# Physiologically motivated audio-visual localisation and tracking

**Stuart N. Wrigley and Guy J. Brown**

Speech and Hearing Research Group, Department of Computer Science, University of Sheffield, UK

{s.wrigley,g.brown}@dcs.shef.ac.uk

http://www.m4project.org

http://www.amiproject.org

## Introduction

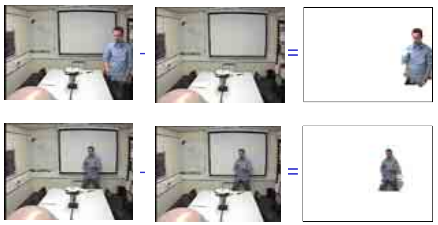
Many studies have employed neural oscillators for single modality segregation. Few have examined their utility in computational models of across-modality binding. Hence, we investigated neural oscillator based audio-visual grouping using a localisation and tracking problem.

Audio information from a KEMAR binaural manikin, visual cues from a single camera, placed directly above the manikin. The goal is to determine the spatial location of an individual participant and track that participant through time.

## Video segmentation

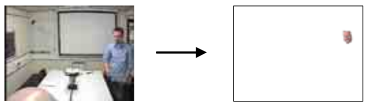
Object and Motion detection

Calculate the frame difference between either reference frame (objects) or previous frame (motion).




Face detection

Contiguous, oval regions of skin coloured pixels. Pixel is skin coloured if it falls within a certain RGB range[1].



For all features, detected regions below a certain size are discarded.

[1] F. Solina *et al.*, "Color-based face detection in the '15 seconds of fame' art installation," in *Proc. Mirage*, 2003.

[2] D. L. Wang, "Primitive auditory segregation based on oscillatory correlation," *Cognitive Science*, 20, 409–456, 1996.

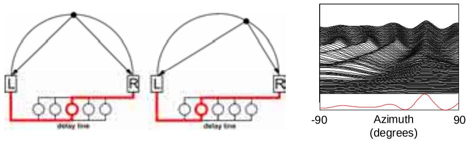
[3] S. de Brouwer *et al.*, "What triggers catch-up saccades during visual tracking," *J. Neurophysiology*, 87, 1646–1650, 2002.

## Audio localisation

Cochlear filtering is performed by 64 gammatone filters with centre frequencies equally spaced on the ERB scale between 50 Hz and 8 kHz.

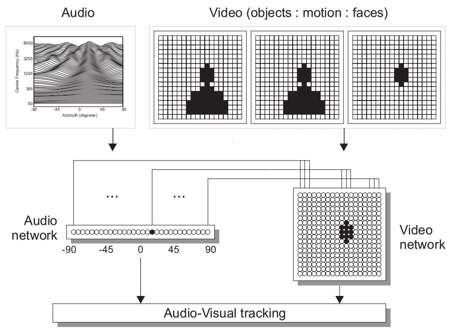
Auditory nerve firing rate is approximated by half-wave rectifying and square root compressing the output of each filter.

Signal ITD estimated by cross-correlation of the left and right auditory nerve response approximations.



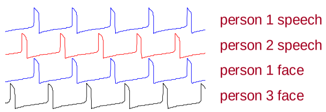

Precomputed ITD:Azimuth mapping used to calculate the signal's lateralisation in degrees.

## A-V Model




## Oscillatory correlation framework

A possible solution to the binding problem is temporal correlation (i.e. synchrony). The oscillatory correlation theory[2] suggests that neural oscillations are responsible for encoding the synchrony between features.

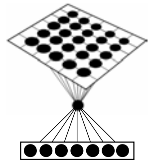



## Neural networks

Video network: 720x576 grid of neural oscillators in which each node corresponds to a particular frame pixel. Excitatory connections are placed between stimulated neighbouring nodes.

Audio network: 181 neural oscillators in which each node corresponds to a particular audio azimuth from -90° to 90°.

Each oscillator feeds excitatory input to the global inhibitor. The global inhibitor, in turn, feeds inhibitory input back to each oscillator. This ensures only one block of synchronised oscillators can be active at any one time.



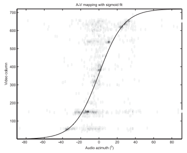
## Audio-Visual mapping

The camera introduces image distortion and does not provide a 180° field of view.

Hebbian learning phase used to learn a mapping between audio azimuth activity and activity in a particular range of video frame columns.

Training data consists of a subject speaking at 10° intervals around the manikin whilst video recorded.

A-V mapping determines the connection weights between nodes in the video network and nodes in the audio network

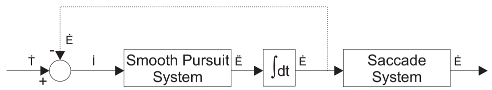


## Oculomotor tracking

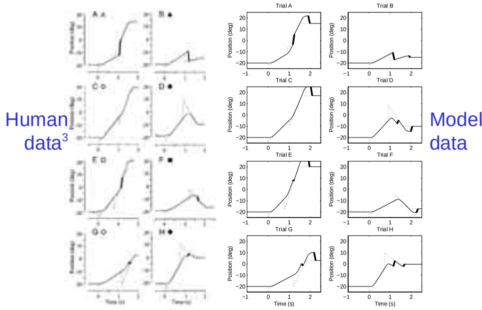
Inspired by the human oculomotor system incorporating smooth pursuit eye movements (< 50 deg/s) and catch-up saccades (> 500 deg/s).

Smooth pursuit modelled as a leaky integrator corresponding to an internal representation of target velocity.

Catch-up saccades overcome delays in the visual pathway.



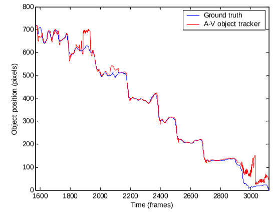

## Oculomotor model evaluation




Dotted line: stimulus; Solid line: eye movement; Thick line: saccade.

## Audio-Visual evaluation

Single participant walking around meeting table a speaking at 10° intervals.




Mean error per frame: -9.8 pixels; Face width: 26 to 46 pixels (dependent on distance).

## Conclusions

A network for audio-visual localisation and tracking has been described which uses audio azimuth (from binaural recordings) and face, motion and object location extracted from video frames.

The neural oscillator system can successfully identify the audio-visual locations of active speakers.

The oculomotor model accurately tracks participants.

Work is currently concentrating on improving the tracking in multi-participant environments by incorporating attentional factors.

M4 is supported by the EU IST Programme (project IST-2001-34485).