Auditory Representations of Speech Sounds in a Neural Model: the Role of Peripheral Processing

R.I. Damper,

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

(email: rid@ecs.soton.ac.uk)

Abstract—

The categorization of speech sounds by the auditory system has been a subject of intense attention over several decades, reflecting its importance to the scientific study of speech perception and the technological development of more human-like capabilities in automatic speech recognition. In previous work, we have firmly established that a two-stage computational model can mimic important aspects of the speech categorization behavior of human and animal listeners. The first stage employs a biologically-motivated 'front-end', modeling the peripheral auditory system, and the second stage is a trainable artificial neural network 'back-end', modeling more central processes. When suitably trained on synthetic stimuli, the two-stage system is able to reproduce the important effects of category formation for the class of initial plosive-stop speech sounds, and movement of the category boundary with place of articulation. Appropriate behavior is maintained across a variety of 'back-end' architectures and associated learning algorithms. The behavior is *emergent* in that it was not explicitly programmed into the model. These facts imply that there is something very basic about categorization behavior.

Unlike real (human and animal) listeners, a software model can be interrogated to find out the contribution of its component parts to the overall behavior. Replacing the auditory front-end by a more prosaic fast Fourier transform analyzer allows us to focus on the contribution of the acoustic-to-auditory transformation to categorization. We find that the front-end processor is not essential to category formation but plays an important part in the boundary-movement phenomenon, by emphasizing important time-frequency regions of the speech signal.

I. 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, understanding speech perception is virtually synonymous with understanding the staged transformations which relate the physically-continuous acoustic stimulation 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 the celebrated 'speech invariance problem'. 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 [1], however:

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

while Kuhl and Miller [2] 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."

While this intervention is difficult or impossible to achieve in experiments using human or animal listeners, it is 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 [3], [4], [5], [6].

In previous work, we have focused on simulated representations of synthetic speech sounds at the level of the auditory nerve. 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 is able to categorize the patterns of simulated auditory nerve activation in a way which mimics the psychophysical behavior of real listeners. The question which then arises, and which we address here, is: how important is the restructuring of information by the peripheral auditory system to the obtained categorization?

The remainder of this paper is organized as follows. In Section II immediately following, we outline our modeling philosophy. In Section III, we review the important aspects of the perception of voicing in syllable-initial consonants, not only by human listeners, but by animals and machines (i.e. software models) also. Since the software simulation replicates the important aspects of the human and animal data, we describe in Section IV an analysis aimed at discovering the auditory features underpinning this behavior. In Section V, we report the result of removing the front-end of the simulation, so as to assess the importance of restructuring of information by the peripheral auditory system. Finally, Section VI presents some discussion and

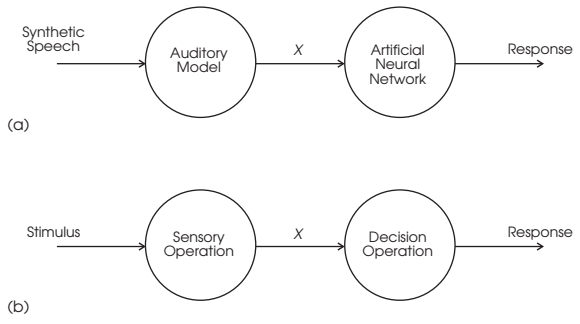

Fig. 1. (a) Two-stage model of the auditory system consisting of a biologically-faithful simulation of the peripheral auditory system and an artificial neural network trained on the auditory nerve firing patterns denoted X. This processing scheme can be usefully related to the framework of signal detection theory (b) which clearly separates sensory and decision operations.

conclusions.

II. Approach to Neural Modeling

Artificial neural network (ANN) modeling in the 'parallel distributed processing' style [7] and the biologically-faithful modeling of real neural systems [8], [9] are typically seen as orthogonal approaches to the simulation of intelligent behavior. The former abstracts away putatively irrelevant complexities of cell electrophysiology and interconnection, allowing systems of practically large size to be studied, but at the expense of abandoning biological fidelity. The latter retains fidelity but computational complexity dictates that only relatively small neural systems (of known structure and function) can be considered [10].

We employ both modeling approaches, treating them as complementary. Exploiting detailed anatomical and physiological knowledge of the peripheral auditory system, the first stage of the simulation is a biologically-faithful model of the transduction of sound from the pinna to the mammalian auditory nerve [11]. The second stage is a trainable ANN which models essentially unknown details of central auditory system function at a high level of abstraction. (See Shamma [12] for earlier, similarly-motivated work in this area.) Figure 1(a) shows a schematic representation of this processing scheme and relates it to a traditional signal-detection theoretic viewpoint [13] in which a sensory process produces a (unidimensional) variate X which is the basis for subsequent decisions (Figure 1(b)). Signal detection theory is important and relevant here [14], [15], because its clear separation of sensory and decision operations focuses attention on the locus of categorization. That is, does it take place at the sensory stage (so that X is discrete) or at the decision stage (so that X is continuous)?

Here, X is the (continuous) firing pattern of auditory nerve activity computed in response to the synthetic speech stimuli developed by Abramson and Lisker [16] – their so-called voice onset time (VOT) continuum. The ANN is trained on these patterns ('neurograms') and, hence, acts as a 'synthetic listener'. The perception of the Abramson and Lisker VOT stimuli by human and animal listeners has been much studied and, as reviewed immediately below, a good deal is known about this. Accordingly, we are able to verify that the two-stage model acts essentially indistinguishably from a real listener. In this work, we remove the auditory front-end – the sensory processing part – and train the ANN on a more direct representation of the stimulus, in order to assess the role of the auditory periphery in categorization.

III. Real and 'Synthetic' VOT Perception

It has been known for many years now that the VOT continuum is perceived 'categorically', i.e. perception changes abruptly from 'voiced' to 'unvoiced' as VOT is increased uniformly, and discrimination is far better between categories than within a category. Hence, labeling functions are non-uniform and discrimination functions are non-monotonic. An intriguing finding is that such categorical behavior is also observed in non-human listeners – a result which 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.

In now-classical work, Kuhl and Miller [2] obtained labeling curves for English speakers and for chinchillas in response to bilabial (/ba-pa/), alveolar (/da-ta/) and velar (/ga-ka/) stimuli in which VOT was varied. These revealed labeling functions in which there was a sharp transition from a high number of 'voiced' judgements to a low number as VOT increased. The functions were well fitted by a probit (sigmoid). Taking the 50% points as the boundaries between voiced and unvoiced categories, there was a phoneme boundary-shift effect with place of articulation such that the boundary moves from about 25 ms through 35 ms to 42 ms as the place of constriction in the vocal tract moved back from bilabial through alveolar to velar. Also, the chinchillas exhibit boundary values not significantly different from the humans (although the curves are less steep).

In previous work, we have employed ANNs as synthetic listeners. A variety of neural models has been studied: brain-state-in-a-box associative networks [15]; competitive-learning networks[5]; multilayer perceptrons (MLPs) [3], [15] and single-layer perceptrons (SLPs) [5]. In all cases reported here, the ANNs were perceptrons trained by back-propagation [17]. In this earlier work, networks were trained on a neurogram representation of the Abramson and Lisker stimuli.

Neurograms were computed as follows. Stimuli were applied to the auditory model at time $t = 0$ at a simulated level of 65 dB SPL. The times of firing ('spikes') of each of the simulated auditory-nerve fibers were noted. There are 128 of these and, because of the tonotopic organization of the auditory system, each can be associated with a particular 'center' frequency (CF). Activity before $t = 0$ is spontaneous, as is that in channels with CF index 1..8 (for reasons to do with the bandwidth of our auditory filters at low frequency). Fuller details are given in [3], [11]. (Thanks to the detailed modeling of phenomena such as middle-ear transmission, basilar membrane dynamics, mechanical-to-neural transduction, neural tuning and tonotopic organization, rate saturation, two-tone suppression etc., it is possible to demonstrate that our computed responses agree very well in all details with the physiological recordings (in cat) of auditory nerve responses to synthetic /ba/ and /da/ syllables (0 ms VOT) by Miller and Sachs [18], as redrawn by Shamma [19].) Spikes were then counted in a (12×16) analysis window stretching from -25 ms to 95 ms in 10 ms steps in

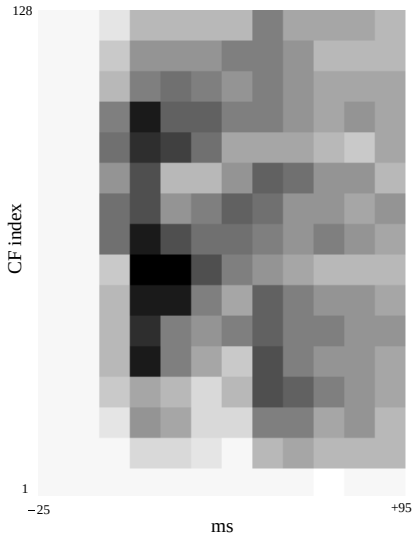

Fig. 2. Typical 'reduced' neurogram in the (12×16) matrix form presented to the neural networks: / ba / stimulus, 40 ms VOT. There is very significant data-reduction relative to a representation which retains the CF and time identity of each spike.

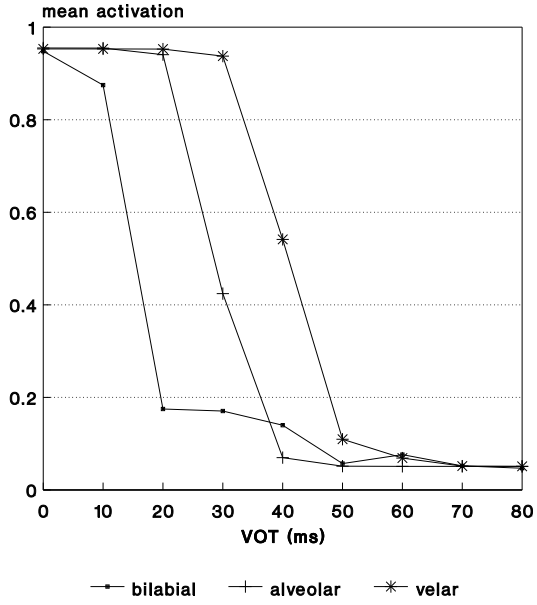

Fig. 3. Mean output activation versus VOT for MLPs with 2 hidden units trained on neurograms from 0 ms and 80 ms endpoints. The boundary placement as a function of place of articulation mimics that seen in 'real' (human and animal) listeners.

the time dimension and from 1 to 128 in steps of 8 in the CF dimension. Figure 2 shows a typical such 'reduced' neurogram in response to the / ba / stimulus with 40 ms VOT. We refer to this as 'reduced' because there is clearly very considerable data reduction relative to a representation which retains the individual CF and time identity of each and every spike. Some such reduction is, of course, necessary if the parameters (connection weights and thresholds) of the ANNs are to be reliably estimated from limited training data. The reader should nonetheless bear in mind that all fine timing information has effectively been eradicated from the inputs to the ANNs.

Separate networks were constructed for each of the three stimulus series (bilabial, alveolar and velar). Each had 192 (12×16) input units, a variable number (h) of hidden units, and a single output unit (with sigmoidal activation function) to act as a voiced/unvoiced detector.

As in the Kuhl and Miller study with chinchillas (which had to be trained to respond appropriately to the stimuli), the MLPs were trained on 50 repetitions of the endpoint stimuli (0 and 80 ms VOT) and tested on 50 repetitions of the full range of values (0 ms to 80 ms in 10 ms steps), so that generalization was tested on the intermediate (10 ms to 70 ms) stimuli. Because the auditory model is probabilistic in nature (as a result of its simulation of mechanical-to-neural transduction in the cochlea), stimulus repetition produced non-identical neurograms. This is convenient, because it allows us to generate sufficient training data for our purposes. Target outputs were 1 for the voiced (0 ms VOT) stimuli and 0 for the unvoiced (80 ms VOT) stimuli. Training used a learning rate of 0.01, a momentum of 0.7, and a range of ± 0.02 for the initial, random weights; and was terminated when the average squared error per training pattern was 0.0025.

Figure 3 shows typical labeling functions obtained by averaging output activations over the 50 stimulus presentations. In this case, there were $h = 2$ hidden units, but results were insensitive to the value of h. This is illustrated by Figure 4 in which essentially the same curves are obtained with single-layer perceptrons (SLPs) having no hidden units whatsoever ($h = 0$). The form of these labeling functions was insensitive to the initial random weight settings for the back-propagation training. That is, labeling functions like these – with the correct order of boundary shift – were consistently obtained over several repetitions of the training.

Figures 3 and 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. Thus, the neural model is clearly capturing the 'essence' of categorical perception. The behavior 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.

IV. Analyzing the Model

Unlike real (human and animal) listeners, the computational model can be systematically manipulated and probed to deter-

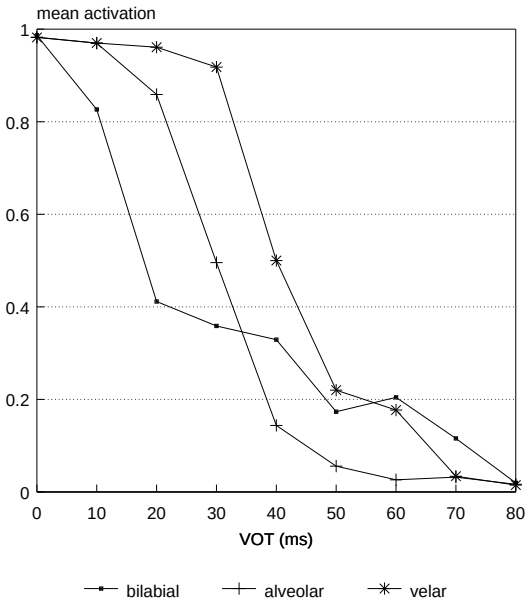

Fig. 4. Mean output activation versus VOT for single-layer perceptrons trained on neurograms from 0 ms and 80 ms endpoints.

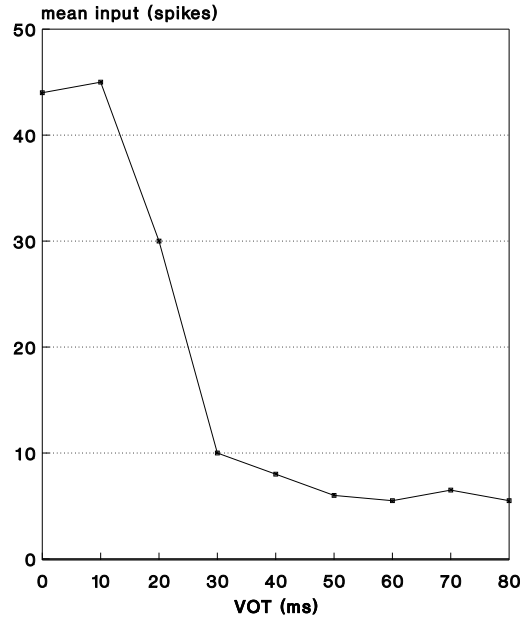

Fig. 5. Mean spike-count input for the 5 SLP nodes with maximal product of input and *positive* weight versus VOT for the alveolar series.

mine the basis of its behavior. In this regard, the finding that no hidden units are necessary in order to simulate CP is very important. A major attraction of the single-layer perceptron architecture is that it is straightforward to identify the areas of the neurogram which contribute to the observed categorization behavior. That is, all connections are direct from the neurogram to the SLP output node, without the complication of intervening hidden units.

A simplified contribution analysis [20] was conducted by identifying the connections 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, was found to produce more meaningful results. The analysis considered positive and negative weight values separately. (Note that the input values are spike counts and are always positive.) Highest products of input and *positive* weight are located around the low-frequency region (the four frequency channels covering CF indices 8 to 48 in the model, corresponding to 73 to 675 Hz) just after acoustic stimulus onset where voicing activity varies maximally as VOT varies. (This is perhaps not surprising as this the region of the perceptually important first formant ($F1$) transition.) The precise location of this region shifts in the three nets (bilabial, alveolar, velar) in the same way as the boundary point. Averaging the inputs to the 5 SLP nodes with the largest positive product across all 50 stimuli at each value of VOT produces the pattern in Figure 5 (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 a sigmoidal activation function as there is in the neural model, so indicating that categorization behavior is not merely a trivial consequence of specific details of the ANN architecture. Similar findings obtain for the bilabial and velar series.

A similar plot focusing on the large absolute 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 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, count spikes in this region. But how important is restructuring of information by the peripheral auditory system to this mechanism?

V. Training on an Acoustic Representation

The previous section has illustrated the merits of an approach whereby an artificial neural network is trained on a set of neurograms obtained from repeated application of end-point stimuli – and is then analyzed to find what it has learned in terms of its connection weights. We can therefore assess the importance of the auditory front-end by removing it from the simulation. We note at this point, however, the potentially important fact that there is only one (synthesized) example of each endpoint stimu-

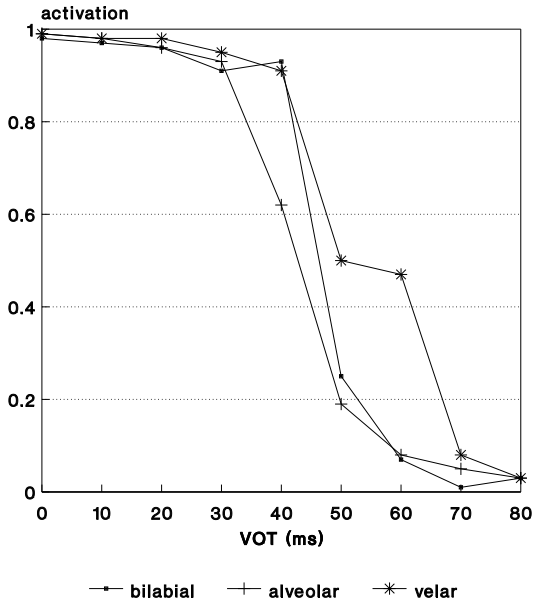

Fig. 6. Output activation versus VOT for single-layer perceptrons trained on FFT-processed 0 ms and 80 ms endpoint stimuli. The boundary shift with place of articulation is abolished.

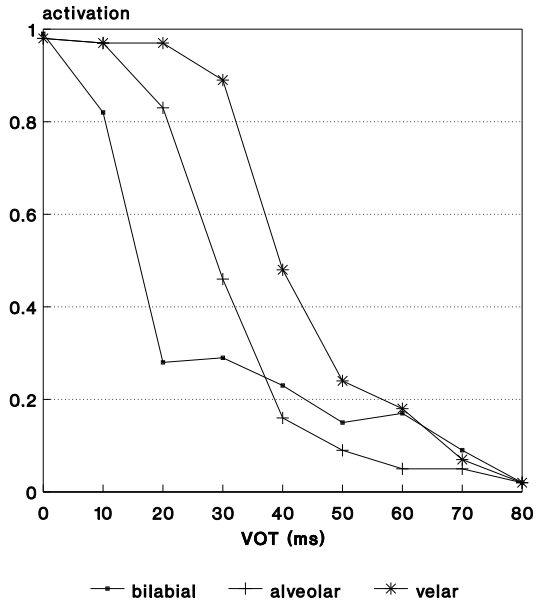

Fig. 7. Output activation versus VOT for single-layer perceptrons trained on averaged neurograms for 0 ms and 80 ms endpoint stimuli. In spite of the probable under-training, the boundary shift with place of articulation is restored.

lus. Thus, we have very little data on which to train the ANNs. We return to this matter below.

Since the waveform representations of the stimuli are, because of their high redundancy, inherently unsuitable for input to the neural network, they were pre-processed by fast-Fourier transform (FFT). That is, the power spectral densities of the stimuli were computed using a 256-point fast Fourier transform over 25.6 ms frames (the sampling rate was 10 kHz) 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 again summed in a (12×16) analysis window stretching 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, to form the input to the nets. So, in this case, the frequency dimension is divided up linearly (in Hz) rather than approximately logarithmically according to CF.

Figure 6 shows a typical result for an SLP trained on the FFT-processed endpoints. As can be seen, the sharp labeling function is retained, but the boundary shift is abolished. It seems, therefore, that the auditory front-end – mimicking sensory processing – is essential to the proper simulation of CP. There are, however, some important differences between the ways that the labeling functions of Figure 6 were produced relative to those depicted earlier: namely the paucity of training data in the present case, and the different frequency scalings.

As there is only one training instance for each endpoint and each place of articulation, it is likely that the networks trained on FFT-processed data are in fact *under*-trained. Further, it may be important that the network sees some appropriate statistical distribution of the data during training. To test the importance of such factors to correct boundary shift, new sets of training data were constructed, consisting of single, averaged instances of the endpoint neurograms. Figure 7 shows a typical result: as can be clearly seen, in spite of the probable under-training, the correct boundary shift with place of articulation is restored. The indication is that there is indeed an important role for "restructuring of information" by the peripheral auditory system.

This, however, is contrary to the conclusion of [6]. In this latter work, the spectral energy in the acoustic signal (FFT-processed) was evaluated for the same putatively-important time-frequency region as used in the spike-counting model. Figure 8 shows the result, which is a very fair approximation to the 'correct' labeling functions, with appropriate boundary shift, except that the category boundaries are too long by about 10 ms. This is a consequence of using the simulated neural responses to place the analysis region without regard for the propagation delay through the peripheral auditory system. Although this delay is actually a function of frequency, it does not deviate too much from about 10 ms for the frequencies of interest here. Taking proper account of the propagation delay results in approximately correct boundaries.

VI. Discussion and Conclusions

How can the apparent contradiction between this last finding and the inability of the SLPs trained on FFT-processed data to exhibit correct boundary shifts be resolved? It seems that the speech perception mechanism must focus on particular time-

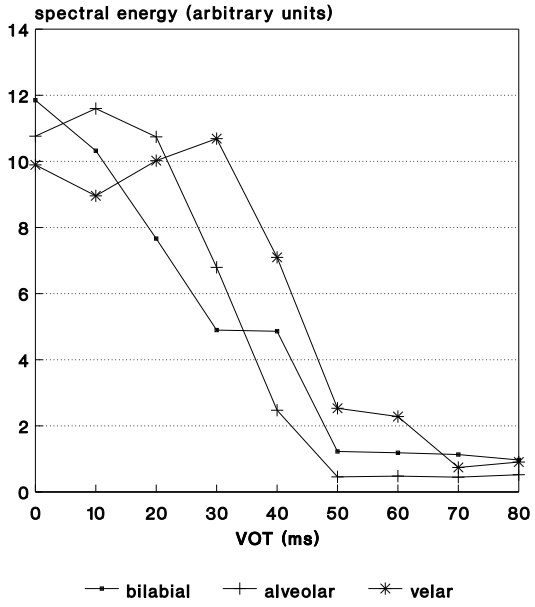

Fig. 8. Spectral energy in the acoustic stimuli evaluated over a time-frequency range corresponding to that for the spike-counting model (Figure 5).

frequency regions in order to make discriminations between the categories of speech sounds studied here. Processing by the periphery renders these regions more prominent in the auditory nerve representation. Thus, neural networks trained on an auditory representation are able to make the necessary discriminations. On the other hand, networks trained on the acoustic data have difficulty because the important information, although present, is insufficiently prominent.

In conclusion, important aspects of the voiced/unvoiced categorization of synthetic syllable-initial stop consonants are reproduced by a two-stage simulation of the auditory system. This behavior is *emergent* – it is not explicitly programmed into the model – and no fine timing information is necessary. Unlike real (human and animal) listeners, the computational model can be systematically manipulated to determine the basis of its behavior. This reveals information in the region of first formant ($F1$) onset is vital to the perception of voicing for these stimuli. The peripheral auditory system plays an important part in emphasizing this region in the neural representation.

VII. Acknowledgements

The VOT stimuli used here 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.

Thanks also to my student Mat Gore for programming and data-analysis effort.

References

[1] A.Q. Summerfield, "Differences between spectral dependencies in auditory and phonetic temporal processing: Relevance to the perception of voicing in initial stops," *J. Acoust. Soc. Am.*, Vol. 72, 1982, pp. 51–61.

[2] P.K. Kuhl and J.D. Miller, "Speech perception by the chinchilla: Identification functions for synthetic VOT stimuli," *J. Acoust. Soc. Am.*, Vol. 63, 1978, pp. 905–917.

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

[4] R.I. Damper, "Connectionist models of categorical perception of speech," *Proc. IEEE Int. Symp. Speech, Image Processing & Neural Networks, Vol. 1*, Hong Kong, 1994, pp. 101–104.

[5] R.I. Damper, S. Harnad and M.O. Gore, "A computational model of the perception of voicing in initial stop consonants," *J. Acoust. Soc. Am.*, submitted.

[6] R.I. Damper, "A biocybernetic simulation of speech perception by humans and animals," *Proc. IEEE Int. Conf. Systems, Man & Cybernetics (SMC '97), Vol. 2*, Orlando, FL, 1997, pp. 620–625.

[7] D.E. Rumelhart and J.L. McClelland (editors), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition (2 Vols.)*, Cambridge, MA: Bradford Books/MIT Press, 1986.

[8] R.J. MacGregor, *Neural and Brain Modeling*, London, UK: Academic, 1987.

[9] C. Koch and I. Segev (editors), *Methods in Neuronal Modeling: From Synapses to Networks*, Cambridge, MA: MIT Press, 1989.

[10] T.W. Scutt and R.I. Damper, "Computational modelling of learning and behaviour in small neuronal systems," *Proc. Int. Joint Conf. Neural Networks*, Singapore, 1991, pp. 430–435.

[11] M.J. Pont and R.I. Damper, "A computational model of afferent neural activity from the cochlea to the dorsal acoustic stria," *J. Acoust. Soc. Am.*, Vol. 89, 1991, pp. 1213–1228.

[12] S.A. Shamma, "Speech processing in the auditory system. II: Lateral inhibition and the central processing of speech invoked activity in the auditory nerve," *J. Acoust. Soc. Am.*, Vol. 78, 1985, pp. 1622–1632.

[13] D.M. Green and J.A. Swets, *Signal Detection Theory and Psychophysics*, New York: Wiley, 1966.

[14] N.A. Macmillan, H.L. Kaplan and C.D. Creelman, "The psychophysics of categorical perception," *Psych. Rev.*, Vol. 84, 1977, pp. 452–471.

[15] R.I. Damper and S. Harnad, "The psychophysics of synthetic categorical perception," *Percept. Psychophys.*, submitted.

[16] A. Abramson and L. Lisker, "Discrimination along the voicing continuum: Cross-language tests," *Proc. 6th Int. Cong. Phonetic Sciences, Prague, 1967*, Academia, Prague: 1970, pp. 569–573.

[17] D.E. Rumelhart, G.E. Hinton and R.J. Williams, "Learning representations by back-propagating errors," *Nature*, Vol. 323, 1986, pp. 533–536.

[18] M.I. Miller and M.B. Sachs, "Representation of stop consonants in the discharge pattern of auditory-nerve fibers," *J. Acoust. Soc. Am.*, Vol. 74, 1983, pp. 502–507.

[19] S.A. Shamma, "Speech processing in the auditory system. I: The representation of speech sounds in the responses of the auditory nerve," *J. Acoust. Soc. Am.*, Vol. 78, 1985, pp. 1612–1621.

[20] D. Sanger, "Contribution analysis: A technique for assigning responsibilities to hidden units in connectionist networks," *Connection Sci.*, Vol. 1, 1989, pp. 115–138.