Roger K. Moore  
University of Sheffield

Some insights into talker-listener-environment coupling, energetics and the contrastive particulate structure of spoken language

Back in the early 1980s, my small research group at the Royal Signals and Radar Establishment (consisting of myself, Martin Russell and Mike Tomlinson) were investigating advanced forms of dynamic time warping (DTW). In particular, we were studying the detailed spectro-temporal relationships revealed by DTW between different versions of the same and contrasting utterances, and we came up with two key techniques for modelling the patterning that we observed: ‘timescale variability analysis’ (TVA) and ‘discriminative networks’ (DNs). TVA was a forerunner of what became widely known as ‘duration modelling’, and DNs were effectively a very early form of sub-word modelling. All of the research was conducted using speech that had been parameterised using the front-end of a military-specification channel vocoder (effectively a 27 channel filter bank). The vocoder not only provided the advantage of real-time speech analysis (so we were able to build real-time ASR systems), but it also offered the bonus that any manipulated speech patterns could be replayed through the channel vocoder synthesiser – and that is what we did on a regular basis, not to generate speech per se, but simply to understand the pattern structures that were embedded in our early statistical models. We were therefore quite surprised when we discovered that we could vary the generated output along various continua automatically such that one word could be transformed to sound like another or, much more interestingly, that a word could be transformed to sound less like another, and that the latter manipulations sounded clearer (as if the speaker was making more effort)! We had, of course, stumbled across a practical demonstration of what Bjorn Lindblom would subsequently publish as his theory of hyper and hypo speech (H&H).

Since those early days, I have held a continuing belief that talkers actively manage their speech production to suit the communicative context (including the listener(s) and the environment), and that such teleological behaviour was the source of much unexplained variability. It is for this reason that Lindblom’s H&H theory figures strongly in my ‘predictive sensorimotor control and emulation’ (PRESENCE) model of speech in which I introduce the concept of ‘reactive speech synthesis’ – a synthesiser that dynamically adjusts its output as a function of the perceived effect on the listener.

In this talk I will discuss my current thinking in this area, touching on Robin Hofe’s investigation into H&H using ‘AnTon’ (his animatronic tongue and vocal tract) and Mauro Nicolao’s research into ‘speech synthesis by analysis’, but also speculating about (i) the wider implications of dynamic coupling between talkers, listeners and their communicative environments, (ii) the fundamental role that energetics plays in conditioning the behaviour of living systems, and (iii) the special consequences for the evolution of a high information-rate low degree-of-freedom system such as spoken language.


