*Presented to One Day Meeting for Young Speech Researchers, University College London, UK, Thursday 14th April 2005*

# Physiologically motivated audio-visual localisation and tracking

*Stuart N. Wrigley and Guy J. Brown, Department of Computer Science, University of Sheffield, 211 Portobello Street, Sheffield, S1 4DP*

The localisation and tracking of individual speakers in multi-participant conversational interaction is an important step toward the automatic analysis of meetings. The position and movements of participants provides information about group dynamics as well as giving an indication of meeting events. An audio-visual localisation and tracking system is presented which draws its inspiration from neurobiological processing. The meeting room is recorded using a KEMAR binaural manikin and a single camera placed directly above the manikin. Source localisation from the binaural audio and face, object and motion locations from the video frames are used as input to two linked neural oscillator networks. The strength of the connections between the two networks determines the mapping between activity at a particular audio azimuth and activity at a particular visual frame column. A Hebbian learning rule is used to establish the connection strengths. The combined network segments the video and audio features and then produces audio-visual groupings on the basis of common spatial location. The audio-visual groupings are tracked through time using a mechanism based upon that of the human oculomotor system which incorporates smooth pursuit and saccadic movement.