Presented to the British Society of Audiology Short Papers Meeting on Experimental Studies of Hearing and Deafness,
University College London, Thursday and Friday, September 16th and 17th 2004

# Audio-visual source localization and tracking using a network of neural oscillators

*S.N. Wrigley and G.J. Brown, Speech and Hearing Research Group, Department of Computer Science, University of Sheffield, Regent Court, 211 Portobello Street, Sheffield, S1 4DP*

The objective of the MultiModal Meeting Manager (M4) project is to produce a system to enable structuring, browsing and querying of an archive of automatically analysed meetings recorded in a room equipped with multimodal sensors. For applications such as video conferencing, it may be desirable to automatically localise, and then track, an individual speaker. This allows a reduction in the video bandwidth requirement since only the portion of the frame surrounding the speaker needs to be transmitted.

This work focuses on using video features from a single camera in combination with binaural audio cues captured from a KEMAR manikin to localize and track a subject within a physiologically plausible framework.

The video cues consist of three features which can be extracted using simplistic image analysis techniques: binary masks representing the presence of objects, motion and faces. Regions of the video frame containing objects or motion can be obtained by subtracting the current frame from a reference frame or the previous frame respectively. Regions containing faces are found by identifying contiguous areas of skin coloured pixels which are oval in shape.

A single audio cue is extracted from the binaural recordings – source azimuth. Sound sources located away from a subject's midline reach the ears at slightly different times – interaural time difference (ITD). The ITD of a signal, and hence its azimuth, can be estimated by calculating the cross-correlation of the left and right audio channels.

The core of the model comprises two linked networks of neural oscillators which perform audio-visual segregation on the basis of oscillatory correlation (Wang, 1996). One network is two-dimensional in which each node corresponds to a video frame pixel and the second network is one-dimensional in which each node represents activity at a particular audio azimuth.

Since the camera causes distortion to the image and does not provide a 180 degree field of view, a Hebbian learning phase has been incorporated to learn a mapping between audio azimuth activity and activity in a particular range of video frame columns.

The network successfully groups video and audio activity when at the same position and segregates incongruous audio and video data. Work is also concentrating on employing attentional processes within the oscillator networks to investigate physiologically plausible tracking behaviour and competition between segregated sources.

**Acknowledgement**
This work was conducted as part of the MultiModal Meeting Manager (M4) project which is funded by the EU IST Programme (project IST-2001-34485).

**Reference**
Wang, DL (1996). Cognitive Sci 20:409-456.