



P2-6

**A model of multi-resolution auditory scene analysis**

*Sue Harding and Georg Meyer*

*MacKay Institute of Communication and Neuroscience, Keele University, UK  
s.m.harding@cns.keele.ac.uk, g.f.meyer@cns.keele.ac.uk*

Human listeners are adept at understanding speech in the presence of other sounds, whereas this task remains difficult for computational speech recognition systems. A theoretical framework that may be used to explain human performance is auditory scene analysis (Bregman, 1990). The auditory environment is considered to be a scene that contains multiple sound sources, which can be segregated into separate perceptual streams using primitive, low-level cues. Suggested cues include frequency (both fundamental frequency, F0, and formant frequencies), intensity, and location (via interaural time difference, ITD, and interaural intensity difference, IID) as well as common amplitude modulation of frequency components. Computational auditory scene analysis (CASA) models (Cooke and Ellis, 2001) typically use such cues to decompose the environment into low-level features that are then grouped into separate streams prior to a recognition stage.

We argue that the perceptual data does not support this strictly hierarchical approach but is more consistent with a model that uses a primary, low-resolution spectro-temporal representation for speech in a passive pattern matching stage, augmented by secondary, high-resolution representations containing detailed information. Many of the representational features that underlie stream segregation and speech pattern matching are mutually exclusive: low-level segregation requires high-resolution representation in multiple domains while robust speech pattern matching requires a representation that removes these sources of speaker and environmental variability. For example, cues needed for segregation on speaker position or F0 require high temporal or spectral resolution representations, while speech pattern matching would probably need information on formant transitions and amplitudes, which would be extracted from a more coarse spectro-temporal representation. Much experimental evidence (see Harding and Meyer, 2001) supports such a low-resolution representation for speech, showing that human listeners can tolerate massive distortions of speech signals with little reduction in intelligibility. We assume that a pattern-matching process acts on a low-resolution representation of the incoming auditory signal, comparing the representation with known patterns and highlighting those parts of the signal that do not match the patterns well. High-resolution analysis would occur in parallel and provide cues for segregation, localisation and speech segmentation; these detailed representations could be used both to resolve ambiguities in the low-resolution patterns and to analyse unmatched portions of the signal.

We have developed a simple computational model using hidden Markov models and a smoothed cepstral representation of speech that segregates the incoming auditory signal according to the similarity of the signal to previously learned speech patterns. We present the results of testing this model on vowel-nasal syllables combined with non-speech sounds such as pure tones and noise. The model performs particularly well, even at low signal-to-noise ratios, under circumstances in which the approximate position of the phoneme boundary is known. [Supported by EPSRC studentship award no. 99304828 and EU project SPHEAR.]

A. S. Bregman (1990) *Auditory Scene Analysis: The Perceptual Organization of Sound*. MIT Press.

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

Harding S. and Meyer G. (2001) A case for multi-resolution auditory scene analysis. Aalborg, Denmark: *Proc. Eurospeech 2001* 1, 159-162.