ON THE LEARNABILITY OF THE VOICING CONTRAST FOR INITIAL STOPS

R.I. Damper　　*S.R. Gunn*

Department of Electronics and Computer Science,
University of Southampton, Southampton SO17 1BJ, UK

ABSTRACT

The categorical perception (CP) of syllable-initial stop consonants has been intensively studied using psychophysical procedures over many decades. However, computational models consisting of an auditory 'front end' and a learning system as a 'back end' convincingly mimic the essentials of CP. Unlike real listeners, such models can be systematically manipulated to uncover the basis of their categorisations. In this paper, we explore the use of modern inductive learning techniques in simulating CP.

1. INTRODUCTION

A vital part of understanding speech perception is understanding the transformations which relate physically-continuous acoustic stimulation to the discrete code of phonetic percepts. It is immeasurably easier to observe the restructuring of information in a software model of auditory processing than in experiments using human or animal listeners. We have worked for several years on such models, trying to understand the mechanisms of the categorical perception (CP) of voicing in syllable initial stop consonants [1, 2, 3, 4]. Our models have a physiologically-realistic 'front end' producing simulated firing patterns in response to synthetic speech sounds at the level of the auditory nerve. A trainable neural network 'back end' then learns to categorise these patterns. We find that subtle aspects of the psychophysical behaviour of real listeners are mimicked by these models.

However, the neural network paradigm has fundamental shortcomings in terms of learning theory, which mean that our results could be artefactual. Most important, there is only one instance of each specific synthesised speech token, so that networks are inevitably undertrained. Here we use the modern inductive inference technique for small sample sizes of support vector machines [5] to simulate CP.

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 [6] investigated the perception of voicing in syllable-initial stop consonants by English listeners as voice-onset time (VOT) 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. As a consequence, labelling (identification) functions are non-linear, 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. Taking the 50% points on the labelling functions as the voiced/unvoiced boundaries, then as the place of

VOT IN MSEC