The Integration of Auditory and Visual Motion Signals: Neural Summation vs. Independent Decisions

Georg F. Meyer and Sophie M. Wuerger
Centre for Cognitive Neuroscience
School of Psychology
University of Liverpool, UK
georg@liverpool.ac.uk, sophiew@liverpool.ac.uk

Abstract
It is well known that stationary signals are integrated across modalities. In this paper we review three sets of experiments that show that motion perception also benefits from signals in multiple modalities. In the fist experiment we show that supra-threshold auditory motion induces a consistent response bias in a visual motion detection task. The second experiment shows that the integration of global motion signals is well described by a probability summation model. The integration of cross-modal local motion signals that are matched in position and speed, on the other hand, is predicted by a neural summation model (experiment 3). We conclude that cross-modal signals have to be co-localised for effective low-level motion integration.

1. Introduction
Objects or events that we perceive are usually defined by correlated information in multiple sensory modalities. This correlation is exploited by the perceptual system to direct attention to objects or events that share common cues. Stein and colleagues [1] showed that neurones in the superior colliculus respond to stimulation from multiple modalities and that the receptive fields of these neurones typically are approximately spatially aligned in all modalities. Furthermore, the response characteristics of multi-modal neurones show non-linear interactions between the different modalities, particularly near the threshold of the respective unimodal signals. Spatio-temporally coherent stimulation leads to a response enhancement, whereas simultaneous but spatially incoherent stimulation can lead to a response depression.

The neurophysiological data are mirrored in behavioural studies, which show numerous cross-modal links in endogenous and exogenous spatial attention for coincident [2,3] and sequential [4] stimuli in the auditory and visual modalities. The inferior and superior colliculus, which have been implicated in the processing of multi-modal spatio-temporal cues, have also been shown to contain neurones that respond selectively to motion signals [5]. The behavioural effects seen for co-localised signals should therefore also be seen for motion stimuli if the neural processing if this information is analogous. The aim of this study was to establish further evidence for the perceptual integration of auditory and visual motion signals. In a series of experiments we explore the mechanisms of audio-visual motion integration. We show evidence for bias effects as well as for two independent motion integration mechanisms in human observers.

2. Audio-visual Motion Capture
The first experiment explores whether a supra-threshold auditory motion signal introduces a response bias in a visual motion direction identification task.

2.1. Methods
The visual stimulus consisted of a random dot kinematogram (RDK), a black square (16 degrees visual angle) in which 500 white dots (0.05 degrees visual angle) move from a random start position along a line trajectory in a random direction with a speed of 7.4 deg/sec. The proportion of dots that move in a coherent direction can be varied to yield the percept of directed global motion. Figure 1a shows examples where there is no coherent motion (left), where 50% of the dots move coherently (center) and where all dots move in a coherent direction (right). In this and all other experiments described in this paper, the relative proportion of motion signals moving coherently was systematically varied to measure the threshold at which motion is detected. The auditory motion signals were created by cross-fading a white noise of 1 sec duration between two loudspeakers as shown in figure 1b.

The task of the observer in experiment 1 was to identify the direction of visual motion in the RDK. Auditory and visual motion could either be matched or mismatched in location (Fig. 1c, condition A compared to condition B and C) and could move at same or different speeds (Fig. 1c, compare condition A with D).
3. Global Motion Signals: Probability Summation

The aim of our second experiment was to assess whether we find an increase in sensitivity rather than a response bias when motion signals in both modalities are available. We also test whether audio-visual motion integration is restricted to ecologically valid motion signals.

3.1. Methods

We presented two intervals, one containing noise only (always audio-visual noise), the other containing the motion signals (auditory, visual, or bimodal). The 10 subjects had to indicate which of two intervals contained motion in either modality. Auditory and visual motion could either move in the same directions (consistent) or in opposite direction (inconsistent). The auditory and visual motion signals were the same as in experiment 1, except that auditory motion coherence was modulated by adding a pedestal of white noise to the cross-faded auditory signal component.

2.2. Results

Figure 2 shows the raw data and the fitted psychometric functions for 10 observers for the matched speed and position condition (Condition A). When auditory motion but no visual motion is present (fig. 2a, c; zero visual motion coherence), subjects report in 60% of the cases that the visual motion is moving in the direction of the auditory signal. If auditory motion had no effect on the perceived visual motion, we would expect the lower asymptote of the psychometric functions at 50% (chance performance). This is the case for control condition (Fig. 2b).

The data demonstrates that the visual motion bias induced by supra-threshold auditory motion is consistent with the direction of the auditory motion (audio-visual motion capture). We found the same results when auditory and visual motion were not matched in location or speed (Fig. 1, conditions B,C,D). See [6] for more details.

Figure 3: Relative frequency of correct identification for auditory, visual and bimodal stimuli for the consistent and inconsistent motion condition.

3.2. Results

Figure 3 shows the motion discrimination performance as a function of coherence in the visual and auditory signal domain. Two-dimensional psychometric func-
tions were fitted to the auditory, the visual and all the intermediate data points and thresholds (81% correct identification) for auditory and visual motion identification were derived (see fig. 3). Neural summation predicts that the threshold contour form a straight line, probability summation predicts curved lines (see also Fig. 4). A goodness-of-fit test revealed that the probability summation model accounts for the data quite well whereas neural summation must be rejected which is consistent with previous findings [7]. The data for inconsistent and consistent motion do not differ from each other, which implies that this late integration is not direction selective [8].

4. Localised Motion Signals: Neural Summation

Behavioral data derived from global motion signals are consistent with a probability summation model, which assumes separate processing in the two modalities and integration at a decision level. This is consistent with the finding that the motion signals in both domains do not need to be matched in position, speed or direction to be integrated. This finding appears to be inconsistent with physiological data [9, 10] that suggest that stationary multi-modal signals are integrated at an early neural processing stage provided the signals match in position and are coincident. The aim of this experiment was to test whether strictly co-localised signals are required for neural summation.

4.1. Methods

We presented two intervals, one containing only noise (auditory-visual dynamic noise), the other containing a motion signal plus noise. The motion signal could be visual, auditory or binodal. Ten subjects were asked to judge which of two intervals contained motion by pressing the appropriate button on a response box. Local motion signals were generated by using 31 LEDs and loudspeakers that were arranged along a 180 degree arc. The motion signals described a 90 deg arc in either the left or right frontal hemifield of the observer. Auditory and visual components moved independently; hence the signals could move in the same (D+) or opposite direction (D-) and could be in the same (H+) or in different hemifields (H-). A schematic diagram is given in fig. 5. The signals always moved at a speed of 30deg/sec. Signals were presented in a background of noise, fig. 6. Interleaved Quest procedures [11] were used to measure 24 thresholds.

4.2. Results

We did not find significant differences in the thresholds as a function of the hemi-field of signal presentation or for the direction of motion. The data shown in fig. 7 are therefore pooled across these conditions. To compensate for variation in the auditory and visual thresholds between subjects the data is normalized to the individual audio and visual thresholds.

The different behavioural thresholds for the consistent (H+D+) and inconsistent conditions suggest that two integration processes are used. A neural summation process accounts for the consistent motion signal thresholds while a probability summation process predicts the thresholds seen in the other conditions [12].
4.3. Audio-Visual Receptive Fields

A neural summation model assumes that raw motion signals are integrated across the two modalities within a narrow spatial kernel. We measured this receptive field size by systematically displacing audio and visual motion signals that moved in the same direction and at the same speed. The experimental conditions were the same as in the previous experiment. Threshold data for five subjects is shown in fig. 8. The data shows that the motion signals in both modalities have to fall within 20 degrees to be effectively integrated.

5. Conclusions

Our experiments show that motion detection thresholds for audio-visual signals are significantly lower for locally consistent motion signals than for global motion signals or for inconsistent signals. This means that effective integration of audio-visual signal will require either real sound sources and visual objects or very high quality simulations. This is likely to prove either a fascinating challenge or a real headache for developers of systems that exploit the much higher sensitivity to multi-modal data in context such as virtual or augmented reality systems.

6. Acknowledgements

This work was supported by the EU TMR projects Spear and Hoarse as well as by the Royal Society. Most of this work was carried out in collaboration with F. Roehrbein, M. Hofbauer, K. Schill, and C. Zetzsche, Munich / Bremen University.

7. References