systems group at the University of Southampton. His research interests include computer vision, active contours, support vector machines and neural networks.

Mat Gore was born in Leeds, England, in 1973. He graduated with an honours degree in Computer Science from the University of Bristol in 1995. He then worked for two years at the University of Southampton researching computational techniques to model aspects of speech perception. Mat Gore is currently working in IT development for Credit Suisse First Boston in London.

List of Figures

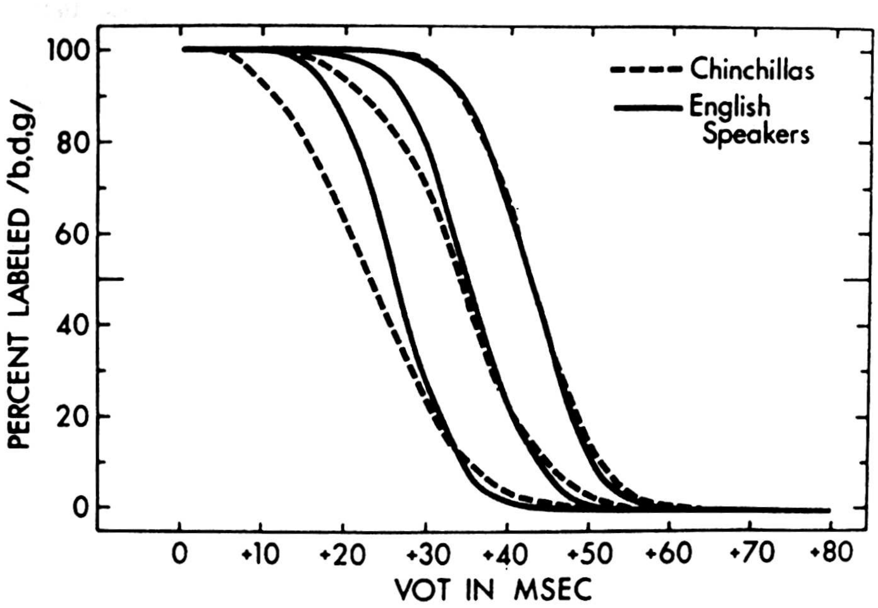

Figure 1: Labeling curves for syllable-initial stop consonants varying in voice-onset time (VOT) for human and chinchilla listeners. Smooth curves have been fitted to the raw data points by probit analysis. Reprinted with permission from Kuhl and Miller, "Speech perception by the chinchilla: Identification functions for synthetic VOT stimuli", *Journal of the Acoustical Society of America*, 63(3), March 1978, 905–917. Copyright 1978 Acoustical Society of America.

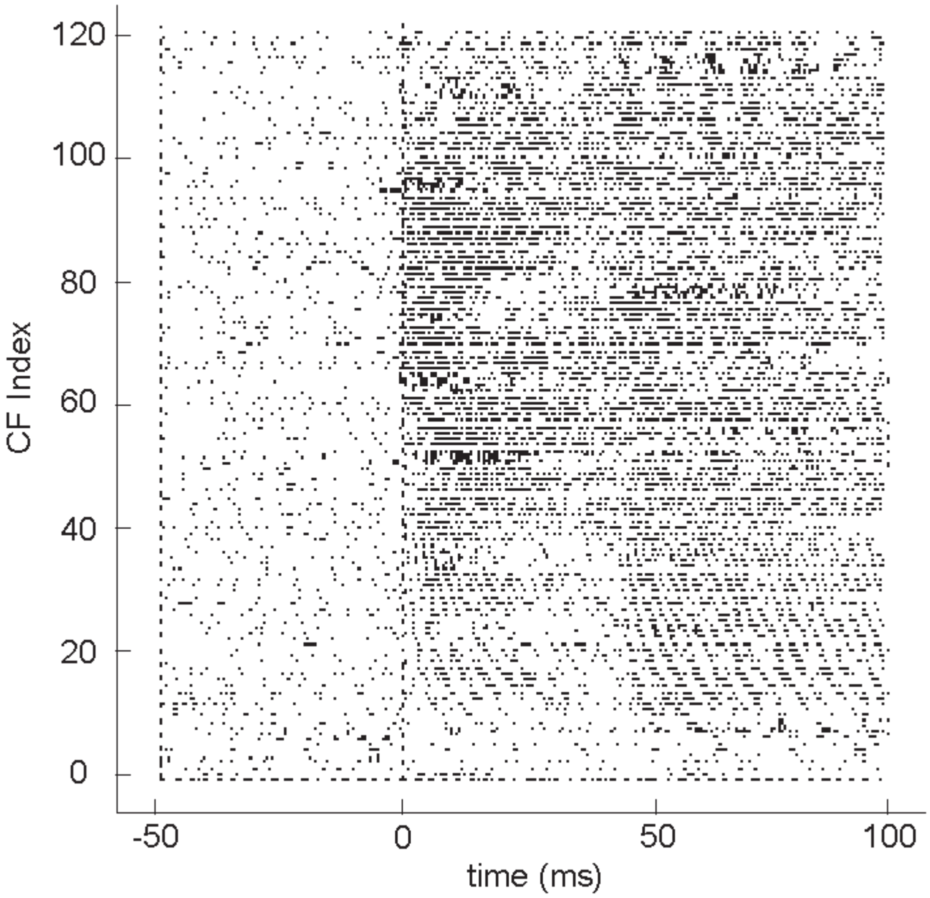

Figure 2: Response of the P-D model to the bilabial stimulus with 40 ms VOT in the form of a neurogram. Each dot depicts the firing of a neuron at the indicated CF and time. Stimulus onset is at 0 ms. CF index of 1 corresponds to 50 Hz and CF index of 128 to 5 kHz. Activity before stimulus onset is spontaneous, as is that in CF channels 1 to 8 (because the dc gains of the auditory filters with low CF are too large to be ignored).

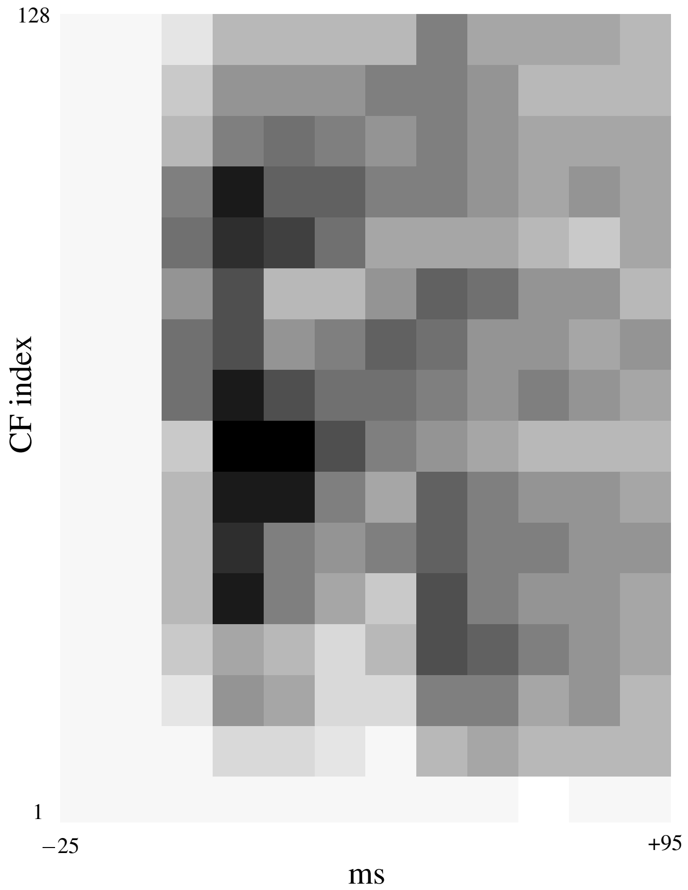

Figure 3: Typical 'reduced' neurogram in the (12 × 16) matrix form presented to the neural networks: bilabial stimulus, 40 ms VOT. There is very significant data reduction relative to the representation which retains the CF and time identity of each spike as in Fig. 2.

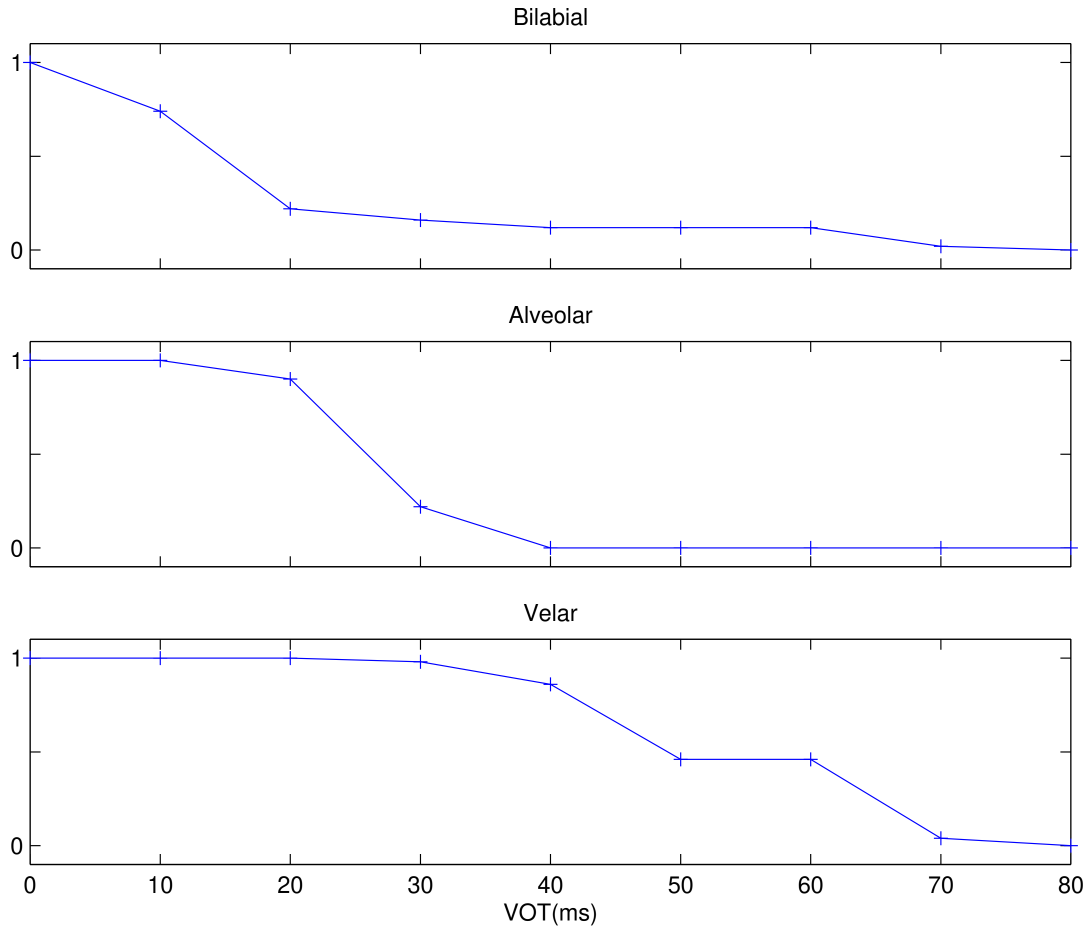

Figure 4: Mean output activation versus VOT for single-layer perceptrons with hard classification, trained on neurograms from 0 ms and 80 ms endpoints. The phoneme boundary movement with place of articulation is convincingly simulated.

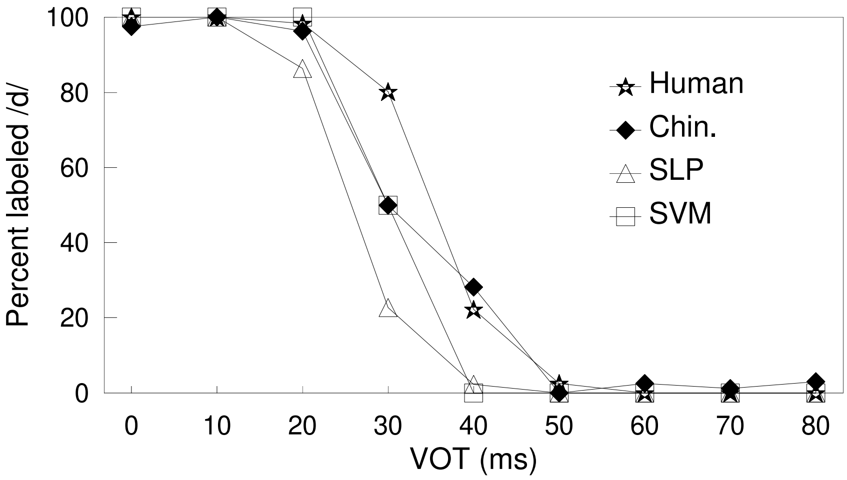

Figure 5: Composite labeling functions for the alveolar series for humans, chinchillas, typical single-layer perceptron and support vector machine. The human and animal data are taken from Fig. 3 of reference [3].

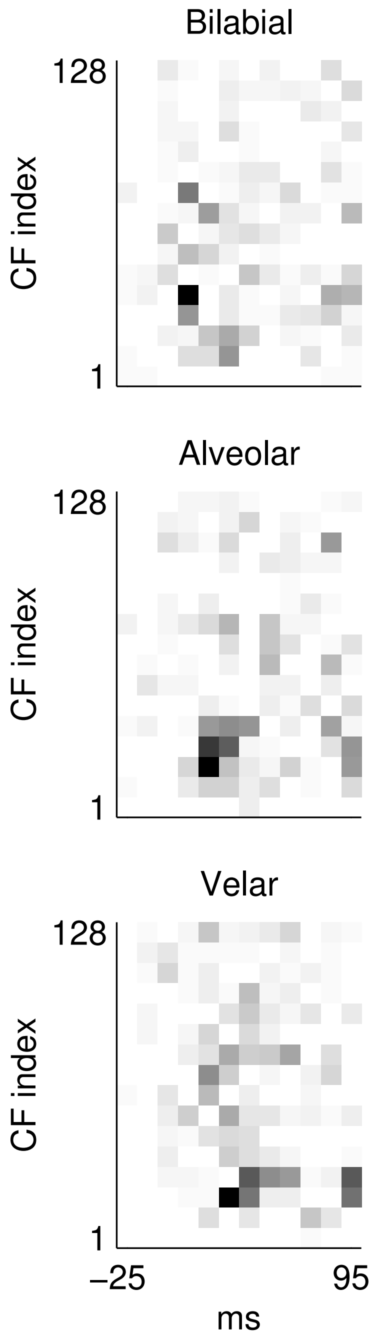

Figure 6: Gray-scale depiction of the squared weight vectors of each perceptron, $\mathbf{w}^2$, for the three stimuli series.

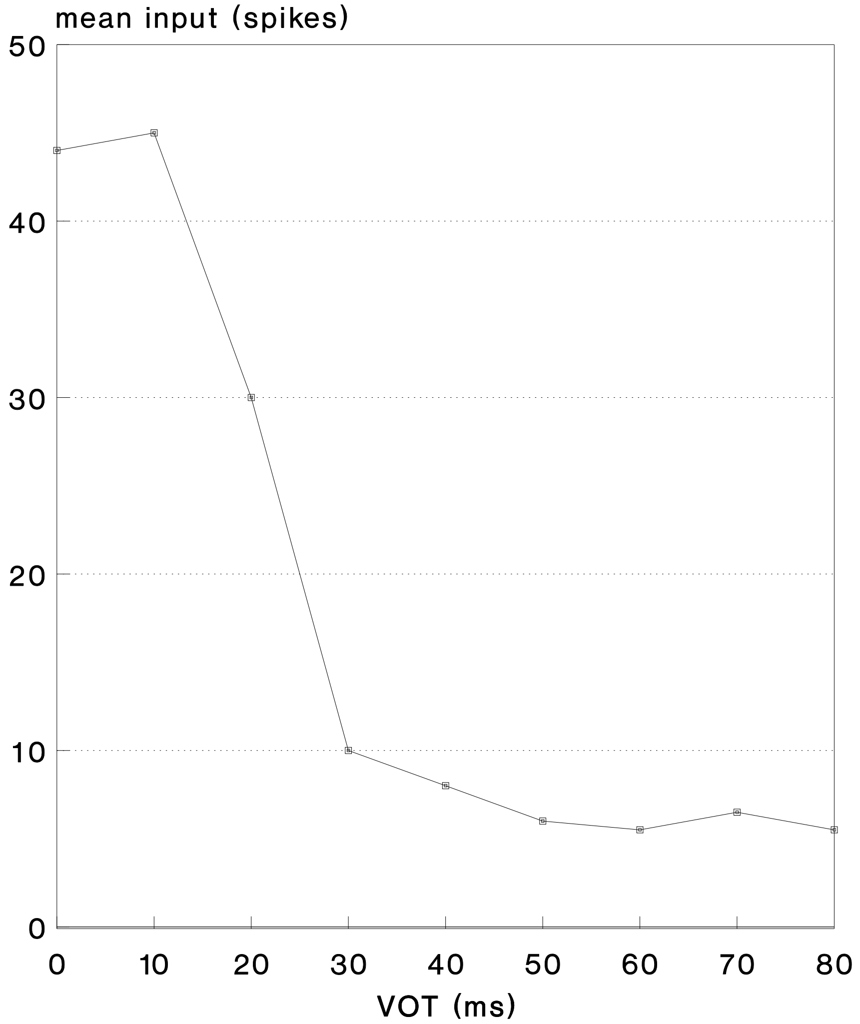

Figure 7: Mean spike-count input for the five SLP nodes with maximal product of input and *positive* weight versus VOT for the alveolar series.

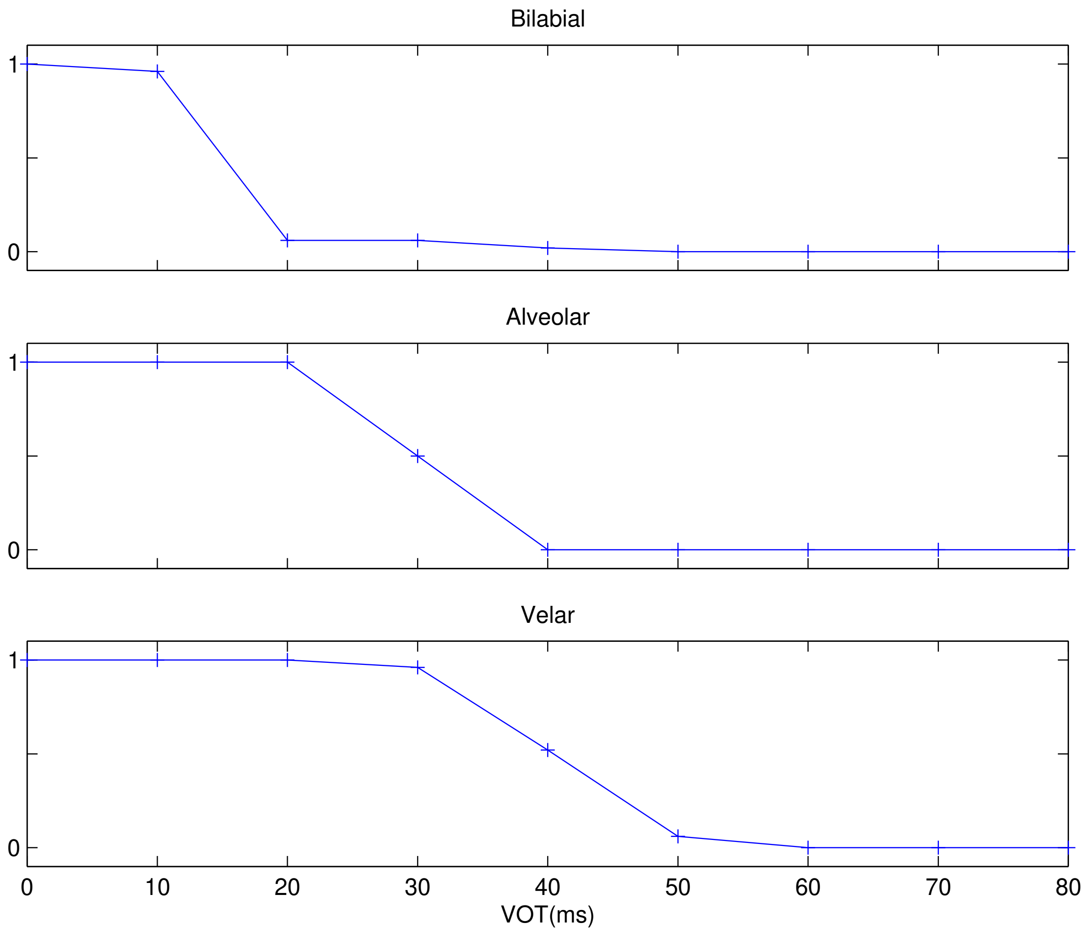

Figure 8: Labeling curves for the hard SVM classifier. Category boundaries are essentially identical to those for the perceptron.

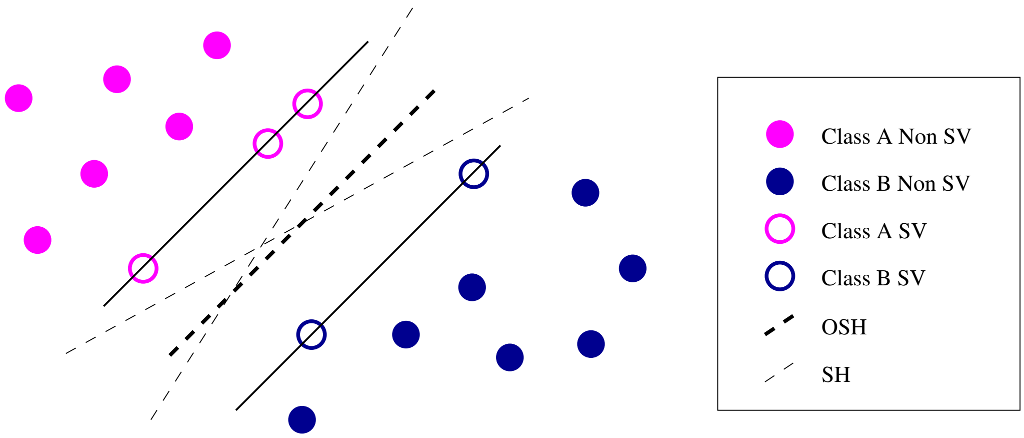

Figure 9: Simplified (two-dimensional) visualization of the support vector classification technique for a two-class (A, B) problem. Support vectors (SVs) lie on the margin. In the separable case, there are no training data in the region between the two margins. The optimal separating hyperplane (OSH – shown dotted in bold) is parallel to both margins, and equidistant from them. We use the normal vector to these three hyperplanes as the basis of knowledge extraction. There are other separating hyperplanes (SHs) but these are not optimal in the sense of satisfying equation (3), and would not be expected to generalize as well.

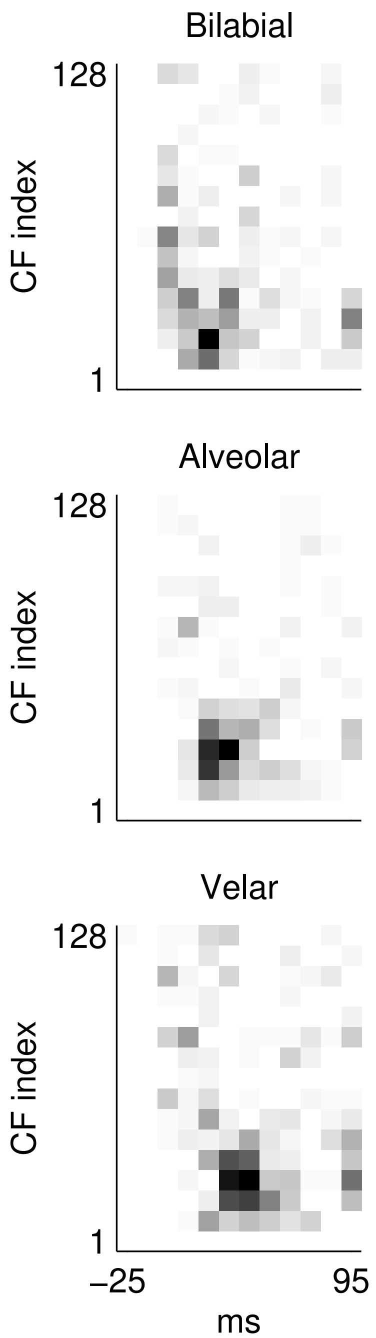

Figure 10: Gray-scale depiction of the squared weight vectors of the hard SVM, $\mathbf{w}^2$, for the three stimuli series.

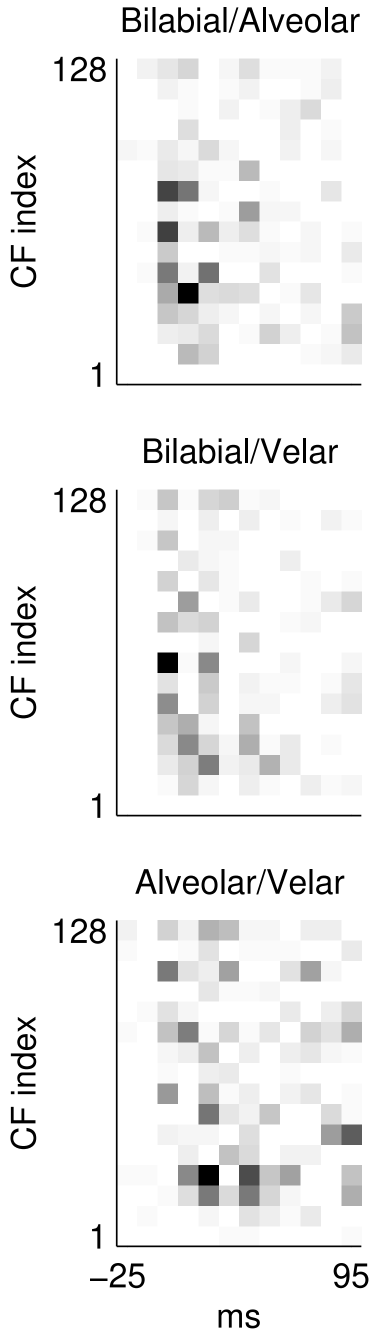

Figure 11: Gray-scale depiction of the squared differences between weight vectors of the hard SVM for the three stimuli series.

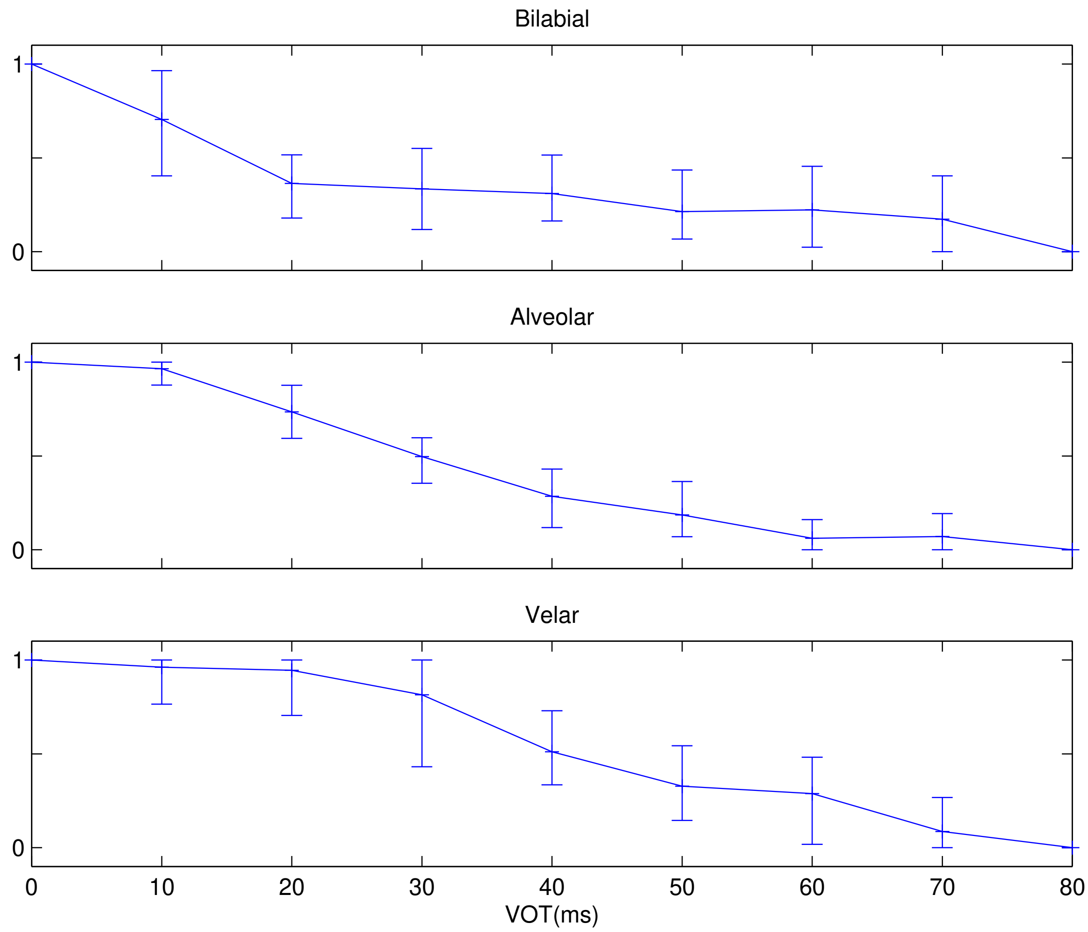

Figure 12: Labeling curves (with error bars) for the soft SVM classifier. Again, category boundaries are essentially identical to those for the perceptron and the hard classifier.

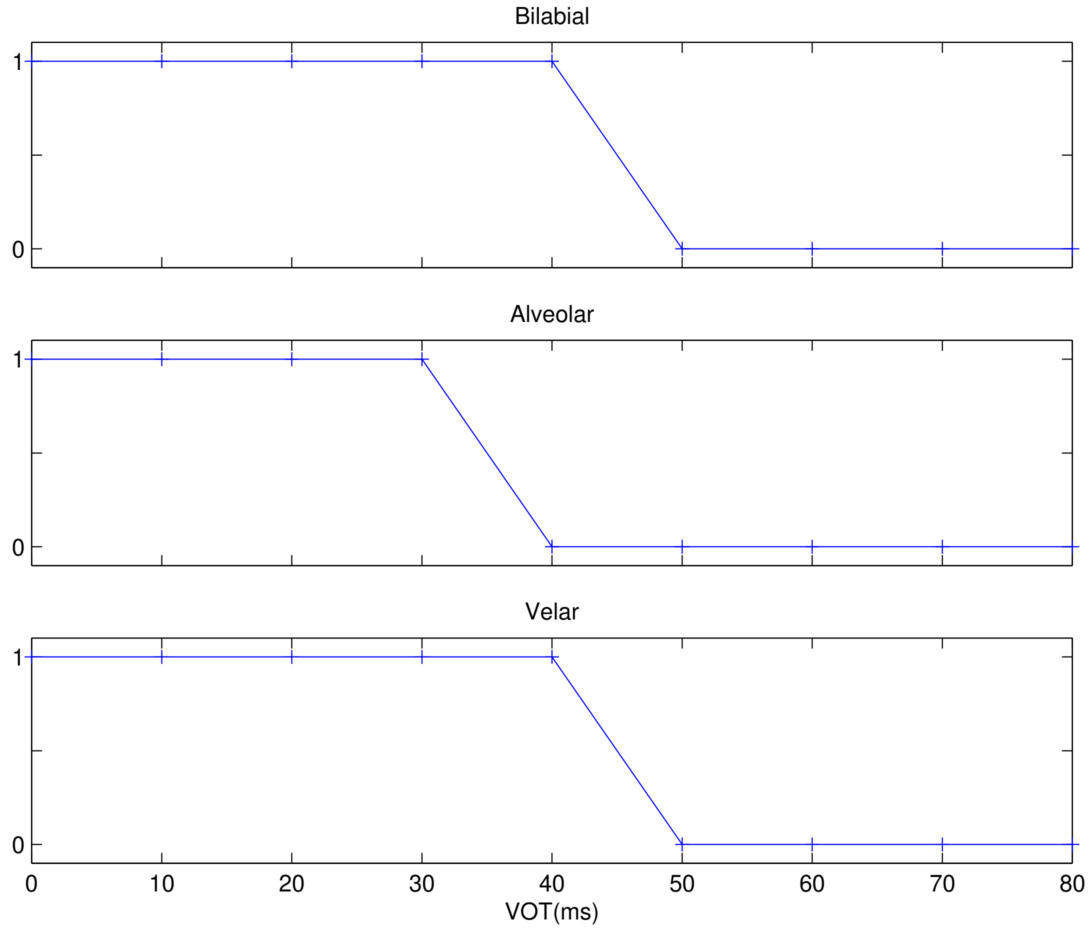

Figure 13: Labeling curves obtained from a hard classifier with decision boundary constructed as the bisector (in 192-dimensional space) between FFT-analysed endpoints. Correct movement of the boundary with place of articulation is not maintained, indicating that some aspect of the auditory transformation is essential to realistic simulation of this effect.

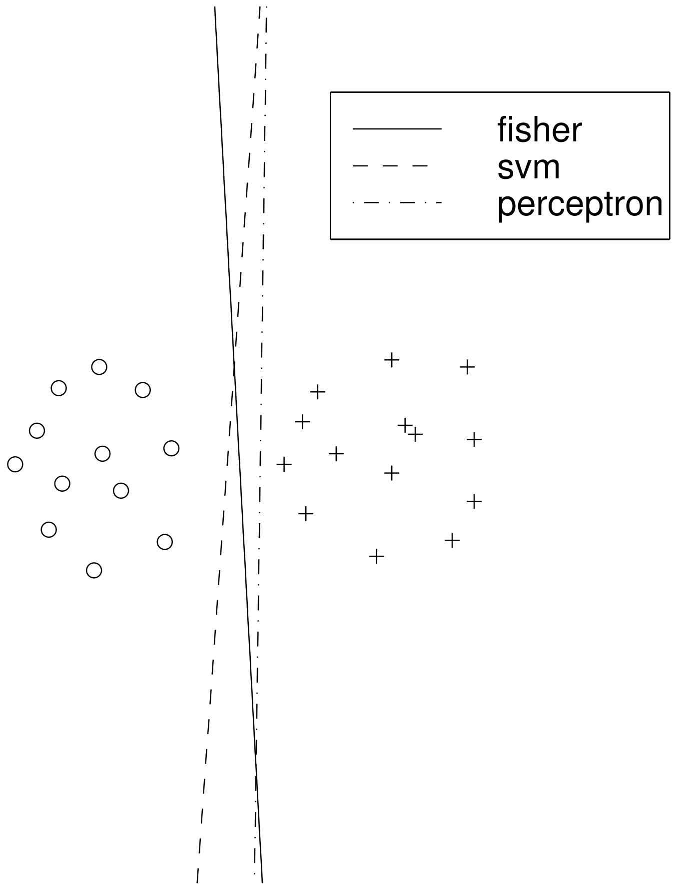

Figure 14: Illustration of differences between Fisher linear discriminant and separating hyperplanes of a support vector machine and a typical perceptron.

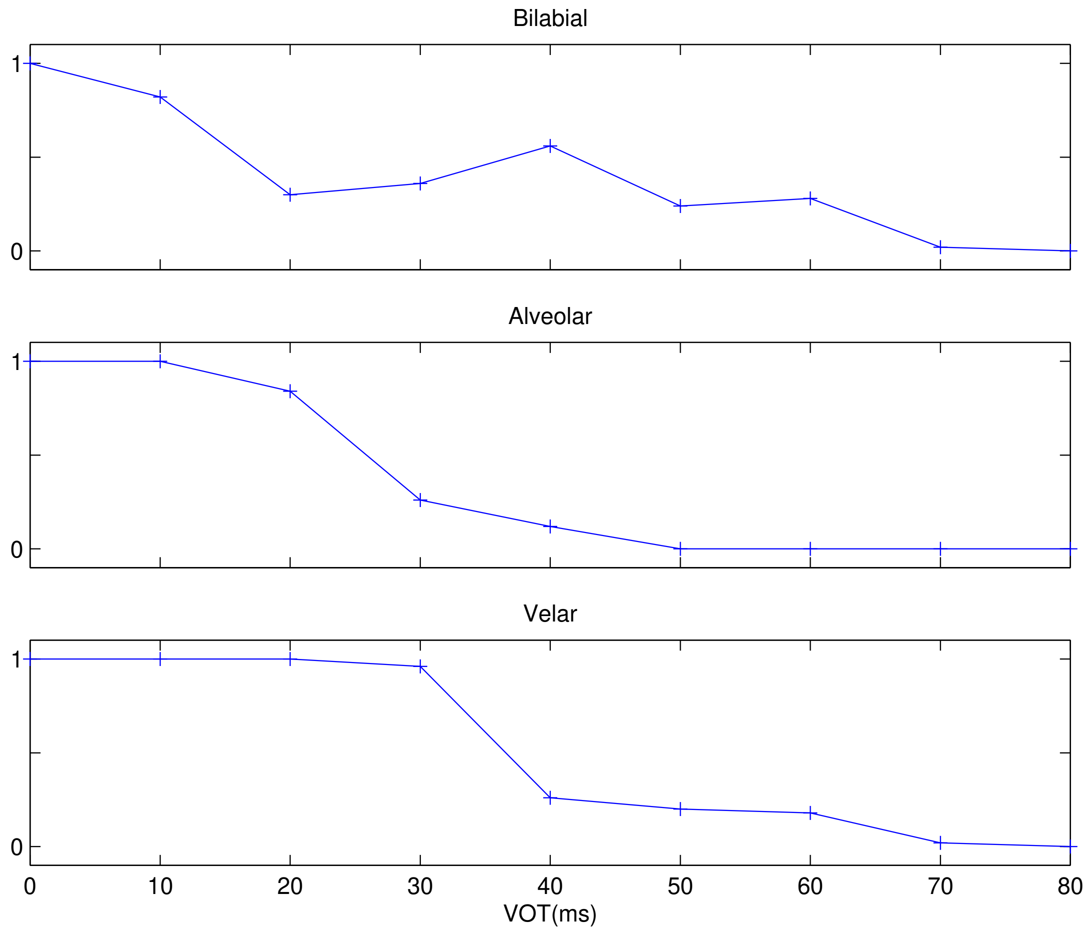

Figure 15: Labeling curves obtained from a Fisher linear discriminant analysis of the three stimuli series.

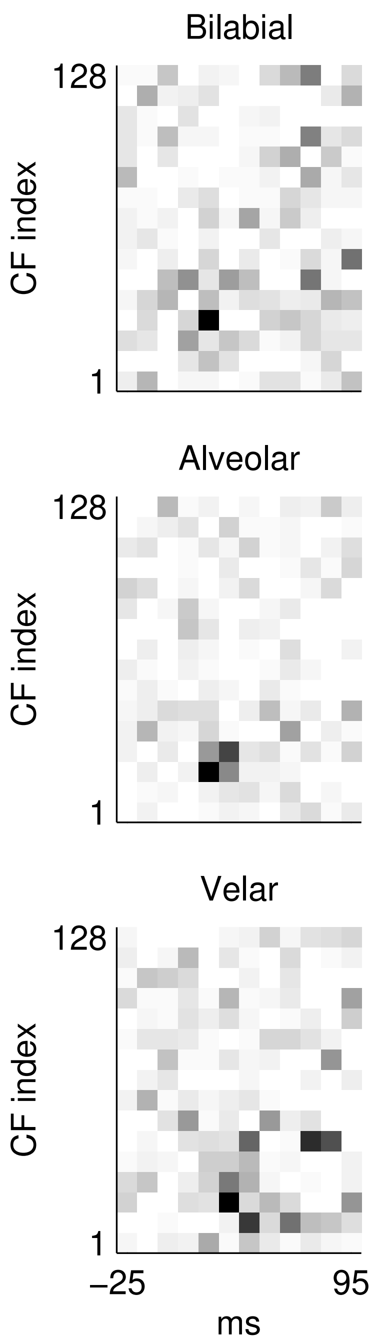

Figure 16: Gray-scale depiction of the squared weight vectors of the Fisher linear discriminant for the three stimuli series.