The Role of the Auditory Periphery in the Categorization of Stop Consonants

R.I. Damper

Department of Electronics and Computer Science, University of Southampton, Southampton S017 1BJ, United Kingdom

Abstract: A proper understanding of speech perception requires an understanding of the staged restructuring of information that occurs as an acoustic stimulus is transformed by the auditory system into a phonetic percept. It is now firmly established that a physiologically-faithful computer simulation of peripheral processing feeding into a trainable artificial neural network (ANN) is capable of reproducing the important aspects of the categorization of initial stop consonants into voiced and unvoiced classes. Correct behavior is found to be insensitive to the ANN architecture and precise training scheme used, suggesting that such categorization is very basic to the repertoire of auditory processing strategies. Unlike a human or animal listener, a computational model can be easily manipulated and interrogated to discover the underlying basis of category formation. In particular, the physiological-based simulation of the auditory periphery can be dramatically simplified and the effect on categorization observed. It is found that proper modeling of frequency scaling and neural adaptation are essential to correct simulation of the well-known shift of phoneme boundary with place of articulation for stop consonants. The role of the auditory periphery in this case seems to be to emphasize the region of first formant information around the time of voicing onset.

INTRODUCTION

The categorization of speech sounds by the auditory system has been a subject of intense attention over several decades e.g. [1, 2], 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 [3, 4, 5, 6, 7], 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. It is certainly not necessary to posit 'speech is special' or direct perception explanations when a very simple learning system is adequate to reproduce results from real listeners. In the words of [6], these results "support the notion that phonetic percepts are directly recoverable from acoustic and auditory representations."

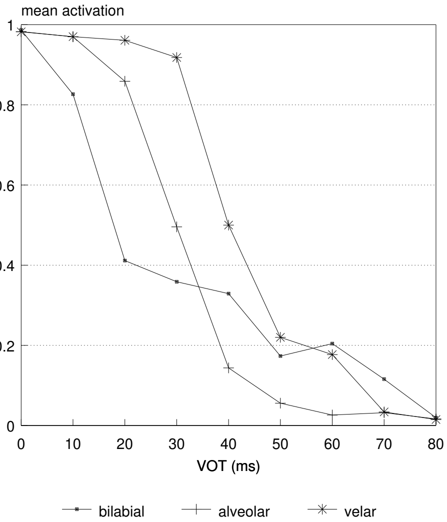
INTERROGATING THE COMPUTATIONAL MODEL

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. In particular, we focus here on the contribution of the auditory front end by replacing it by a more prosaic fast Fourier transform (FFT) analyzer. This allows us to determine the part played by the acoustic-to-auditory transformation in categorization.

Figure 1(a) shows the labeling functions obtained by plotting the mean output activation of single-layer perceptrons (SLPs) – trained by back-propagation on simulated auditory nerve representations of /ba-pa/, /da-ta/ and /ga-ka/ 0 ms and 80 ms voice onset time (VOT) endpoints – when tested over the full range of VOT. The result is classical categorization with the usual shift of the phoneme boundary according to place of articulation. When the SLP is trained on FFT-transformed acoustic data, however, the boundary shift is abolished (Figure 1(b)). Thus, the front-end processor is not essential to category formation but apparently plays an important part in the boundary-movement phenomenon.

Using single layer perceptrons, it is straightforward to identify those inputs which contribute to the observed categorization behavior as all connections are direct from the inputs to the (single) output node, with no intervening hidden nodes. In this work, inputs are integrated time-frequency regions of (simulated) auditory activity obtained by summing spike counts within 192 pre-defined time-frequency cells.

Consider the SLP trained on the outputs of the auditory model. The contribution of each of the 192 inputs was assessed by taking the average product of input spike count and corresponding connection weight. This revealed that the significant auditory nerve activity was that occurring just after stimulus onset in the frequency range corresponding

(a)


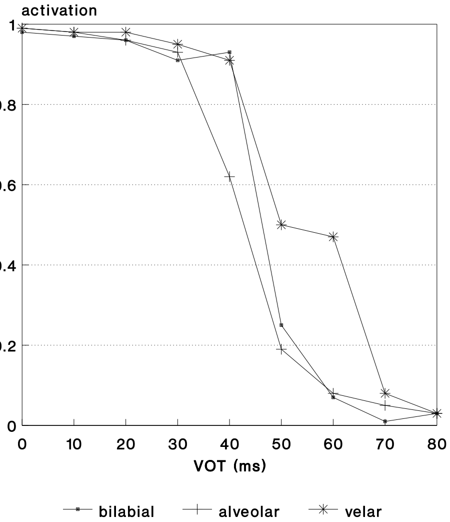
(b)


Figure 1: (a) Labeling functions in terms of mean output activation of SLPs trained on simulated auditory nerve representations of /ba-pa/, /da-ta/ and /ga-ka/ 0 ms and 80 ms VOT endpoints (b) The phoneme boundary shift with place of articulation is removed when the perceptrons are trained on FFT-transformed acoustic data.

to first-formant transition. In fact, categorization can be simply explained in terms of a central mechanism which focuses on particular time-frequency regions of auditory nerve activity and, in essence, counts spikes.

So why do we not get correct categorization with the model trained on FFT-transformed stimuli? By further analysis, we find that the regions of auditory nerve activity which are so important to the learned categorization of Figure 1(a) are not well represented in the FFT data. The auditory periphery performs frequency warping and also emphasizes onsets by the mechanism of neural adaptation. Both of these are important to robust extraction of the key features which underlie the categorization of initial stops.

ACKNOWLEDGMENT

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

REFERENCES

[1] S. Harnad (ed.), *Categorical Perception: The Groundwork of Cognition*, Cambridge, UK: Cambridge University Press, 1987.

[2] A.M. Liberman, *Speech: A Special Code*, Cambridge, MA: MIT Press, 1996.

[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, 7–41 (1990).

[4] R.I. Damper, "Connectionist models of categorical perception of speech," *Proceedings of IEEE International Symposium on Speech, Image Processing and Neural Networks, Vol. 1*, Hong Kong, 101–104, 1994.

[5] R.I. Damper, "A biocybernetic simulation of speech perception by humans and animals," *Proceedings of IEEE International Conference on Systems, Man and Cybernetics, Vol. 2*, Orlando, FL, 620–625, 1997.

[6] R.I. Damper, S. Harnad and M.O. Gore, "A computational model of the perception of voicing in initial stop consonants," *Journal of the Acoustical Society of America*, submitted.

[7] R.I. Damper and S. Harnad, "The psychophysics of synthetic categorical perception," *Perception and Psychophysics*, submitted.