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

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

MultiModal Meeting Manager

Introduction

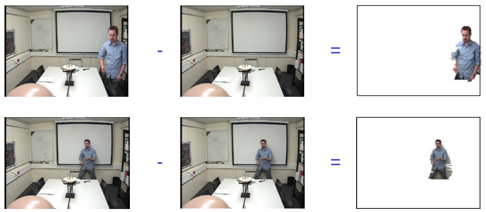
The objective of the M4 (multimodal meeting manager) project is to produce a demonstration system to enable structuring, browsing and querying of an archive of automatically analysed meetings recorded in a room equipped with multimodal sensors.

Goal: produce a system capable of localising and tracking one or more speakers using both binaural audio and video cues in a physiologically plausible manner.

Video segmentation

Object and Motion detection

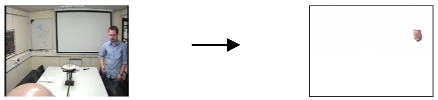
Calculate the frame difference between either reference frame (objects) or previous frame (motion).



Face detection

Contiguous, oval regions of skin coloured pixels. An RGB pixel is classified as skin if (Solina *et al.*, 2003†):

R>95 && G>40 && B>20 && — *eliminates gray*
(max_{RGB} - min_{RGB})>15 && — *ensures fair complexion*
abs(R-G)>15 &&
R>G && R>B — *red component must be the largest*



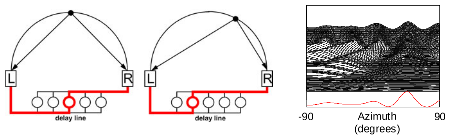
† Solina *et al.* (2003). Color-based face detection in the "15 seconds of fame" art installation. *Proc. MIRAGE 2003.*

Audio localisation

Cochlear filtering is performed by 128 gammatone filters with centre frequencies equally spaced on the ERB scale between 50 Hz and 8 kHz.

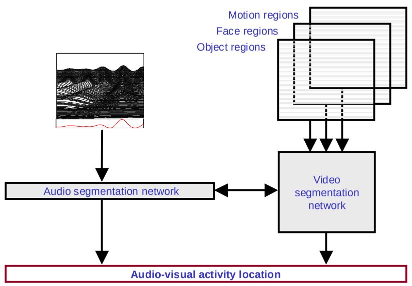
Auditory nerve firing rate is approximated by half-wave rectifying and square root compressing the output of each filter.

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



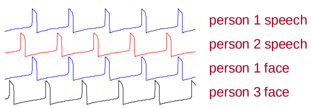
Precomputed ITD:Azimuth mapping used to calculate the signal's lateralisation in degrees.

A-V Model



Oscillatory correlation framework

A possible solution to the binding problem is Temporal Correlation (i.e. synchrony). The oscillatory correlation theory (Wang, 1996) suggests that neural oscillations are responsible for encoding the synchrony between features.



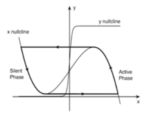
Relaxation oscillators

Reciprocally connected excitatory unit and inhibitory unit whose activities are represented by **x** and **y**.

Conceptually, an oscillator can represent the mean activity of a population of neurons or the behaviour of a single neuron's membrane potential and ion channels.

$$\dot{x} = 3x - x^3 + 2 - y + I_0$$

$$\dot{y} = \varepsilon\left[\gamma\left(1 + \tanh\frac{x}{\beta}\right) - y\right]$$

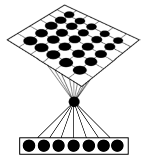


Neural networks

Video network: 72x58 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.



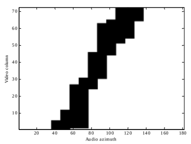
Audio-Visual mapping

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.

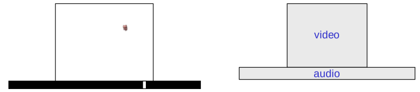
A-V mapping determines the connection weights between nodes in the video network and nodes in the audio network



Segmentation results

The network successfully groups video and audio activity when at the same position and segregates incongruous audio and video data.

Example of consistent A-V



Example of inconsistent A-V

Future work

The video feature of motion can be used to enhance the reliability of the audio azimuth estimates.

Initially, the amount of motion could simply be used to control the degree of smoothing applied to azimuth estimates of *n* previous and subsequent time frames. High motion = low smoothing.

Ultimately, the video motion information could be integrated into the azimuth estimation algorithm and used to determine the degree of temporal integration for particular azimuth ranges.

Work is also concentrating on employing attentional processes within the oscillator networks to investigate physiologically plausible tracking behaviour and competition between segregated sources.

Conclusions

A network for audio-visual segmentation and segregation 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.

Work is currently concentrating on improving the robustness of audio azimuth estimates and incorporating attentional factors.

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