



P3-2

**Temporal integration as a consequence of multi-source decoding**

*Jon Barker<sup>1</sup>, Martin Cooke<sup>1</sup> and Dan Ellis<sup>2</sup>*

*1: Department of Computer Science, University of Sheffield, UK*

*2: Department of Electrical Engineering, Columbia University, USA  
j.barker@dcs.shef.ac.uk, m.cooke@dcs.shef.ac.uk, dpwe@ee.columbia.edu*

How do listeners integrate evidence for speech in order to support reliable identification? Much of our everyday listening experience is set against a background of other sources, and the evidence for a speech target frequently manifests itself as scattered time-frequency islands of high signal-to-noise ratio. Individual fragments, such as formant portions, typically contain insufficient information. However, it appears that relatively few fragments are needed to constrain speech hypotheses to a manageable number (Cooke and Green, forthcoming). Robust speech perception in noise appears possible if we could determine which fragments belong together. A potential solution exploits auditory scene analysis principles (Bregman, 1990; Cooke and Ellis, 2001), which seek to group evidence based on 'primitive' processes such as common onset and harmonicity. While computational instantiations of these techniques have been applied to simultaneous fragments with some success, it is hard to see how temporally-disparate elements can be integrated using such constraints. For instance, it has proved difficult to apply interpolation or extrapolation procedures in sequential grouping of formants and harmonics. Another view (Remez *et al.*, 1994) suggests that speech lacks sufficient coherence to enable primitive grouping, and that instead, listeners call upon prior knowledge of speech for successful interpretation. Such 'schema-driven' grouping is also part of Bregman's conception, but there it works in concert with primitive processes.

To test these ideas, we have developed a computational framework for sound source identification which decodes speech in the presence of arbitrary noise backgrounds (Barker *et al.*, 2001). The initial stage involves a partitioning of the signal into regions which appear to originate from the same source, based on primitive grouping factors. The second stage performs temporal integration of these fragments by employing a set of source models. In fact, the multi-source decoder explores all possible combinations of fragments in a computationally-efficient manner. In addition to a search of the space of speech model sequences, as performed by the usual Viterbi search employed in ASR, the algorithm carries out a simultaneous search through the space of fragment speech/background assignments. The likelihood of individual hypotheses is evaluated using missing data techniques (Cooke *et al.*, 2001). We present results illustrating the action of the decoder in situations where fragments are grouped using primitive features, and contrast this with its interpretation where no such features are available. [Supported in part by EC ESPRIT long term research project RESPITE (No. 28149).]

Bregman, A.S. (1990) *Auditory Scene Analysis*. Cambridge, MA: MIT Press.

Barker, J., Cooke M.P., and Ellis, D.P.W. (2001) Integrating bottom-up and top-down constraints to achieve robust ASR: The multisource decoder, *Proc CRAC*, Aalborg

Cooke, M.P., Green, P.D., Josifovski, L., and Vizinho, A. (2001) Robust automatic speech recognition with missing and uncertain acoustic data. *Speech Communication* 34, 267-285.

Cooke, M.P., and Green, P.D. (forthcoming) Auditory organisation and speech perception: pointers for robust ASR. In S. Greenberg and W. Ainsworth (eds.), *Listening to Speech*. Oxford: Oxford University Press.

Cooke, M.P., and Ellis, D.P.W. (2001) The auditory organization of speech and other sources in listeners and computational models. *Speech Communication* 35, 141-177.

Pearce, D., and Hirsch, H.G. (2000) The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. *Proc. ICSLP '00*, Beijing, 4, 29-32.

Remez, R. E., Rubin, P. E., Berns, S. M., Pardo, J. S., and Lang, J. M. (1994) On the perceptual organization of speech. *Psychological Review* 101, 129-156.