Modeling the Acoustic-to-Auditory Transformation for Stop Consonant-Vowel Syllables

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


Corresponding author: R. I. Damper
Presenting author: R. I. Damper
Emails: {rid|srg}@ecs.soton.ac.uk
Telephone: +44 23 80 594577
FAX: +44 23 80 594498
First choice: speech and language (B)
Second choice: audition (B)
A key question for the study of spoken language is: How is the continuous sound pressure wave of speech decoded by a listener into the discrete percept of a particulate linguistic message? It is well known that this decoding exhibits a high degree of invariance to the acoustic effects of speaker, phonetic context etc. Significant efforts have been, and still are, aimed at determining the mechanisms of sound-to-symbol decoding, with particular emphasis on stop consonant perception where the degree of invariance is especially marked. Unsurprisingly, the traditional research method employed in this field has been human perception studies. Less usually, signal analysis (e.g., Kewley-Port 1982; Soli 1983) and animal psychophysics (e.g., Kuhl and Miller 1978; Kluender, Diehl, and Killeen 1987; Sinnott and Williamson 1999) have been used. However, computational modeling offers a valuable, complementary, yet under-exploited approach to the study of speech perception.

In previous work over many years (Damper, Pont, and Elenius 1990; Damper 1998; Damper, Gunn, and Gore 2000; Damper and Harnad 2000), we have studied extensively the emergence of phonetic categories in a variety of computational models in which a biologically-inspired ‘front end’ produces an auditory time-frequency representation which acts as input to a connectionist ‘back end’ learning system. Past work has concentrated on modeling the perception of the voicing distinction in the particular synthetic stop consonant-vowel (CV) stimuli of Lisker and Abramson (1970). This has shown that realistic modeling of the movement of phoneme boundary with place of articulation (which is a characteristic of these stimuli) is dependent on realistic modeling of auditory preprocessing, confirmed the relative importance of formant transition information over onset information, suggested that the ‘low-frequency property’ (Stevens and Blumstein 1981) is primary and (together with the animal psychophysics work) comprehensively denied the necessity to invoke ‘special mechanisms’ to explain human speech perception.

Major limitations of the work so far are the restriction to the voicing distinction and the sole use of the Lisker and Abramson stimuli. In this paper, we will report on modeling the place distinction in CV stimuli. Not all synthetic CV series show the same boundary shift with place of articulation (e.g., Kluender 1991) and synthetic stimuli do not capture the context variability of real speech. Hence, we will also report on modeling work with the real speech data of Nossair and Zahorian (1991).

References


