Journal of Eye Movement Research
1(2):4, 1-17

Audio-visual integration during overt visual attention

Cliodhna Quigley
Neurobiopsychology, Institute of Cognitive Science, University of Osnabrück

Selim Onat
Neurobiopsychology, Institute of Cognitive Science, University of Osnabrück

Sue Harding
Speech and Hearing Group, Department of Computer Science, University of Sheffield

Martin Cooke
Speech and Hearing Group, Department of Computer Science, University of Sheffield

Peter König
Neurobiopsychology, Institute of Cognitive Science, University of Osnabrück

How do different sources of information arising from different modalities interact to control where we look? To answer this question with respect to real-world operational conditions we presented natural images and spatially localized sounds in (V)isual, Audiovisual (AV) and (A)uditory conditions and measured subjects' eye-movements. Our results demonstrate that eye-movements in AV conditions are spatially biased towards the part of the image corresponding to the sound source. Interestingly, this spatial bias is dependent on the probability of a given image region to be fixated (saliency) in the V condition. This indicates that fixation behaviour during the AV conditions is the result of an integration process. Regression analysis shows that this integration is best accounted for by a linear combination of unimodal saliencies.

Introduction

Operating under normal everyday conditions, the human brain continuously deals with often highly complex streams of input from different modalities. Attentional mechanisms allow the selection of a subset of this information for further processing. Overt attentional mechanisms, which involve the directed movement of sensory organs, play a particularly important role in human vision. Saccadic eye movements normally occur several times a second, and allow a sequential sampling of visual space by bringing different portions of the scene onto the fovea for maximum processing by the high density of photoreceptors found there. Eye movements are tightly linked to attention (Hoffman & Subramaniam, 1995; Awh, Armstrong & Moore, 2006) and measuring where people look when presented with an image is a useful means of characterising what makes a visual stimulus pre-attentively interesting, or salient (Parkhurst, Law & Niebur, 2002). One of the most influential models of overt visual attention is currently the saliency map model (Koch & Ullman, 1985; Itti & Koch, 2001). Saliency map models decompose a stimulus into some constituent features (luminance contrast or red-green colour contrast, for example), calculate the centre-surround differences at different scales within each feature space, and then linearly combine these topographically aligned feature values to yield the salience measure for each region of the stimulus. The most salient points selected by the model can be verified by comparison with human fixation behaviour in response to the same stimuli. Unimodal visual (Itti and Koch, 2001) and auditory (Kayser et al., 2005) models have been implemented, and their performance has been shown to correlate with human behaviour (Parkhurst et al., 2002; Kayser et al., 2005). As such, studying what makes a stimulus spatially and temporally salient under natural conditions can broaden our understanding of attentional mechanisms.

Although the integration of features within single modalities and the role of this integration in capturing attention has been actively studied, research into similar crossmodal integration processes is scarce. We know from psychophysical studies that spatially and temporally overlapping multimodal stimuli elicit faster and more

DOI 10.16910/jemr.1.2.4

ISSN 1995-8692

accurate responses in target detection tasks compared to unimodal stimuli (e.g., Frens et al., 1995; Corneil et al., 2002). These benefits are attributed to an integration of the individual modalities, and not just statistical facilitation due to parallel processing of the unimodal information, as the measured crossmodal responses are on average found to be faster or more accurate than the best measured unimodal response. The corresponding neural mechanisms have also been investigated, in particular the multisensory integrative neurons of the superior colliculus (Meredith and Stein, 1986; Bell et al., 2005), an area which plays an important role in orienting behaviours such as eye and head movements (Sparks, 1986). The response of collicular cells depends on the intensity of the stimulus: while many cells' multisensory responses can be modelled as a summation of the responses to the individual unisensory cues, a super-additive response to multisensory stimulation with near-threshold stimuli is also a well-known phenomenon (Stanford et al., 2005). Furthermore, reconsideration of the way in which early sensory processing areas are organised (Wallace et al., 2004) and recent work clearly showing multisensory interactions in early sensory areas (e.g., Kayser et al., 2007) have introduced new questions about when and where crossmodal integration might take place (Foxe & Schroeder, 2005), and how we should go about measuring and characterising it (Kayser & Logothetis, 2007). Finally, a great deal of the research to date uses very simple and controllable stimuli, and although it is undoubtedly of benefit to use parametric stimuli and a rigorously defined task, the reliance on simple artificial stimuli and reaction time or detection tasks has been criticised (De Gelder & Bertleson, 2003). In summary, we still lack an understanding of how information from different modalities is integrated for the control of attention, particularly under natural conditions.

We are interested in using subjects' visual behaviour to infer something about the underlying integrative process: How is the information from different modalities combined in order to guide overt attention? We propose three hypotheses for this process, illustrated in Figure 1. The end of this process will always involve the selection of a visual location to be attended, via a winner-takes-all operation (WTA). Each hypothesis can be characterised by the shape of the interaction surface created when the unimodal and multimodal saliencies, namely the probability of a given region of space to be fixated under the corresponding stimulation condition, are plotted against each other (right column of Fig. 1). The first hypothesis we consider (Maximum) involves a direct competition between modalities, and the maximally salient region among both modalities is chosen for fixation. This is reflected in a sublinear integration surface. There is no integration at work here; it is a pure competition between unimodal saliencies. Such a competition or race model is an important benchmark in cross-modal research, and in reaction time experiments a violation of the model is taken as an indication of an integrative process (Miller, 1982). We will use the Maximum hypothesis analogously here. The second hypothesis (Linear Integration) is of a linear integration of the unimodal saliencies and can be modelled as a weighted summation, with a corresponding planar surface used to model the multimodal saliency as a function of unimodal saliencies. As mentioned above, multimodal responses measured from single neurons in superior colliculus have been found to approximate the sum of the constituent unimodal responses, and a statistically optimal linear integration has also been found in human estimates of object properties provided by multimodal cues (Ernst & Banks, 2002). The final proposed scenario (Supra-linear Integration) involves an expansive non-linearity. Such a multiplicative interaction has been reported at the neural level in superior colliculus in response to threshold artificial stimuli, as mentioned

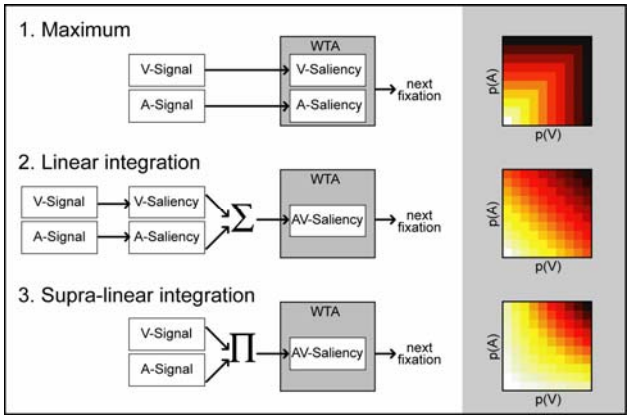

Figure 1. Three hypotheses for audiovisual interaction. The processing leading up to the winner-takes-all selection of the next location to be fixated is shown for each hypothetical integration. The rightmost column (grey background) depicts the corresponding interaction plot for each scheme – visual saliency is plotted on the x-axis, auditory saliency on the y-axis, with colour value (increasing from light to dark) indicating the magnitude of audiovisual saliency for a stimulus region with given visual and auditory saliencies. See text for more detail.

DOI 10.16910/jemr.1.2.4

ISSN 1995-8692

earlier. Alternatively, this pattern of integration could be due to an earlier facilitatory or modulatory effect of the signals themselves, possibly in early sensory cortices before estimates of unimodal saliencies are complete. At the behavioural level, the latency and accuracy of saccades to auditory white noise and LEDSs presented at various signal-to-noise ratios has been found to be best explained by a multiplicative audio-visual integration (Corneil et al., 2002).

Here we report the results of an eye-tracking study carried out using natural visual and auditory stimuli in order to determine the nature of the audiovisual interaction involved in the bottom-up control of human eye movements. By measuring the eye movements of 32 subjects presented with unimodal visual (V), auditory (A) and combined audiovisual (AV) stimuli, we can empirically determine saliency maps for the stimuli presented in each condition, and use these to directly assess the crossmodal interaction. Previous work by our group (Onat et al., 2007) used this approach with natural stimuli: images of forest scenes and bird-songs presented from the left or right of the computer screen. The present experiment expands the categories of visual stimuli used, seeks to avoid any semantic relation between images and sounds, and increases the number of sound sources and their spatial dimensionality by using a simulated auditory space.

Methods

Subjects and Recording

Participants were recruited by advertisement at the University of Osnabrück, and were screened for hearing deficits before taking part in the experiment. Eye movement data of 32 subjects (13 male; aged 19-36, median age 23) was recorded. All had normal or corrected-to-normal vision, normal hearing, and were naive concerning the exact purpose of the experiment. Informed written consent was given by each participant, and the experiment was conducted in accordance with Declaration of Helsinki guidelines. Subjects received payment for participation (€5) or credit towards a university course requirement.

All experiments took place in a dimly-lit room dedicated to eye tracking. The Eye Link II head-mounted camera-based eye tracker (SR Research, Ontario, Canada) was used to record binocular eye movements in pupil-only mode, at sampling rate of 500 Hz and nominal spatial resolution of 0.01°. Default velocity, acceleration, motion and duration thresholds (30°s-1, 8000°s-2, 0.1°, 4ms) were used to define saccades.

A chin rest was offered to subjects to assist them in remaining as still as possible during the experiment, and was used by all participants. Before the experiment began, the system was calibrated using an array of 13 fixation spots presented serially in a random order, until an average absolute global error of less than 0.5° was reached. Subjects were also required to fixate a central fixation point before each stimulus presentation, which allowed for online correction of any small drift errors during the experiment. The background colour of the calibration and drift correction screens was set to the mean luminance of the images used in the experiment.

The experiment was run using a Python program running on a high-speed computer (Apple Mac Pro: Apple, CA, USA). Visual stimuli were shown on a 30” display (Apple Cinema HD display: Apple, CA, USA) at a resolution of 2560 × 1600 pixels. The images did not fully extend to the horizontal edges of the display and these areas were covered with an aperture made from thick black card. Subjects were seated 60 cm from the display with the centre of the screen approximately at eye level, meaning that the images subtended approximately 44° × 36° degrees of visual field. Auditory stimuli were presented through in-ear binaural earphones (ER-4B: Etymotic Research, IL, USA) via an external audio interface (Transit High-Resolution Mobile Audio Interface: M-Audio, CA, USA) at a comfortable listening volume chosen by the subject before the experiment using a white-noise test stimulus.

Stimuli

Images were of natural indoor or outdoor scenes captured using a high-quality digital camera (DSC-V1 Cyber-shot: Sony, Tokyo, Japan). 24 greyscale bitmap images (1944 × 1600 pixels, 12 indoor and 12 outdoor scenes) were used in the experiment (see Fig. 4 for sample stimuli) and were chosen with the aim of avoiding central presentation of any distinctive objects, while creating scenes that were interesting along their full spatial extent.

Each four-second auditory stimulus contained a single sound sustained over the complete duration, and were

DOI 10.16910/jemr.1.2.4

ISSN 1995-8692

deliberately chosen so that the underlying auditory scene could not be well-identified. 18 basis sounds, selected for their steady-state spectrogram and verified by a human listener, were taken from a collection recorded for the experiment at the University of Sheffield, which were created by manipulating various household and office objects in a repetitive and continuous way, for example dragging a hairbrush over a cork surface, or slowly unfolding a piece of aluminium foil. The sounds are described in more detail in the Appendix. A later experiment with three naive listeners, who did not take part in the eye-tracking experiment, asked participants to first describe each sound in their own words, and then to judge whether pairs of sounds were the same or different (18 pairs, of which 12 were randomly paired mismatched sounds, and 6 were matched pairs, with matching balanced over subjects). The first part of this test confirmed that the sounds could not be easily identified, with the three subjects adequately describing the manipulation and the objects for only 8, 8, and 4 of the 18 sounds, respectively. The range of responses given was very broad, including "crackling fire" and "footsteps in undergrowth", and thus suggests that the sounds were suitably ambiguous and could be used with all visual stimuli without introducing a specific semantic bias. However, listeners could still differentiate between the sounds, with all three participants correctly classifying all pairs as same or different. As opposed to white noise, the sounds were perceived distinctly and can be considered natural.

The recordings of auditory stimuli were made in an acoustically isolated booth (IAC 400-A: IAC, Staines, UK) using a single microphone (B&K type 4190: Bruel & Kjaer, Denmark). The recorded signal was pre-amplified (Nexus 2690 microphone conditioning amplifier: Bruel & Kjaer, Denmark) and then digitised at 25 kHz (RP2.1 real-time processor: Tucker-Davis Technologies, FL, USA). The signals were later re-sampled to 44.1 kHz using polyphase filtering, as implemented in Matlab's inbuilt resample function. All further pre-processing and analysis was carried out in Matlab (The MathWorks, MA, USA).

The chosen sounds were normalised to approximately equal loudness using the a-weighting curve (IEC, 1979). The next step was to process the sounds so that they could be listened to through headphones and perceived as arising from a set of locations in 3-D space (for a more detailed account of the following procedure and alternatives, see Blauert, 1997). This "spatialisation" of the sounds was achieved using RoomSim (Campbell et al., 2005), an open source acoustic room simulation program. Given the dimensions of a rectangular room, the sound absorption properties of the room's surfaces, and the location of sensor and sound source, RoomSim computes an impulse response (IR) which models the source-to-sensor acoustic transformation for that particular environment. The program can also be used to create binaural room IRs by additionally incorporating a head-related transfer function (HRTF), which models the direction-specific spectral response characteristics of the two ears of a human listener. In order to use the computed IRs to spatialise a sound signal, the mono sound signal is simply convolved with each of the binaural IRs (corresponding to the two ears). These two resulting sounds are then used as the channels of a single sound that can be played through earphones.

Four different artificial sound sources were used, each with 20° horizontal (azimuth) and 22.5° vertical (elevation) offset relative to the listener, roughly corresponding to the four corners of the computer screen.

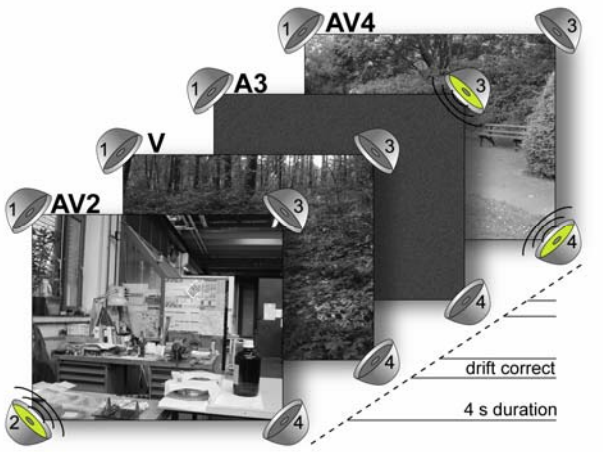

Figure 2. Procedure for the eye-tracking experiment. Each of the 96 trials lasted for four seconds and consisted of either an image (V), a sound with white noise image displayed (A1-4), or simultaneously presented image and sound (AV1-4). Sounds were spatialised to one of four locations, corresponding approximately to the corners of the screen (1: top-left, 2: bottom-left, 3: top-right, 4: bottom-right). Drift correction was performed before presentation of each stimulus. See text for further details.

DOI 10.16910/jemr.1.2.4

ISSN 1995-8692

We used the dimensions of the room in which the experiment was conducted, and chose surface absorption properties that produced realistic-sounding reverberation. A generic HRTF from the CIPIC collection (Algazi et al., 2001) was used, and a binaural room IR was created for each of the four sound sources. Each of the 18 sounds was convolved with each IR, yielding 72 spatialised auditory stimuli. As it was not known whether participants would be able to discern the difference between these artifical sources, each subject's spatialisation ability was determined in a post-experimental test as described below.

Experimental Procedure

An illustration of the eye-tracking experiment is shown in Fig. 2. The experiment consisted of unimodal (auditory: A, and visual: V) and multimodal (AV) conditions. Each auditory stimulus had one of four locations, which amounted to four different auditory and audiovisual conditions (A1-4, AV1-4). Participants' eye movements were measured in 96 four-second trials, of which 24 were visual, 24 (6 × 4) were auditory, and 48 (12 × 4) were audiovisual. Each image was seen three times by each subject (1 × V, 2 × AV), and each of the 18 sounds was heard once at each of the four possible sound locations. Static white noise images were displayed during auditory trials in order to limit subjects' fixations to the same spatial range as V and AV trials. The order of stimulus presentation and the assignment of auditory stimuli to A and AV trials were randomly generated for each subject, with the single constraint (to aid balancing within subjects) that repeated appearances of images in AV trials were accompanied by sounds from sources with common azimuth or elevation. This was a free-viewing experiment, with participants told only to "study the images and listen to the sounds carefully".

After the eye-tracking experiment, participants were briefly interviewed to determine whether they had realised that the sounds differed in spatial location, and were informed of the details of the spatialisation. In order to determine whether they had accurately perceived the spatial location of the auditory stimuli, participants then

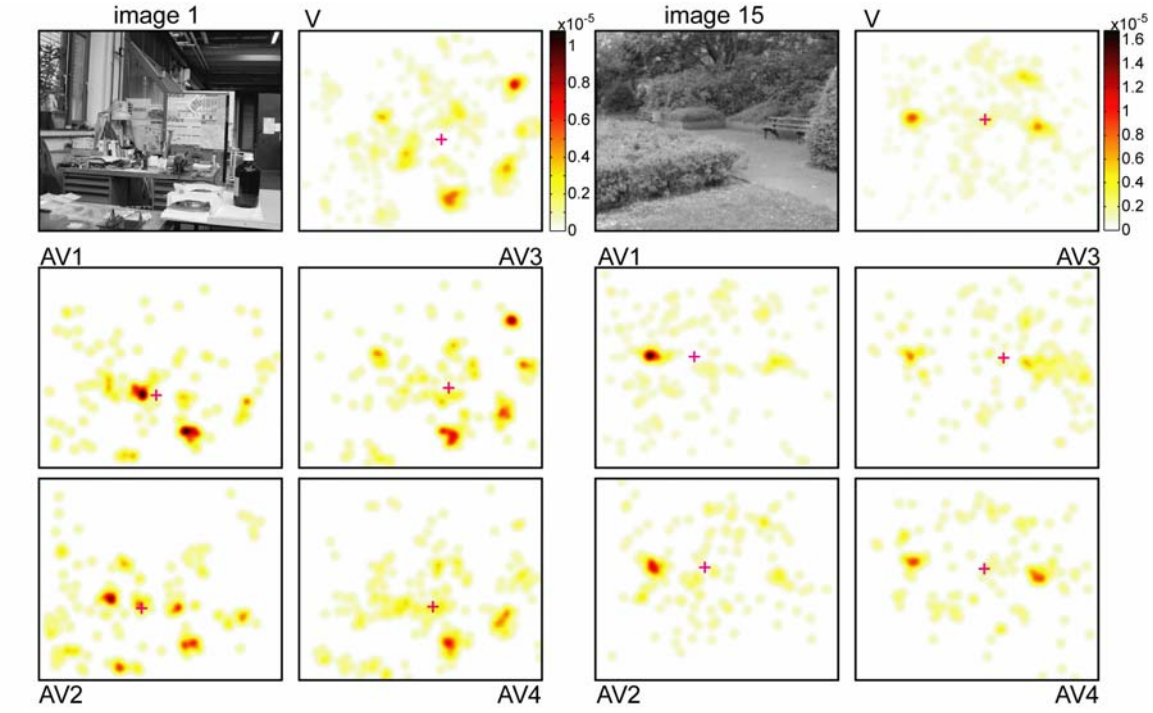

Figure 3. Data from two representative images. Two of the images used in the experiment are shown (original resolution was 1944 x 1600 pixels), along with the image pdfs constructed from eye movements of all subjects presented with this image in V and AV conditions. Probability of fixation is denoted by increasing colour; see colour bars provided with each image for corresponding scale (for visualisation purposes, V and AV pdfs are equalised to the same range). Centre of gravity of each pdf is indicated by a cross.

DOI 10.16910/jemr.1.2.4

ISSN 1995-8692

completed a four-alternative forced choice (4AFC) task, in which they listened again to each sound heard in the experiment and indicated their estimation of its location (upper-left, lower-left, upper-right, lower-right) by button press on marked keys on the numerical pad of a standard keyboard. The sounds were presented in a new random order for each subject, with the constraint that two different spatialisations of the same sound did not directly follow each other. Finally, subjects were debriefed about the purpose of the experiment and any questions they had were answered. The entire experiment lasted less than an hour, and was conducted in either English or German, depending on the participant's language preference.

Data Analysis

The recorded fixation data was used to create probability density functions (pdfs), which quantify the probability of a region of space to be fixated by a subject, as follows. For each condition, subject, and stimulus, the pixel location of each recorded fixation was marked on a raw fixation map. To smooth the data and to compensate for the finite precision of the eyetracker, these raw maps were convolved with a Gaussian of unit integral (full width at half maximum: 1.5° visual angle). First fixations were not included due to the bias introduced by the central fixation spot which preceded each trial. Image pdfs for V and each of the 4 AV conditions (corresponding to the 4 sound locations) were produced by averaging over subjects (see Fig. 3 for examples); subject pdfs were made by averaging over images (Fig. 4). In the case of the auditory conditions, pdfs were made by averaging over subjects and sounds, resulting in a total of 4 pdfs.

Pdfs were characterised by their centre of gravity, and centre of gravity shifts between V and AV conditions, used to characterise the shift in fixation density due to auditory stimulation, were calculated by simple subtraction. The horizontal influence of sounds was evaluated by comparing the distributions of horizontal shifts grouped by common azimuth, i.e. shifts measured using AV pdfs with sound sources located to the top-left and bottom-left (AV1 and AV2) vs. those arising from AV pdfs with right-side sound stimulation (AV3 and AV4). The vertical auditory effect was evaluated by comparing distributions of vertical shifts grouped by common AV elevation. Image pdfs reveal consistent biases in subjects' fixation behaviour when presented with the corresponding stimulus, and can thus be considered as empirical saliency maps. This plays an important role in subsequent analysis.

Two approaches were taken to model the audiovisual interaction and to test the hypotheses described in the Introduction. First, a multiple regression analysis was used to fit the unimodal saliencies (the empirical saliency maps constructed from the measured gaze data: p_A and p_V in the equation below) with the measured multimodal saliency (p_{AV}) according to the following equation:

$$p_{AV} = \alpha_0 + \alpha_1 \cdot p_V + \alpha_2 \cdot p_A + \alpha_3 \cdot p_{A\times V} \quad (1)$$

Least squares regression was used to solve this equation separately for each AV stimulus, yielding 24×4 (images×sound source locations) estimates for each coefficient (α_0 to α_3). The last term of the equation, $p_{A\times V}$, is the normalised (to unit integral) element-wise product of the unimodal saliencies p_A and p_V, and is assumed to approximate the outcome of a multiplicative cross-modal interaction.

The second approach involved creating a model of the interaction from the measured data, which was then evaluated in terms of the candidate interaction scenarios detailed above. As mentioned earlier, V and AV image pdfs constitute empirical saliency maps, and each point of the saliency map for a given image provides us with the saliency of that portion of stimulus space under the respective input condition. We also have a value for the saliency of visual space measured in the auditory condition, yielding a triplet of A, V, and AV saliency values for each pixel of each image. Data-driven integration plots were made by ordering these triplets in a three-dimensional V×A×AV space, thereby disregarding their spatial location and the stimulus from which they originated. Because of the high resolution of the images involved, the individual saliency maps were downsized by a factor of 4 (to 486×400 pixels) before creating the interaction plots. Saliency maps generally tend to have few highly salient areas, meaning the higher saliency areas of the V×A×AV space are sparsely populated. For this reason, we restricted our analysis to a subset of the space, which was chosen in order to maximise the amount of data included in the model, while still requiring adequate density of data for estimation of audiovisual saliency values (see Results for more details). Data within this range were binned, and the mean of each bin was used as an estimate for p_{AV}, and the bin centres

DOI 10.16910/jemr.1.2.4

ISSN 1995-8692

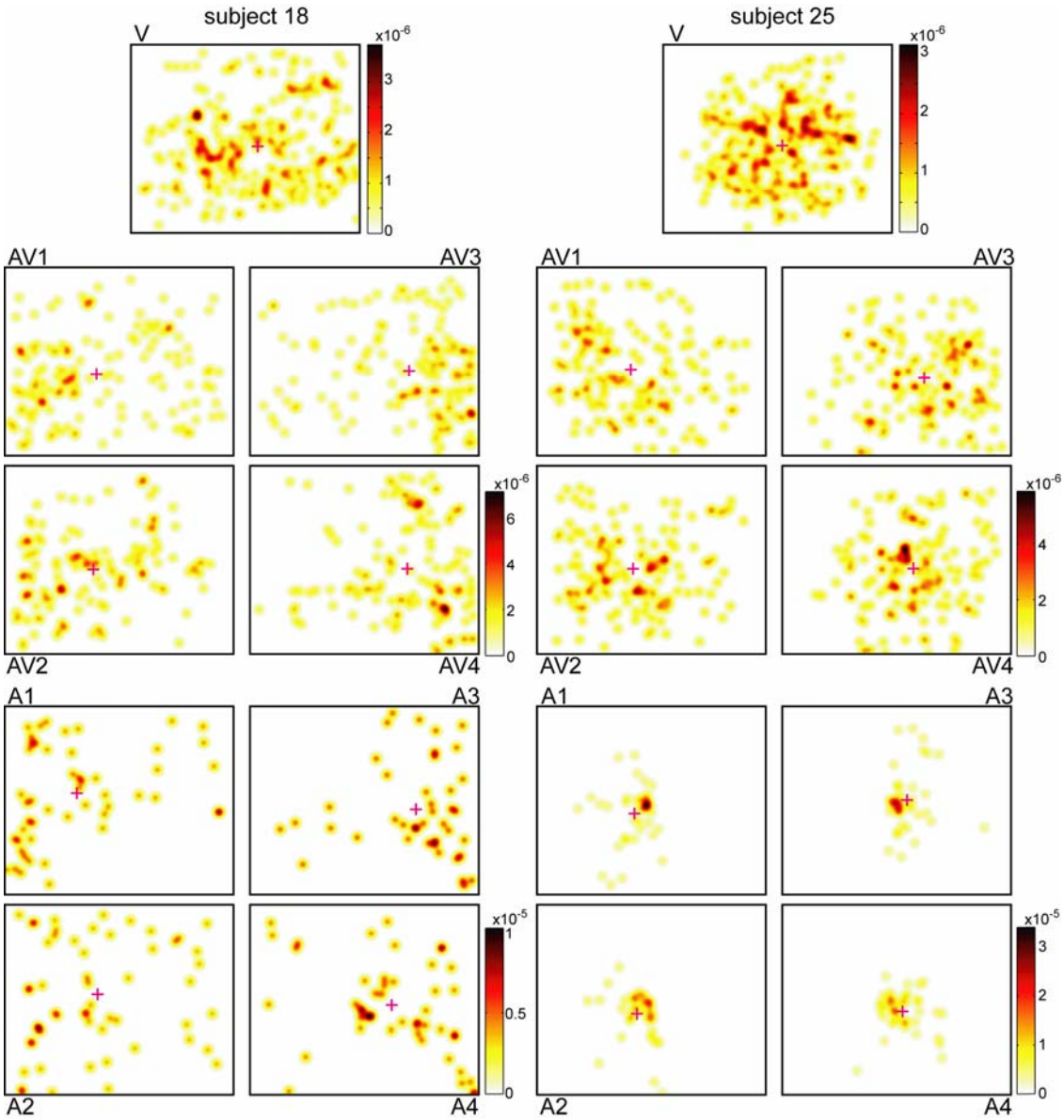

Figure 4. Data from two representative subjects. Pdfs constructed from eye movements of separate subjects are shown for each experimental condition: unimodal visual (V), unimodal auditory (A1-4), and multimodal audiovisual (AV1-4). Each pdf includes fixations made by the subject when presented with all stimuli of the corresponding condition. Probability of fixation is denoted by increasing colour; see colour bars provided with each condition and subject for corresponding scale. Centre of gravity is indicated on each pdf by a cross.

as estimates for p_A and p_V . A least squares analysis – weighted by the inverse variance of each bin – was then carried out to estimate the coefficients of the following equations:

$$p_{AV} = \beta_0 + \beta_1 \cdot p_V + \beta_2 \cdot p_A \qquad (2)$$

$$p_{AV} / G(p_{AV}) = \gamma_0 + \gamma_1 \cdot p_V / G(p_V) + \gamma_2 \cdot p_A / G(p_A) + \gamma_3 \cdot p_{A\times V} / G(p_{A\times V}) \qquad (3)$$

with G(•) indicating the geometric mean. The first equation models the multimodal saliency as a linear integration of unimodal saliencies, while the second

DOI 10.16910/jemr.1.2.4

ISSN 1995-8692

includes the multiplicative term $p_{A\times V}$. In the multiple regression analysis of (1) above, this term, which was created by multiplying the unimodal pdfs, could be normalised to unit integral in order to bring it into a range comparable to the unimodal pdfs. Here, the saliencies no longer belong to a structured probability distribution, so the saliency data belonging to each condition were normalised by their respective geometric mean ($G(p_C)$) in order to bring the saliency values into the same exponent range.

Results

Success of Sound Spatialisation

According to participants' responses in the brief interview after eye-tracking was completed, a horizontal difference in the auditory stimuli was clear to 25 subjects (almost 80%), but only 5 of those participants noticed any vertical difference in the spatial location of the sounds played during the experiment. A more objective measure of sound localisation ability is provided by the 4 alternative forced-choice task (4AFC). Performance in the 4AFC task was determined for each subject by calculating the percentage of correct responses given overall and for each sound location separately (Fig. 5A). In order to ensure that these values were not contaminated by any bias in response (e.g. always responding upper-left for any sounds originating from the left), the percentage of correct answers was also determined as a proportion of the number of times each response was given and is shown in Fig. 5B.

As can clearly be seen from Figure 5A, participants' overall success rates are far beyond a chance level of 25% (mean success rate: 65.8 ± 9.4% SD; by source: 1: 68.2±18.6%; 2: 60.8±14.8%; 3: 69.6±16.8%; 4: 64.6±14.5%). However, the majority of errors were due to a misjudgement of the elevation of a sound, with only 7 subjects making any errors regarding azimuth, e.g. assigning a sound to a source on the left instead of to the correct or incorrect right-lying source (4 subjects made only a single mistake; 3 subjects made 2, 3 and 9 errors respectively). Evaluating the performance of individual subjects, we found that each participant performed statistically better than chance (25%) when overall sucess is considered (exact Binomial test: $p<10^{-6}$). Taking into account the ease with which the subjects could discriminate horizontally between sources, chance performance in the vertical direction would yield a distribution centred at 50% success. Looking specifically at vertical discrimination, 21 of the 32 participants were beyond chance performance of 50% ($p<0.05$), but 11 subjects did not perform significantly better than chance. At the population level, considering each sound source separately, mean success rates of subjects were significantly higher than a 50% chance level (one-tailed one-sample t-test, by source: $p<10^{-5}$, $p<10^{-3}$, $p<10^{-6}$, $p<10^{-5}$). It seems then, that the spatialisation of the sounds was a success for this group of subjects, and that

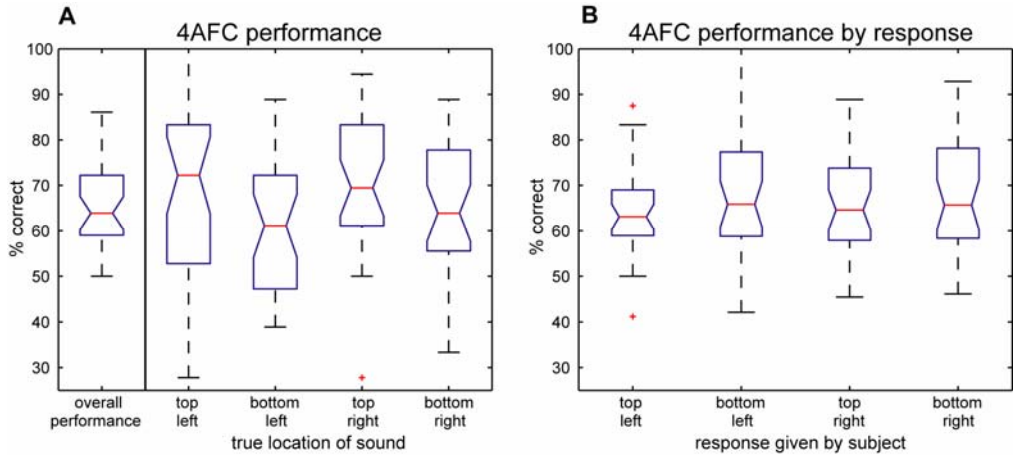

Figure 5. Success of spatialisation. The results of the four alternative forced-choice task (4AFC). A: Success of all subjects is given as the percentage of sounds from each location that were correctly localised. Left, overall results; right, results by source location. The horizontal red line within each box shows the median percentage correct of 32 subjects. Boxes extend to the lower and upper quartile values of each distribution, with whiskers showing the range of the data. Outliers are shown as red crosses. B: Percentage of times each response was given correctly.

DOI 10.16910/jemr.1.2.4

ISSN 1995-8692

they could, for the most part, discriminate sound location both in azimuth and elevation.

Effect of Auditory Stimulation on Fixation Behaviour

Looking at the example subject pdfs shown in Figure 4, we see that the fixation density of subjects tends to shift towards the location of the accompanying sound in multimodal trials. In order to systematically examine this effect, we compared the centre of gravity shifts between the V and AV pdfs belonging to each subject, and also separately for each image.

Figure 6A provides an overview of the centre of gravity shifts measured between each multimodal condition (i.e. grouped by sound location) and the corresponding visual condition, in polar co-ordinates. It is clear that there is a horizontal effect, and the mean shifts for each condition over both images and subjects seem to suggest that there may also be a vertical difference. Examining horizontal and vertical fixation shifts separately in Cartesian co-ordinates, we find that the horizontal effect is highly significant over both images (Wilcoxon rank-sum test, $p<10^{-14}$) and subjects ($p<10^{-12}$), as can be seen in Figure 6B. The vertical effect is however less evident. Statistically, it reaches significance only when the effect is considered over images ($p=0.0126$; over subjects: $p=0.0987$). In order to evaluate whether this constitutes a trend, the distributions of vertical shifts between conditions grouped by common elevation were also compared (i.e. vertical difference among left sources vs. right sources). For this grouping, the null hypothesis that both distributions of vertical shifts arose from the same underlying population cannot be rejected, with $p=0.9562$ for images and $p=0.9639$ over subjects. This gives us some indication of the strength of the vertical trend that is seen over images and subjects. In conclusion, fixation in multimodal conditions is biased towards the auditory source location.

Time Course of Effect

Next we determined when the effect was strongest during stimulus presentation. The two-dimensional Kolmogorov-Smirnov two-tailed two-sample test (Fasano & Franceschini, 1987; implemented according to the algorithm given in Press et al., 1992) was used to assess spatial differences between fixation distributions of the following condition pairs: firstly, each AV condition with V (namely AV1 vs. V; AV2 vs. V; AV3 vs. V; AV4 vs. V); secondly, each A condition with all the remaining A conditions (i.e., A1 vs. [A2,A3,A4]; A2 vs. [A1,A3,A4]; etc.). A temporal bin size of 250 ms was used, yielding 16 bins. Temporal intervals for which the null hypothesis was rejected ($p<0.05$) in at least 6 of the 8 pairs were considered to belong to the temporal interval of interest. As mentioned earlier, first fixations were not included in analysis, which left too few fixations in the first bin for some comparisons to be performed. The temporal interval of interest extended from the second bin (250ms) until 2500ms after stimulus onset, after which time subjects' viewing behaviour in different conditions became less spatially directed. The time course of the effect seen here is similar to that found in our previous study (Onat et al., 2007). Only fixations lying within this interval of interest were subject to further analysis.

Is it a True Integration?

The first of the hypotheses under consideration (see Introduction) is that the unimodal saliency maps do not combine; rather, the maximally salient point of the unimodal maps is chosen as the next fixation target. If the overt behaviour is driven exclusively by auditory information, then little or no similarity should be seen between visual and audiovisual fixation maps corresponding to the same image. In order to reject this hypothesis, and to confirm that the process at work does indeed integrate visual and auditory information, we measured the similarity between image-matched V and AV saliency maps using two approaches: correlation, with correlation coefficients squared to yield the r^2 statistic; and the Kullback-Leibler divergence, which quantifies the divergence of a posterior probability distribution (here an AV image pdf) from a prior distribution (in this case, a V pdf), meaning that lower values indicate greater similarity. Control distributions were provided by comparing pairs of V and AV saliency maps from different images.

Figure 7A depicts the r^2 values calculated for image-matched V and AV image pdfs (shown in light grey) and the control distribution created by comparing image-shuffled pairs (dark grey). r^2 for the image-paired V and AV empirical saliency maps are significantly higher than the control values ($p<10^{-23}$, Wilcoxon rank-sum), indicating that the distribution of fixations for each image is similar under purely visual and combined audiovisual stimulation. Figure 7B shows the Kullback-Leibler (KL)

DOI 10.16910/jemr.1.2.4

ISSN 1995-8692

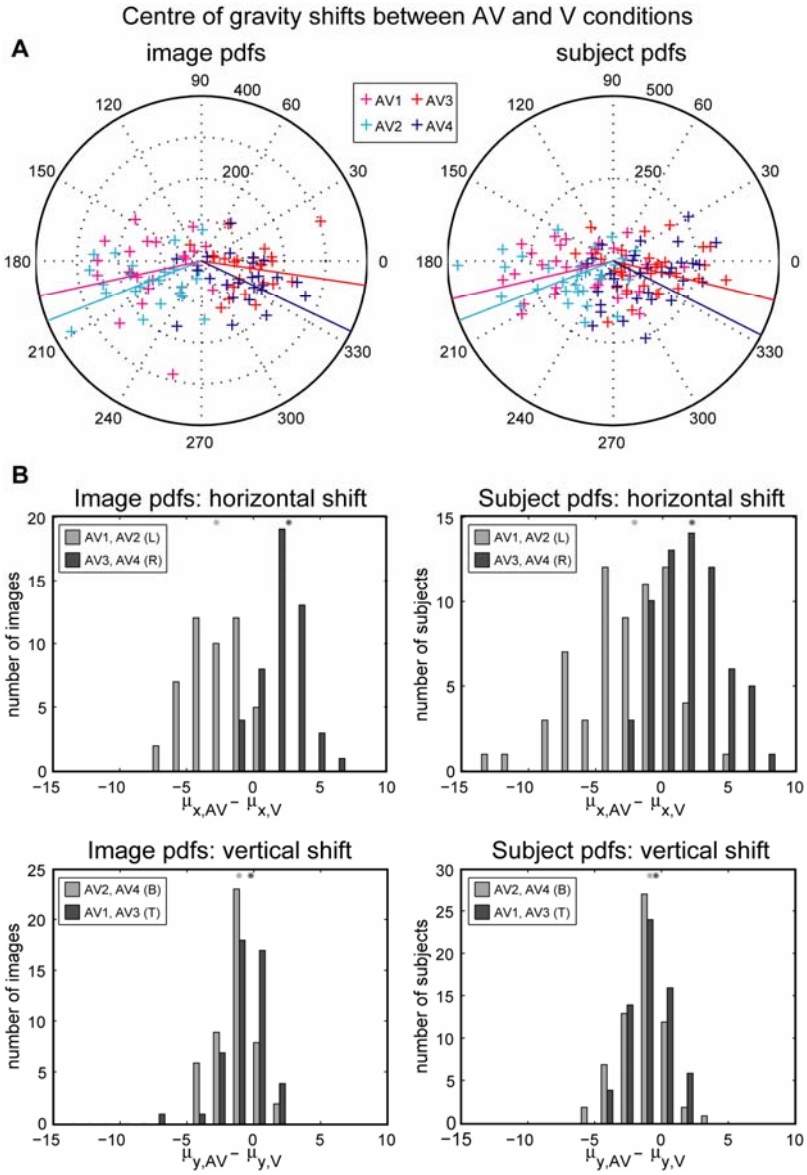

Figure 6. Shift in fixation towards sound location. The shifts in fixation density between visual (V) and each audiovisual condition (AV1-4) are calculated by a simple subtraction of the centres of gravity of image- or subject-matched visual and audiovisual pdfs. A: The distribution of fixation shifts are shown here in polar co-ordinates, for all image pdfs (left plot, N=24) and subject pdfs (right plot, N=32). Each point depicts the shift between a V and AV pdf for a single subject or image, with audiovisual condition coded by colour (see legend). The (vector) mean over all subjects or images for a single condition is drawn as a single line – the magnitude of these lines is purely for visualisation purposes and does not represent any variance within conditions. B: Histograms of horizontal (upper plots) and vertical (lower plots) shifts in fixation density for all image pdfs (left plots, N=24) and subject pdfs (right plots, N=32), calculated separately from Cartesian coordinates of centres of gravity. For horizontal shifts, the data are grouped by common azimuth of sound source (top-left and bottom-left sound sources in light grey vs. top-right and bottom-right in dark grey), with positive values describing shifts to the right. In the case of vertical shifts, data are grouped by common elevation (bottom/top sources in light/dark grey), and positive values indicate an upward shift. Values are given in visual degrees, and the median value of each distribution is indicated by an asterisk.

DOI 10.16910/jemr.1.2.4

ISSN 1995-8692

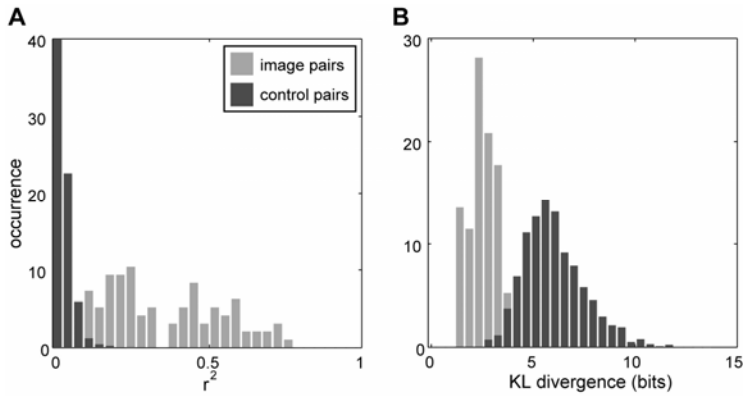

Figure 7. Fixation behaviour is similar in V and AV conditions. A: The distribution of r2 values calculated for image-paired V and AV saliency maps (light grey) and control pairings of V and AV image pdfs (dark grey). B: The distributions of Kullback-Leibler divergences of AV from V image pdfs, again with actual image pairings shown in light grey, and control pairings in dark grey. Actual and control distributions are significantly different in both cases.

divergence values. Again, the actual and control distributions are significantly different ($p<10^{-23}$), and the low values of the image-matched V/AV pairs suggest that fixation behaviour in the multimodal AV conditions does not differ greatly from that in the image-only V conditions. This similarity in fixation behaviour indicates that the visually salient information remains salient in AV conditions. Coupled with the evidence for an auditory effect seen above, this strongly supports the idea that there is indeed an integration at work.

How Much of Multimodal Saliency is Explained by the Unimodal Saliencies?

In order to evaluate the strength of contribution of unimodal visual and auditory saliency to the multimodal saliency of each stimulus, we used multiple regression analysis (see Methods). Applying this analysis to each combination of image-ordered A, V, AV and constructed A×V fixation maps, we created 24×4 models. The range of the goodness-of-fit of these image-specific models (r^2) was from 0.047 to 0.763, centred at a median of 0.308. As a control, we also created simpler models in which each image-based AV map was modelled as a linear combination of A and V saliency maps only. This yields the same pattern of results, and in fact, the two r^2 distributions do not differ significantly (2-sample Kolomogorov-Smirnov test, p=0.99).

The regression coefficients describing the contribution of each A, V, and constructed A×V saliency map are shown in Figure 8. All distributions are significantly different to each other (two-sample two-tailed t-test, $p<10^{-23}$ in all cases). The values for the first coefficient of the equation (α_0), which defines the intercept of the fitted model, are normally distributed and located close to but significantly different to 0 (one-sample two-tailed t-test, $p<10^{-23}$; mean ± standard deviation: $8.8\times10^{-8} \pm 5.5\times10^{-8}$), and are not shown in Figure 8. Looking at the distributions of the unimodal coefficients, we see that visual saliency contributes most to the multimodal saliency (α_1: 0.6188 ± 0.2047), followed by auditory saliency (α_2: 0.1454 ± 0.1239). The

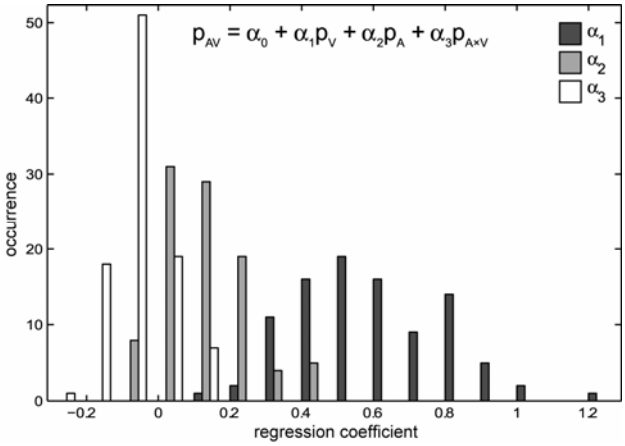

Figure 8. Histogram of the coefficients determined by least squares regression for the equation shown. Regression was performed for ecah AV stimulus, resulting in a total of 24 × 4 (image × sound location) estimates for each coefficient. Intercept coefficients (α_0) are centred close to 0 and are not shown. Visual coefficients (dark grey) show the greatest contribution to multimodal saliency, followed by auditory coefficients (light grey), and then the smaller and slightly negative multiplicative factor (white).

DOI 10.16910/jemr.1.2.4

ISSN 1995-8692

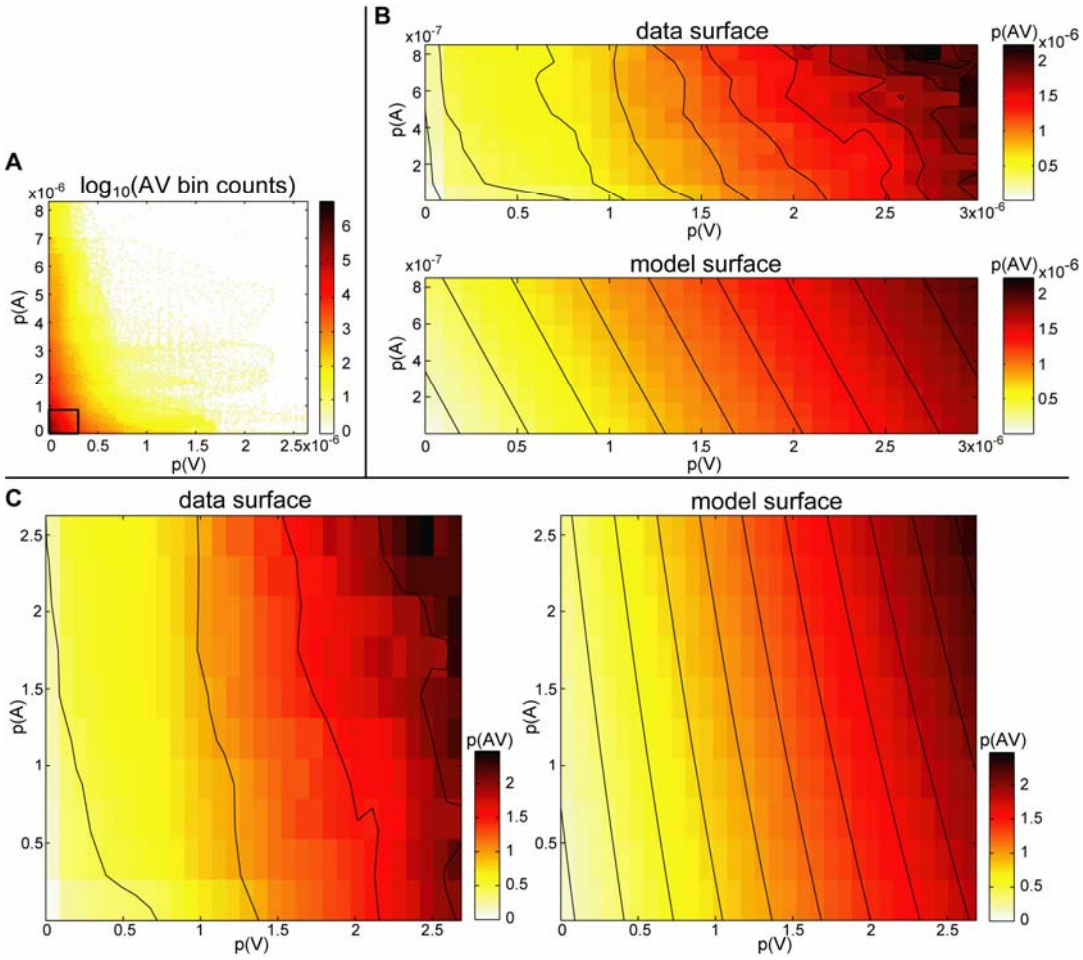

Figure 9. Interaction plots. A: The size of the complete data set is shown. Visual saliency is plotted on the x-axis and auditory on the y-axis. The number of measured AV data points falling within each bin of this space is indicated by the colour of that bin (log scale). Higher saliency areas of the space are more sparsely populated, and the data range chosen for further analysis is outlined in black. B: The resulting interaction plot for the chosen subset of data (corresponding to around 90% of total available data) is shown with iso-contours in the upper figure, with the expected value of multimodal saliency of each bin shown in colour. The shape of this surface describes the type of interaction process involved. Below is the model constructed from a linear combination of unimodal saliencies, weighted according to the coefficients calculated from (2). C: The data-driven interaction plot is shown again on the left, this time with unimodal and multimodal saliencies normalised by their geometric mean. On the right, the model resulting from the coefficients calculated from (3) is shown.

multiplicative term provides only a small, slightly negative contribution (α_3: -0.0395 ± 0.0821). This suggests that a linear combination of the unimodal saliencies makes a substantial contribution to the measured multimodal saliency.

What is the Nature of the Integration?

As explained in the Methods section, the measured saliency maps were used to create a data-driven model which captures the relationship between unimodal and multimodal saliencies. The complete V×A×AV space (100×100 V×A bins), which contains all saliency triplets, is shown in Figure 9A. The lack of data in the areas furthest from the origin prohibits a meaningful weighted regression analysis of the complete dataset, so we limit our analysis to a subset of the complete range of unimodal auditory and visual saliencies for which each bin contains at least 1000 estimates of p_{AV} (outlined in black in Figure 9A). This represents a reasonable compromise between including as much data as possible in the model and minimising the error of estimation of

DOI 10.16910/jemr.1.2.4

ISSN 1995-8692

multimodal saliency for each bin. The chosen range (auditory: [0, 0.85×10^{-6}]; visual: [0, 3×10^{-6}]) is bordered by a box in Figure 9A and includes approximately 90.8% of the available data. The data within the range of interest were binned again (into 10 × 30 bins), with the expected value of each bin used as an estimate of p_{AV} (data surface in Fig. 9B) and then subjected to the weighted regression analyses of Equations 2 and 3. The coefficients acquired according to weighted least squares solution of (2) were: $\beta_0 = 0.0000$, $\beta_1 = 0.5554$, $\beta_2 = 0.3270$. As in the multiple regression result, the unimodal visual saliency outweighs the auditory contribution to the multimodal saliency. In order to quantitatively evaluate the AV surface of this integration plot in terms of the remaining hypotheses (Linear Integration and Supra-linear Integration, corresponding to planar and expansively nonlinear surfaces, respectively), a model of the multimodal saliency (model surface in Fig. 9B) was constructed using the unimodal saliency maps and these coefficients. The goodness-of-fit of the model and the data, quantified by the coefficient of determination r^2, is 0.9772, indicating that this simple linear combination of unimodal saliencies provides a good model for the corresponding multimodal saliency.

In comparison, a weighted least squares solution of (3) provided the following coefficients: $\gamma_0 = 0.1032$, $\gamma_1 = 0.6432$, $\gamma_2 = 0.0846$, $\gamma_3 = 0.0434$, which were also used to construct a model (shown in Fig. 9C), yielding a slightly higher r^2 of 0.9814. Here the visual information again dominates, but the auditory contribution decreases substantially, and a slight effect of the multiplicative term ($p_{A\times V}$) is also seen. The intercept γ_0 is also noticeably higher. As mentioned in the Methods section, the terms in (3) were normalized by their respective geometric means in order to bring them into comparable range. Although the introduction of the multiplicative term was the main motivation for this normalization, it is also important to note that the ranges of the contributing unimodal saliencies are not equal, which is reflected in the aspect ratio of Fig. 9B. Normalising the visual and auditory saliencies changes their relative ranges (as can be seen in the data plot of Fig. 9C), which explains the decrease seen between β_2 and γ_2.

Comparing the r^2 for the two models, it appears that the linear combination of unimodal saliencies provides a good model for the measured multimodal saliencies. Not a great deal is gained by the inclusion of the multiplicative term, which leads to the conclusion that the second of our candidate hypotheses – a linear integration – provides a good explanation of the process underlying crossmodal integration for the control of eye movements.

Discussion

We investigated the integration of auditory and visual information during overt attention by measuring the eye movements of 32 subjects, who were asked to freely explore natural stimuli in visual, auditory and audiovisual conditions. Analysis of the recorded eye movements revealed a systematic shift of gaze towards the sound location of multimodal stimuli. Auditory information thus has an effect on the selection of candidate fixation points. Furthermore, the pattern of fixation was found to be similar for images shown in V and AV conditions, implying that visually salient regions remain salient under multimodal stimulation. Taken together, this suggests that simultaneously presented visual and auditory information integrate before selection of the next fixation point. Reconsidering the three alternative processes detailed in the Introduction, we can now reject the first (Maximum) hypothesis of competing modalities, in which the maximally salient information wins out in the selection of the next target of overt attention. Next, multiple regression analysis of the saliency maps grouped by stimulus suggested that a linear combination of unimodal saliencies could adequately explain the multimodal saliency (supporting the Linear Integration hypothesis). Finally, the interaction plot constructed from the data was found to fit extremely well with a linear model, and little was gained by including the multiplicative term corresponding to the Supra-linear Integration hypothesis. These results therefore support our second hypothesis of a linear integration of visual and auditory information in the bottom-up control of human eye movements.

The auditory stimuli had been processed to differ in spatial location using a generic non-individualised technique, and based on reports in the literature (e.g., Wenzel et al., 1993), it was expected that subjects would have difficulty in vertically discriminating the locations of these simulated sounds. However, the 4 alternative forced-choice task results revealed that overall, all subjects performed beyond chance level. This first attempt at simulating auditory sources differing in spatial

DOI 10.16910/jemr.1.2.4

ISSN 1995-8692

location was a success, and the spatialisation approach used could be further developed to cover more of the stimulus space, allowing a more systematic examination of horizontal and vertical effects of auditory stimulation. There is certainly room for improvement, for example regarding the more subtle, subjective aspects of auditory perception, such as feeling that the sound is truly situated outside the head and not just directionally presented through headphones. One factor that can help to make sounds more realistic is reverberation, which was included in the simulation process we used to synthesise our stimuli. Fixation selection is an unconscious process, and it is not obvious whether and how any improvement in subjective auditory aspects, such as externalisation, would affect visual behavior. In addition, there is a known trade-off between the directional accuracy of sound localisation and the authenticity of the artificial auditory experience (Shinn-Cunningham et al., 2005), which was not explored in any great depth here. Nonetheless, the results of the 4AFC show that subjects could discriminate between the spatial locations of different sound sources. Additionally, fixation was shifted towards the sound source, showing a behavioural effect of the spatialised sounds. For the present purposes, this suffices to show that the stimuli were indeed perceived to differ in spatial location.

Some further aspects of the auditory stimuli also warrant discussion. First, the sounds were created to be reasonably stationary over time, and were additionally fixed to a single spatial location for their duration. It can be argued that natural sounds are often transient in time and space, and that they serve only to redirect attention. However, the analysis used here is dependent on the empirical saliency maps created from the recorded visual behaviour of many subjects. The construction of our integration model thus depended on having as much data as possible, which led us to choose continuous, spatially static sounds in order to prolong any spatial bias in participants' overt behaviour and to thus sample more data. Examining the progression of this spatial bias over time, we found that the orienting effect of sound on eye movements was strongest between 250 and 2500 ms. Within this temporal interval of interest, there was no indication that processing early after stimulus onset is any different to later processing, but the exact time course of the effect was not investigated any further. With enough data, a future study could address this in detail.

Second, we deliberately aimed to avoid semantic congruency between visual and auditory stimuli. This could of course be done differently. However, introducing meaningful content to the auditory modality also introduces associated problems. On the one hand, a readily identifiable sound might provide a more complex spatial cue by virtue of the listener's world knowledge – birds usually sing from trees, footsteps usually emanate from the ground, etc. In order to create a realistic integration model, the auditory saliency map would have to reflect this, which would require more data. Alternatively, the semantic information contained in the sound could be used to create congruent and incongruent audiovisual stimuli. There are different cases that must be dealt with here. In congruent stimuli, the sound-object, for example an animal, might already be visible, i.e. already visually salient. If the sound is incongruent to the presented visual scene, the sound could become salient for reasons of incongruence making incongruence difficult to control. Furthermore, in cases where the sound-generating agent is not visible, this might elicit visual search, which cannot be compared to free-viewing conditions. It is indeed interesting to explore the role of semantic congruence, but there are pitfalls involved, and here we chose to avoid them.

As mentioned earlier, a previous study by our group used the same approach but with visual images of forest scenes and birdsong presented from loudspeakers at the left or right of the computer screen (Onat et al., 2007). In that study there was high semantic congruence between images and sounds, and the same result was found – eye movements during audiovisual stimulation were well explained by a linear integration of visual and auditory saliency maps. There were several further differences between the two experiments. First, concerning the stimuli used in both experiments – the study described here increased the dimensionality of the auditory space, and deliberately avoided semantic correspondance between images and sounds. The means of presenting the stimuli also differed. Here, sounds were created using simulation software and presented through earphones, while the images spanned more of subjects' visual field. Despite these differences, our findings confirm the results of the previous experiment and support the major finding here: that a linear integration of unimodal saliencies provides a good explanation of the process underlying crossmodal integration for the control of overt attention.

DOI 10.16910/jemr.1.2.4

ISSN 1995-8692

In terms of the shift in fixation density towards the sound location during multimodal stimulation, we additionally investigated the effect in the vertical direction. The horizontal effect of sounds played in audiovisual conditions was clear, while the vertical effect was much smaller and did not reach significance when considered over subjects. One reason for this may be in the possible ambiguity in perception of sound locations – the 4AFC task revealed that subjects made most errors in vertically distinguishing sound locations. Alternatively, the stronger spatial biasing by the horizontal auditory component may simply reflect an intrinsic property of human viewers to attend more to horizontal aspects of the visual space, reflecting the spatial layout of the human world. This is certainly an aspect of this study that warrants further investigation, possibly by means of a further increase in the auditory coverage of stimulus space.

Our ultimate goal is to understand attentional processes under fully natural conditions, and this study can be considered as a step in this direction. In terms of visual stimuli, this requires a move beyond static images, and the use of video stimuli introduces a new set of image features in the form of motion features. These have been investigated in a pilot eye-tracking experiment within our group (Açık et al., in preparation), and further research is underway. In addition, naturally behaving observers act within their environment; as well as eye movements, head and body movements are an important part of overt attention. Recent work by our group and others has also begun to address this issue (e.g. Schumann et al., in press). The paradigm used here is certainly amenable to more advanced stimulation techniques, as the required uni- and multi-modal saliencies are measured empirically from the participants' visual behaviour, which removes the need for a robust model of saliency computation from the visual, auditory, or audiovisual stimuli themselves.

Finally, the hypothesis best supported by the data is that a linear integration of unimodal saliencies is involved in the control of eye movements during free viewing of natural stimuli. Previous work in our group (Schumann et al., 2007) has also shown that within the visual modality, the pair-wise integration of the information provided by visual features is also well-modelled by a linear integration. Taken together, these results hint that there might be a general principle of linear integration underlying the combination of information in the brain. As pointed out earlier, neurons in the Superior Colliculus have been found to integrate unimodal stimuli in a linear fashion, but this depends on stimulus efficacy, with stimuli close to threshold levels inducing a supra-linear integration. Working with artificial stimuli, some researchers have found the same pattern of inverse effectiveness (Corneil et al., 2002) when auditory and visual stimuli are embedded in noisy backgrounds, but others using different artificial stimulation methods have failed to do so (e.g. Frens et al., 1995). Here we used natural auditory and visual stimuli and found that a linear combination of unimodal saliencies well explains the integration process underlying overt visual attention. It remains to be seen whether further factors, such as a manipulation of semantic congruency, can have an effect on the integrative processing scheme found here.

Acknowledgements

This research was part of the EU STREP project Perception on Purpose (POP).

Appendix

Table 1
A description of the auditory stimuli used in the experiment. The name of the original sound file is given, along with a short description of the sound content.

Filename	Description
Bag_1	Heavy plastic bag, crumpled in hand
Beads_11	Plastic beads, poured from cup into box
Beads_6	Hand run through plastic beads in box
Cake_2	Small straw broom, crushed in hand
Cups_3	Thermos flask lid in plastic cup, shaken
Eraser_4	Whiteboard eraser, moved along arm
Foil_1	Aluminium foil, scrunched up and opened
Jug_7	Pen rubbed on outside of small ceramic jug
Kitchen_3	Egg beater, hit with plastic pen
Marble_5	Glass marble, rolled inside cork pot stand
Plates_3	Heavy plastic picnic plates, moved in hands
Sandpaper_3	Sheet of sandpaper, waved around

DOI 10.16910/jemr.1.2.4

ISSN 1995-8692

Spoons_2	Several spoons, moved against each other
Spoons_6	Spoons, moved against each other
Tags_10	Plastic tags, scraped on cork surface
Tags_2	Plastic tags, rubbed together
Teabag_2	Plastic bag containing tea leaves, squeezed
Tin_6	Empty coffee tin, pushed along book cover

References

Açık, A., Bartel, A., & König, P. (in preparation). Motion cues in overt visual attention.

Algazi, R. V., Duda, R. O., Thompson, D. M., & Avendano, C. (2001). The CIPIC HRTF database. In Proceedings from WASSAP '01: *2001 IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*. 99-102.

Awh, E., Armstrong K. M., Moore T. (2006). Visual and oculomotor selection: Links, causes and implications for spatial attention. *Trends in Cognive Sciences, 10*(3), 124-30.

Bell, A.H, Meredith, A.M., Van Opstal, J.A., & Munoz, D.P. (2005). Crossmodal integration in the primate superior colliculus underlying the preparation and initiation of saccadic eye movements. *Journal of Neurophysiology, 93*(6), 3659-3673.

Blauert, J. (1997). An introduction to binaural technology. In R. Gilkey and T. R. Anderson (Eds.), *Binaural and Spatial Hearing in Real and Virtual Environments*. Mahwah, NJ: Lawrence Erlbaum.

Campbell, D. R., Palomäki, K. J., & Brown, G. (2005). A Matlab simulation of shoebox room acoustics for use in research and teaching. *Computing and Information Systems Journal*, ISSN 1352-9404, 9(3).

Corneil, B. D., Van Wanrooij, M., Munoz, D. P, & Van Opstal, A. J. (2002). Auditory-visual interactions subserving goal-directed saccades in a complex scene. *Journal of Neurophysiology, 88*(1), 438–454.

DeGelder, B., & Bertelson, P. (2003). Multisensory integration, perception and ecological validity. *Trends in Cognitive Sciences, 7*(10), 460-467.

Ernst, M.O., & Banks, M.S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature 415*(6870), 429-433.

Foxe, J.J., & Schroeder, C.E. (2005). The case for feedforward multisensory convergence in early cortical processing. *NeuroReport, 16*(5), 419-423.

Frens, M. A., Van Opstal, A. J., & Van der Willigen, R. F. (1995). Spatial and temporal factors determine auditory-visual interactions in human saccadic eye movements. *Perception and Psychophysics, 57*(6), 802–816.

Fasano, G., & Franceschini, A. (1987). A multidimensional version of the Kolmogorov-Smirnov test. *Monthly Notices of the Royal Astronomical Society, 225*, 155–170.

Hoffman J.E., Subramaniam B. (1995). The role of visual attention in saccadic eye movements. *Perception & Psychophysics, 57*(6), 787–795.

IEC. (1979). *Sound level meters, international standard IEC 651:1979*. Technical report, International Electrotechnical Commission, Geneva, Switzerland, 1979.

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience, 2*, 194-203.

Kayser, C., & Logothetis, N. (2007). Do early sensory cortices integrate cross-modal information? *Brain Structure and Function, 212*(2), 121–132.

Kayser, C., Petkov, C. I., Augath, M., & Logothetis, N. K. (2007). Functional imaging reveals visual modulation of specific fields in auditory cortex. *Journal of Neuroscience, 27*(8), 1824–1835.

Kayser, C., Petkov, C. I., Lippert, M., & Logothetis, N. K. (2005). Mechanisms for allocating auditory attention: An auditory saliency map. *Current Biology, 15*(21), 1943–1947.

Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology, 4*(4), 219–227.

Meredith, M. A., & Stein, B. E. (1986). Visual, auditory, and somatosensory convergence on cells in superior colliculus results in multisensory integration. *Journal of Neurophysiology, 56*(3), 640–662.

Miller, J. (1982). Divided attention: Evidence for coactivation with redundant signals. *Cognitive Psychology. 14*, 247–279.

DOI 10.16910/jemr.1.2.4

ISSN 1995-8692

Onat, S., Libertus, K., & König, P. (2007). Integrating audiovisual information for the control of overt attention. *Journal of Vision, 7*(10), 1–16.

Parkhurst, D., Law, K., & Niebur, E. (2002). Modelling the role of salience in the allocation of visual selective attention. *Vision Research, 42*(1), 107-123.

Press, W. H., Flannery, B. P., Teukolsky, S. A., & Vetterling, W. T. (1992). *Numerical recipes in C: The art of scientific computing*. Cambridge, UK: Cambridge University Press.

Schumann, F., Einhäuser, W., Vockeroth, J., Bartl, K., Schneider, E., & König, P. (in press). Salient features in gaze-aligned recordings of human visual input during free exploration of natural environments. *Journal of Vision.*

Schumann, F., Açık, A., Onat, S., & König, P. (2007). Integration of different features in guiding eye-movements. In Proceedings of the 7th Meeting of the German Neuroscience Society / 31th Göttingen Neurobiology Conference, Neuroforum 2007, Göttingen, Germany.

Shinn-Cunningham, B.G., Lin, I-F, & Streeter, T. (2005). Trading directional accuracy for realism. In *Human-Computer Interaction International 2005 / 1st International Conference on Virtual Reality*.

Sparks, D.L. (1986). Translation of sensory signals into commands for control of saccadic eye movements: role of primate superior colliculus. *Physiol Rev 66*, 118-171.

Stanford, T. R., Quessy, S., & Stein, B. E. (2005). Evaluating the operations underlying multisensory integration in the cat superior colliculus. *Journal of Neuroscience, 25*(28), 6499–6508.

Wallace, M. T., Ramachandran, R., & Stein, B. E. (2004). A revised view of sensory cortical parcellation. *PNAS, 101*(7), 2167–2172.

Wenzel, E. M., Arruda, M., Kistler, D. J., & Wightman, F. L. (1993). Localization using nonindividualized head-related transfer functions. *Journal of the Acoustical Society of America, 94*, 111–123.

DOI 10.16910/jemr.1.2.4

ISSN 1995-8692