A computational theory of everyday[1] speech perception?

Martin Cooke
University of Sheffield

In spite of efforts dating back to the early part of the last century, we are a long way from possessing a successful model of human speech perception (SP). In fact, it is a surprising fact that we don't have many well-developed *unsuccessful* models of SP! One possible cause is that, while much early work in the field (eg at Bell Labs) was aimed at building explanatory models of processes in speech perception, it appears that computational modellers and psychologists drifted apart between 1950 and 1970. This process, facilitated by wider access to digital computers, spawned automatic speech recognition (ASR), and it seems that the link between modelling and SP was weakened and virtually lost. The new field of ASR did not see its goal as explaining listeners' preferences, but instead focussed on increasing hit rates by all means necessary. And rather than embracing new computational possibilities, much research in SP seems content to apply abstract models at best to discuss listeners' performance.

One consequence of this separation of interests is that few studies in speech perception compare human and model performance using the same speech material. Of these, most use stimuli which are unrepresentative of everyday speech: too short, too synthetic or too clean. Of those investigations which do use everyday speech, very few go further than looking at summary statistics (eg recognition rate and confusion patterns) rather than listeners' responses when presented with individual tokens.

It is perhaps, then, no coincidence that the performance of ASR systems in everyday speech conditions remains poor. It is my contention that both ASR and SP would benefit from a tighter integration. ASR is desperately in need of inspiration from SP research, while SP can no longer afford to ignore the closest thing we have to a computational theory, viz ASR.

Concretely, this proposal[2] is to develop an experimental programme to build models of SP which employ existing ASR techniques, together with recent refinements such as data selection, counter-evidence, compensation in the presence of masking, etc. Central to this proposal is the need for a corpus of everyday speech material which is suitable for the differing requirements of ASR (large and trainable) and SP (small, with short, memorable, token sequences, phonetically-balanced). Models would be evaluated by comparing their outputs, on a token-by-token basis, with the majority verdict of a panel of listeners. Pilots experiments at Sheffield using 2 different corpora demonstrate the feasibility of this approach and highlight areas where our understanding of speech perception is deficient.

---

[1] Here, 'everyday speech' means natural utterances produced in conditions of additive noise and reverberation typical of human communication.

[2] Which I see as a project associated with Grand Challenge element c: *computational modelling of human language function*.