Self-Learning and Self-Organization as Tools for Speech Research

R.I. Damper,

Cognitive Sciences Centre

and

Department of Electronics and Computer Science,

University of Southampton,

Southampton SO17 1BJ,

England.

Email: `rid@ecs.soton.ac.uk`

WWW: `http://isis.ecs.soton.ac.uk/`

May 28, 1999

(Commentary on "Linear Correlates in the Speech Signal: The Orderly Output Constraint", *BBS* target article by Harvey Sussman, David Fruchter, Jon Hilbert and Joseph Sirosh.

Abstract

Locus equations offer promise for an understanding of at least some aspects of perceptual invariance in speech, but they were discovered almost fortuitously. With the present availability of powerful machine learning algorithms, 'ignorance-based' automatic discovery procedures are starting to supplant 'knowledge-based' scientific inquiry. Prin-

ciples of self-learning and self-organisation are powerful tools for speech research but remain somewhat under-utilised.

Locus equations were first discovered by Lindblom (1963) but have since been more thoroughly investigated by Sussman and his colleagues. They offer as much promise for understanding the vexed question of invariance, whereby speech sounds are physically modified by their context but are still perceived as members of the same equivalence class (phoneme category), as any proposal yet advanced. The basic notion is that, while features of the speech signal may vary as a result of coarticulation, the *relation* between certain key features may exhibit a consistent and lawful ('invariant') form.

In the target paper, as well as reviewing the utility of locus equations, Sussman *et al.* argue for their neurobiological plausibility based on the potential to build relational, higher-order feature detectors (and thereby category detectors) from the combination-sensitive neurons found in a variety of mammalian and avian auditory systems. The essential argument is (p. 5, line 4) that "... there is no reason to suspect novel processing strategies to have arisen ... for the basic auditory encoding of the acoustic cues signaling feature contrasts in human speech". This is at variance with the early 'speech is special' hypothesis which still has its adherents (e.g. Liberman and Mattingly, 1989; Liberman, 1996). In my view, the consensus of informed opinion is now firmly on the side of Sussman *et al.*: seeing human speech perceptual mechanisms as based on general auditory processing principles, common to a range of species, with specialisation only occurring at a relatively high level. (See also (p. 7) the personal communication from Suga.) Indeed, using a computational modelling approach, we have recently shown (Damper *et al.* submitted) that the placement of phoneme category boundary in human and animal listeners between initial stops (/b/ ,/p/ ; /d/ ,/t/ ; /g/ ,/k/) distinguished by their voice onset time can be replicated by a trivially simple neural processing scheme which only needs to integrate activity over certain time-frequency regions of auditory nerve activity. Yet this aspect of speech perception has attracted enormous attention over decades.

The traditional approach to the speech invariance problem can be characterised as 'manual search'. That is, using knowledge and ingenuity, the experimenter tries to generate some hypothesis about possible invariant features, which is then tested for consistency with available data. Lindblom's discovery of locus equations is very much in this vein. But the approach is inherently unsatisfactory, both because knowledge and ingenuity are always in short supply and because the consistency check with the data is *a posteriori*. With the advent of connectionism 10 or 15 years ago, and the greatly-increased availability of powerful learning algorithms such as error back-propagation (Rumelhart *et al.*, 1986), it is becoming ever commoner to employ data-driven rather than knowledge-driven strategies in virtually all areas of scientific inquiry.

Such 'automatic search', exploiting the self-learning and self-organising capabilities of neural networks, ensures that (provided training is successful, and the network convergences onto the desired behaviour) then only hypotheses consistent with the training data are (implicitly) generated. Further, the search is guided by a general optimisation principle (i.e. it is 'ignorance-based'). Together, these two considerations mean that 'features' in the widest sense which are important to categorisation (but which may not be obvious, e.g. because they are relational) naturally emerge as determiners of network behaviour. Of course, it remains to uncover – by an appropriate analysis of the trained network(s) – the implicit hypotheses which have been automatically discovered. Contrary to the wide-spread belief that neural networks are 'black boxes' whose operating principles cannot be sensibly determined, techniques for doing just this are improving all the time.

Not only was this the approach adopted by Damper *et al.* (see above), Sussman and colleagues also mention (p. 20, line 4 *et seq*) several neural network studies (e.g. Waibel *et al.*, 1989) which – after analysis "to determine which parts of the input were most effective for the task" – confirm the importance of $F2_{\text{onset}}$ and $F2_{\text{vowel}}$ to the identification of stop-consonant place of articulation. Sussman *et al.* themselves employ self-organising Kohonen maps (Kohonen, 1990) to confirm the clear emergence of topologically-ordered regions reflecting the consonant categories on the $F2_{\text{onset}}/F2_{\text{vowel}}$ map but not, for instance, on the $F2_{\text{vowel}}/F3_{\text{vowel}}$ map. Of course, the kind of competitive-learning principle embodied in Kohonen-style self-organisation could as well have been employed to discover this relation *a priori* rather than merely to confirm it *a posteriori*.

In conclusion, Sussman *et al.* present a clear and convincing case for the emergence of higher-order features (exploiting general neural processing, rather than speech-specific, principles) as the basis of at least some of the category invariance observed in speech perception. There is great scope, however, for such features and principles to be discovered automatically in the future. Self-learning and self-organising systems offer a valuable and currently under-used tool for speech research.

References

Damper, R.I, Harnad, S. and Gore, M.O. (submitted). A computational model of the perception of voicing in initial stops. *Journal of the Acoustical Society of America*.

Liberman, A.M. (1996). *Speech: A special code*. Cambridge, MA: MIT Press.

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, **78**, 1464–1480.

Liberman, A.M. and Mattingly, I. (1989). A specialization for speech perception. *Science*, **243**, 489–494.

Lindblom, B. (1963). On vowel reduction. *Report No. 29, Speech Transmission Laboratory*, Royal Institute of Technology, Stockholm, Sweden.

Rumelhart, D.E., Hinton, G.E., and Williams, R.J. (1986). Learning representations by back-propagating errors. Nature, **323**, 533–536.

Waibel, A., Hanazawa, T., Hinton, G., Shikano, K. and Lang, K.J. (1989). Phoneme recognition using time-delay neural networks. *IEEE Transactions on Acoustics, Speech and Signal Processing*, **ASSP-37**, 328–339.