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@dcs.shef.ac.uk, g.brown@dcs.shef.ac.uk

Abstract

An audio-visual localisation and tracking system for meeting scenarios is presented which draws its inspiration from neurobiological processing. Meetings are recorded by 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.

1. Introduction

In order to produce a representation of an object, the brain must provide a solution to the *binding problem*: how does the brain, confronted with many features, encoded in many different regions, draw them all together to form a perceptual whole? This problem arises in regard to feature combination within a single modality (e. g. the binding of edges, textures and colours to form a visual image). However, the binding problem also concerns the broader issue of how to link features in *different* modalities, such as the association of a sound with a visual object and possibly even a smell.

One solution to the binding problem lies in the concept of an *assembly*: a large number of spatially distributed neurons [1]. An individual neuron can be a member of several assemblies (each representing a different perceptual object) and hence a mechanism is needed for identifying which cells belong to which assemblies. von der Malsburg [2] suggests that different assemblies could be distinguished by temporal synchronisation of the responses of their constituent neurons. In this scheme, segregation of perceptual objects is represented by the desynchronisation of different assembly responses, and each assembly is identified as a group of synchronised neurons. The advantage of synchronisation is that the extra dimension of phase allows many simultaneous assemblies, each being desynchronised with the others. In order to avoid the computational complexity of assessing spike train synchronicity, von der Malsburg and Schneider proposed a mechanism in which the mean discharge response of a pool of cells is represented by an oscillator [3]. In this manner, groups of features form wholes if their associated oscillators are synchronised and the oscillations of unrelated wholes are desynchronised.

Although a large number of computational studies have demonstrated the use of neural oscillators for segregation and grouping of objects within a single modality (for a review see [4]), few have examined their utility in computational models of across-modality binding. In light of this, the current paper investigates the use of neural oscillators for audio-visual grouping using a localisation and tracking problem. The audio-visual data was collected as part of the M4 (multimodal meeting manager) project[1], which is concerned with the automatic analysis of meetings. Audio information is acquired from a KEMAR binaural manikin and visual cues from a single camera, placed directly above the manikin. The goal of the system is to determine the spatial location of an individual participant and track that participant through time. Four cues are extracted from the data per frame: azimuth of the dominant audio source and masks for faces, objects and motion in the video frame.

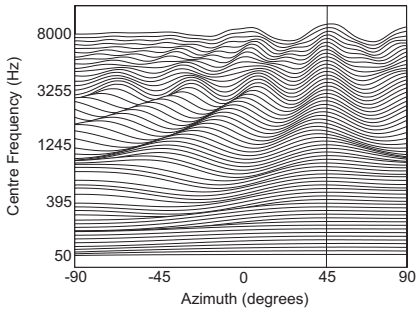

Figure 1: Cross-correlogram of a 155 Hz complex tone spatialised to the right by 45 degrees. The vertical line indicates the position of the 'spine'.

2. Feature extraction

2.1. Audio localisation

The acoustic inputs to each ear of the binaural manikin are sampled at a rate of 48 kHz and are processed by a model of the auditory periphery. The frequency selectivity of the basilar membrane is modelled by a bank of 64 gammatone filters [5] whose centre frequencies are spaced on the equivalent rectangular bandwidth (ERB) scale [6] between 50 Hz and 8 kHz. The auditory nerve response is approximated by half-wave rectifying and square root compressing the output of each filter.

Interaural time difference (ITD) is an important cue used by the human auditory system to determine the direction of a sound source [7]. The conventional technique for estimating the ITD of a signal is by calculating a cross-correlation function using

[1] http://www.m4project.org

the left and right channels in each frequency band:

$$C(i,t,\tau) = \sum_{k=0}^{N-1} a_l(i,t-k)a_r(i,t-k-\tau)w(k). \quad (1)$$

Here, τ is the delay and $a_e(i,t)$ is the simulated auditory nerve activity in channel i at time t for ear $e \in \{l,r\}$. The cross-correlation for channel i is computed using a 80 ms rectangular window $w(t)$ with lag steps equal to the sampling period (20.8 μs), up to a maximum lag of ± 1 ms. We note that this approach is equivalent to the neural coincidence model of Jeffress [8].

Computing $C(i,t,\tau)$ for each channel i gives a cross-correlogram, which is computed at 40 ms intervals resulting in a frame rate of 25 fps to match the video input. Since there may be small time differences between sounds reaching the two ears, channels dominated by a particular source will exhibit a peak at a correlation lag related to the azimuth of the source. For example, Fig. 1 shows the cross-correlogram for a 155 Hz complex tone which has been spatialised to the right by 45 degrees. When the sound source dominates a number of frequency channels, a characteristic 'spine' can be observed at the source azimuth. This can be enhanced by summing across frequency; the largest peak in the summary function then corresponds to the azimuth of the sound source.

2.2. Video features

Since the meetings are conducted in a relatively unchanging environment (i.e., the cameras are stationary and the lighting is consistent), a number of simple (and computationally efficient) video features are employed. The video frame is digitally captured and encoded using the 24 bit RGB colour model. For each pixel, 8 bits each are used for red, green and blue.

Visual objects are detected by calculating the difference between the current frame and a reference frame (usually found at the beginning of a recording when the room is still empty) and motion is detected by calculating the difference between adjacent frames. These difference images are thresholded to produce binary masks. In order to produce a binary mask for face regions, we identify those pixels whose RGB values satisfy the following function [9]:

$$R > 95 \wedge G > 40 \wedge B > 20 \wedge \Delta_{RGB} > 15$$
$$\wedge |R-G| > 15 \wedge R > G \wedge R > B \quad (2)$$

where $\Delta_{RGB} = max(R,G,B) - min(R,G,B)$. In each of the three masks, spurious pixels are discarded by using a region growth algorithm in which a pixel is only kept if its eight immediate neighbours are also 'on'. These candidate regions can still, however, be of any size and shape. To eliminate small regions, all groups whose area is less than a given size are discarded (300 pixels for skin-coloured regions and 3000 pixels for all others). An additional stage is included to produce the final face mask. To ensure that only face shaped (oval) regions remain, the length to breadth ratio is determined and used to discard non-oval regions.

3. Audio-Visual localisation

Two neural oscillator networks represent visual activity (2D network) and audio azimuth activity (1D network). The audio network has 181 nodes each representing an integer azimuth from -90 degrees to 90 degrees; the video network consists of a grid of 720 $\times$ 576 nodes in which each node represents a particular pixel of the binary input mask. The three video features are combined to produce a single binary input mask. If a face region and object region overlap then only the face region is included in the final mask (the object region is discarded). This ensures that face regions take priority over other region types. All remaining regions are included in the input mask. Fig. 2 shows a schematic of the system.

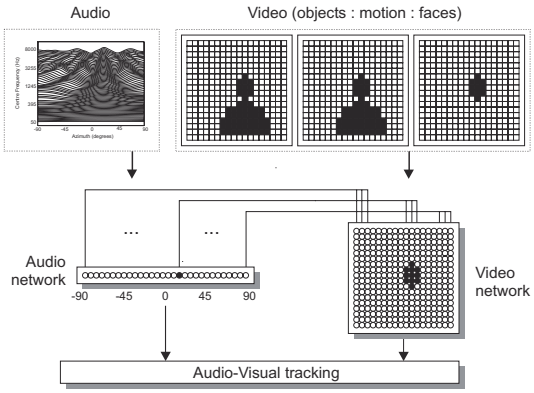

Figure 2: System schematic. The cross-correlogram provides azimuth input to the audio network and a combination of the three video features provide input to the video network. The audio-visual object locations are then used as input to the tracking system.

Each network consists of an array of oscillators based upon LEGION [10]. Within LEGION, oscillators are synchronised by placing local excitatory links between them. Additionally, a global inhibitor receives excitation from each oscillator, and inhibits every oscillator in the network. This ensures that only one block of synchronised oscillators can be active at any one time. Hence, separate blocks of synchronised oscillators (segments) arise through the action of local excitation and global inhibition. Thus, within-network segmentation emerges as a property of network dynamics.

In order to fuse related audio and video activity, the two networks are linked by a number of excitatory connections (placed between azimuth nodes and video columns). The strengths of these A-V connections are determined by a two-stage process. The first stage uses a Hebbian learning rule [11] during a training phase in which repeated, simultaneous video activity at column V and audio azimuth A strengthens the link between audio network node A and video network column V. However, since it is unlikely that the training phase will contain enough activity to generate weights for every possible audio-video pair, the second phase fits a sigmoidal function to the sparse A-V mapping data using the simplex search method [12] (see Fig. 3).

The building block of the network is a single oscillator, which consists of a reciprocally connected excitatory unit and inhibitory unit whose activities are represented by x and y, respectively

$$\dot{x} = 3x - x^3 + 2 - y + I_o + \rho \quad (3)$$

$$\dot{y} = \varepsilon \left[\gamma \left(1 + \tanh \frac{x}{\beta} \right) - y \right]. \quad (4)$$

Here ε, γ, and β are parameters, I_o represents the input to the oscillator and ρ is a noise term which assists desynchronisation among different oscillator blocks. The input I_o to oscillator is a combination of four factors: external input I_r, intra-network

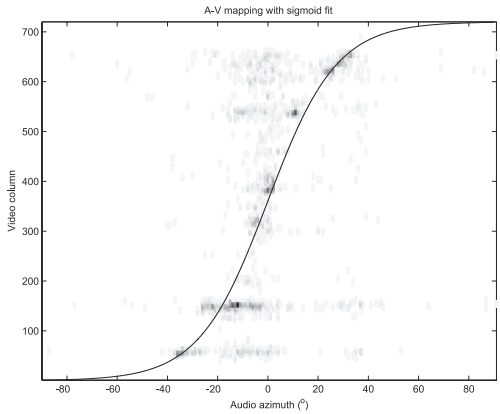

Figure 3: Audio-visual mapping function with coincident audio-visual activity superimposed. Grayscale value represents probability of occurrence.

activity, inter-network activity and global inhibition

$$I_o = I_r + \sum_{k \in N(j)} W_{jk} S(x_k, \theta_x) + \sum_{q \in M(j)} W_{jq} S(x_q, \theta_x) - W_z z. \quad (5)$$

Here, W_{jk} is the connection strength between oscillators j and k, N is the intra-network coupling neighbourhood (four nearest neighbours for video, two nearest neighbours for audio), M is the inter-network coupling neighbourhood (determined by the A-V mapping) and x_k is the activity of oscillator k. An oscillator j is stimulated ($I_r = 0.2$) if its corresponding binary mask input is 'on', otherwise it is unstimulated ($I_r = -5$).

The parameter θ_x is a threshold above which an oscillator can affect others in the network and W_z is the weight of inhibition from the global inhibitor. S is a squashing function which compresses oscillator activity to be within a suitable range,

$$S(m, \theta) = \frac{1}{1 + exp(-K(m - \theta))} \quad (6)$$

where K determines the steepness of the sigmoidal function. The activity of the global inhibitor is defined as

$$\dot{z} = \sigma_\infty - z \quad (7)$$

where $\sigma_\infty = 1$ if $x_k \geq \theta_z$ for at least one oscillator k, and $\sigma_\infty = 0$ otherwise. If $\sigma_\infty = 1$, $z \rightarrow 1$.

Following a short period of time required for the networks to converge on a stable segmentation result, the individual A-V groupings can be determined. Any audio and video network activities which occur at the same time (i. e., their oscillators are synchronised) are said to be grouped (forming 'A-V objects'). Remaining audio or video activity which occurs independently is said to be ungrouped. Any A-V objects are candidates for object tracking.

4. Audio-Visual tracking

Object tracking is implemented using a model inspired by the human oculomotor system. Smooth pursuit eye movements allow primates to follow moving objects with the eyes and are controlled by visual feedback. Such movements are relatively slow (eye velocity usually less than 50 deg/s). In contrast, saccades are fast eye movements (maximum eye velocity greater than 500 deg/s) that allow primates to shift gaze between stationary targets. However, due to delays in the visual pathway, it is occasionally necessary to combine smooth eye movements with 'catch-up' saccades to catch a moving target.

Figure 4: Video frame from the evaluation sequence.

In the model, smooth pursuit initiation occurs when the delayed velocity estimate increases above 1 deg/s. Similarly, termination occurs when the delayed velocity estimate drops below 1 deg/s. The smooth pursuit system is modelled using a leaky integrator to represent the velocity of an object

$$\ddot{E}(t) = g\left(\dot{I} - \dot{E}(t)\right) \quad (8)$$

$$\dot{I} = \left(T(t - \Delta t) - T(t - \Delta t - \delta)\right) fs \quad (9)$$

$$\dot{E} = \dot{E}(t - \delta) + \ddot{E}(t)\delta \quad (10)$$

where T is target position in degrees, $\dot{E}$ is eye velocity in deg/s and $\dot{I}$ is the target velocity in deg/s. t is the time in seconds, fs is the sampling rate and δ is the sampling period ($1/fs$). Δt represents the delay due to the visual pathway (100 ms in this model). In psychophysical studies, the decay of eye velocity after the termination of pursuit has been shown to be characterised by an exponential with a time constant of about 90 ms [13]. Thus, the decay time constant of the leaky integrator was set to approximately 90 ms ($g = 11$ in Equation 8). The maximum eye velocity is limited to 100 deg/s.

The saccade latency (the delay before a saccade occurs after the decision has been made to trigger one) is set to 125 ms [14, p. 1648]. However, in addition to this, a check is made immediately prior to the saccade occurring to confirm that a saccade is still required. Without this check, small and unnecessary saccades were observed. The saccade duration was fixed to be 74 ms [15, p. 1778].

A saccade is triggered when the retinal slip (relative motion of the target with respect to the fovea) is less than 5 deg/s and the positional error is greater than 1 deg ($RS < 5 \wedge PE > 1.1$). During a saccade, Equation 10 becomes

$$\dot{E} = \dot{E}(t - \delta) + \ddot{E}(t)\delta + \dot{J} \quad (11)$$

$\dot{J} = PE/SD$ represents the saccade velocity where PE is the positional error when the saccade is initiated and SD is the saccade duration. Behavioural studies have shown that the smooth motor command is not interrupted during catch-up saccades but is linearly combined with the saccade [15]. Thus, we sum the saccade velocity $\dot{J}$ and the smooth pursuit velocity. In

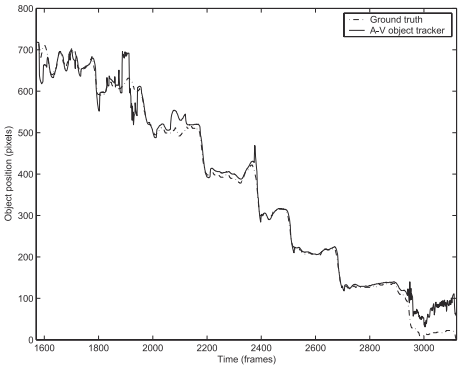

Figure 5: Tracking position over time compared with manually transcribed groundtruth position.

order to correct for the relatively large positional errors which occurred at fixation (the correction of which is impossible in a velocity-only model during fixation), saccades were permitted during fixation.

Information about target position is obtained from the A-V object closest to the current tracking position (if one exists). This is used to update the internal representation of the target motion (velocity and acceleration) and in turn the position of the tracking focus. During the lifetime of an object track, it is unlikely that audio localisation information will always be available: for example, binaural audio localisation is not robust (especially in reverberant environments) and speakers tend to make frequent pauses during speech. In this situation, the tracking algorithm 'backs off' to tracking the nearest video feature until audio information (and hence an A-V object) becomes available.

5. Evaluation

The system was evaluated using a recording of a single participant who moved around the room uttering a short phrase at 10 degree intervals. A representative frame is shown in Fig. 4. Fig. 5 shows the groundtruth and the audio-visual tracker positions for the frames in which the participant was visible. It is evident that the system tracks the participant with high accuracy; indeed, the mean error per frame across the entire sequence was only -9.8 pixels — much less than the width of a face (26 to 46 pixels depending on the distance from the camera). It should be noted that the relatively large errors at the beginning and end of the sequence are largely due to part of the participant's body and/or face being beyond the edge of the frame. In this situation, the automatic tracker 'backs off' to tracking the centre of the remaining body parts whereas the groundtruth shows the position of the visible portion of the face.

6. Conclusions

A neurobiologically plausible approach to participant localisation and tracking has been presented. A number of features are extracted from the audio-visual data which are then segmented and subsequently grouped by two, linked, neural oscillator networks on the basis of the features having originated from the same spatial position. Such A-V objects are then tracked through time using a mechanism which draws heavily on the neurobiological and psychophysical behaviour of the human oculomotor system. The results presented for a single participant show a high degree of accuracy in the system's tracking ability and hence, implicitly, its object detection and localisation. We also intend to use the system to track individual participants in multi-participant environments and current work is evaluating its performance in such situations. Preliminary results are encouraging.

7. Acknowledgements

Funded by the EU IST Programme project MultiModal Meeting Manager (M4; project IST-2001-34485). The authors thank Darren Moore, IDIAP, for his assistance during data collection.

8. References

[1] C. von der Malsburg, "Am I thinking assemblies?" in *Proceedings of the Trieste Meeting on Brain Theory*, G. Palm and A. Aertsen, Eds., Berlin, Germany, 1986.

[2] C. von der Malsburg, "The correlation theory of brain function," Max Planck Institute for Biophysical Chemistry, Göttingen, Germany, Tech. Rep. 81-2, 1981.

[3] C. von der Malsburg and W. Schneider, "A neural cocktail-party processor," *Biol. Cybern.*, vol. 54, pp. 29–40, 1986.

[4] D. L. Wang, "The time dimension for scene analysis," *IEEE Trans. Neural Networks*, in press.

[5] R. D. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, "An efficient auditory filterbank based on the gammatone function," Applied Psychology Unit, University of Cambridge, UK, Tech. Rep. 2341, 1988.

[6] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Res.*, vol. 47, pp. 103–138, 1990.

[7] J. Blauert, *Spatial Hearing — The Psychophysics of Human Sound Localization*. MIT Press, 1997.

[8] L. A. Jeffress, "A place theory of sound localization," *J. Comp. Physiol. Psychol.*, vol. 41, pp. 35–39, 1948.

[9] F. Solina, P. Peer, B. Batagelj, S. Juvan, and J. Kovac, "Color-based face detection in the '15 seconds of fame' art installation," in *Proceedings of Mirage*, INRIA Rocquencourt, France, 2003.

[10] D. L. Wang, "Primitive auditory segregation based on oscillatory correlation," *Cognitive Sci.*, vol. 20, pp. 409–456, 1996.

[11] D. O. Hebb, *Organization of Behavior*. New York: Wiley, 1949.

[12] J. C. Lagarias, J. A. Reeds, M. H. Wright, and P. E. Wright, "Convergence properties of the nelder-mead simplex method in low dimensions," *SIAM Journal of Optimization*, vol. 9, no. 1, pp. 112–147, 1998.

[13] D. A. Robinson, J. L. Gordon, and S. E. Gordon, "A model of the smooth pursuit eye movement system," *Biol. Cybern.*, vol. 55, pp. 43–57, 1986.

[14] S. de Brouwer, D. Yuksel, G. Blohm, M. Missal, and P. Lefèvre, "What triggers catch-up saccades during visual tracking," *J. Neurophysiol.*, vol. 87, pp. 1646–1650, 2002.

[15] S. de Brouwer, M. Missal, G. Barnes, and P. Lefèvre, "Quantitative analysis of catch-up saccades during sustained pursuit," *J. Neurophysiol.*, vol. 87, pp. 1772–1780, 2002.