Cross-modal integration of auditory and visual apparent motion signals: not a robust process

D.Z. van Paesschen

supervised by:
M.J. van der Smagt
M.H. Lamers

Media Technology MSc program
Leiden Institute of Advanced Computer Science (LIACS)
Leiden University, The Netherlands
david.vanpaesschen@gmail.com
September 1, 2009

Abstract

Cross-modal perceptual research is a field of growing interest, after decades of primarily unimodal studies. The existence of multisensory neurons is well documented and previous cognitive studies have shown the effect of one sensory modality on another. Studies on cross-modal integration of auditory and visual motion signals have focused on facilitation of motion detection. The present study uses auditory and visual apparent motion signals to test the effect of sound on the perception of visual motion. In one experiment we show no effect of sound that is temporally coincident and spatially correlated to the visual stimulus, a motion quartet. The second experiment shows no effect when the onset of the auditory stimulus is 118 ms before onset of the frames of the motion quartet. In a third experiment we show no priming effect when the auditory stimulus is presented completely before onset of the motion quartet. Our findings in the first two experiments corroborate results from previous studies on audio-visual integration. It is argued that signals have to be spatially coincident for cross-modal integration to occur, whereas in our experiments the signals were merely spatially correlated.

1 Introduction

For efficient interaction with our environment the information from our different senses has to be integrated. When playing piano for example, the visual image of ones fingers on the keys together with the information from the haptic sensory system when touching or pressing the keys and the sound the piano produces have to be combined to one unified perceptual experience. Not only does information from our different senses converge to a unified representation
of the external world, input from one sensory modality can influence the information gained by the other. Different sensory modalities can interact and alter the processing of one another.

A clear example of visual influence on auditory perception is the ventriloquism effect (Howard and Templeton, 1966) in which an auditory stimulus is misperceived as coming from a direction other than the true direction, due to the influence of a visual stimulus. The influence of visual perception on the auditory one can also be observed in the well-known McGurk effect (McGurk and MacDonald, 1976) in which a single, spoken syllable is misidentified by the listener when watching a human speaker pronounce a conflicting syllable. In the classic experiment the syllable [ba] had been dubbed on to the lip movements for [ga], resulting in the auditory percept of the syllable [da].

An example of the effect of sound on visual perception was given by Shipley (1964) and in the later ‘sound induced illusory flash’ (Shams et al., 2000). Changes in the physical flutter rate of an auditory stimulus were found to induce changes in the apparent flicker rate of a visual stimulus. Another clear example of the effect of sound on visual perception is been given by R. Sekuler, A.B. Sekuler and Lau (1997). Two identical objects moving towards one another, coinciding, and then moving apart can be perceived in two different ways. After coincidence one may see the objects continuing in their original directions or perceive them as having bounced off and moving in reversed direction. The presence of a brief sound at or near the point of coincidence promotes the perception of bouncing. Although the latter experiment shows an effect caused by an auditory signal, both visual and tactile transients are likewise capable of having the bounce-inducing effect (Watanabe and Shimojo, 1998; Watanabe, 2001).

Motion perception is of great importance for interaction with our environment and hence an interesting topic within the field of cross-modal integration. Motion signals can be seen as more complex stimuli than static or transient signals, as they transfer in space over time. For this reason, one could expect that integration of motion signals relies at least partly on other aspects than integration of static or transient signals does. Studies on cross-modal integration of auditory and visual motion signals have focused on facilitation of motion detection (Alais and Burr, 2003; Meyer and Wuerger, 2001; Meyer et al., 2005). Our research builds upon the work of Sekuler et al. (1997) as we focus on the effect of sound on visual motion perception in a bistable image. Instead of using a transient signal, we presented sound that appears to be moving. Depending on spatial and temporal aspects, how does the sound affect the visual motion perception of the bistable image?

### 2 Synchronous grouping

As in the experiments by Sekuler et al. (1997) we used a bistable visual stimulus, a visual image that can be perceived in two mutually exclusive ways. The visual stimulus we used in our experiments is a motion quartet (Ramachandran and Anstis, 1985; Hock et al., 1993), which is a modified version of the apparent motion stimuli by Ternus (1926). A motion quartet is formed by simultaneously presenting two dots corresponding to the diagonally opposite corners of an imaginary rectangle, then presenting two dots on the other corners, then the
first pair again, then the second and so on. The mentioned imaginary rectangle of a classic motion quartet consists of a vertical and a horizontal pair of opposite sides. In most cases, observation of this motion quartet results in a continuously alternating stable movement perception: the two dots are perceived to move either horizontally in opposite directions or vertically (figure 1).

![Figure 1: Example of a classic motion quartet. The two dots are perceived to move either horizontally in opposite directions (left quartet) or vertically (right quartet).](image)

Hock et al. (1993) showed that the aspect ratio of the motion quartet (the vertical divided by the horizontal distance between the dot positions) determines the relative stability of the perceived motion pattern. When the aspect ratio favours the perception of one of the motion patterns, the percept of the favoured motion pattern is more stable than the percept of the unfavoured pattern. The motion quartet has the lowest stability when the aspect ratio is close to 1. Spontaneous switches between the two patterns occur more often for aspect ratios near 1. When the horizontal distance between the dot positions is small, relatively to the vertical distance, the perception of horizontal motion is favoured. When the vertical distance is relatively small, the perception of vertical motion is favoured. This can be seen as a nearest neighbour principle for apparent motion (Shechter et al., 1988; Hock et al., 2003).

Although the aspect ratio that gives equal durations of vertical and horizontal motion perception (i.e. the motion quartet is optimally bistable) is 1 according to the nearest neighbour principle, it is shown that this value can vary between observers (Sterzer and Kleinschmidt, 2005; Kohler et al., 2008). In a first pilot experiment with three subjects, we wanted to establish for which aspect ratios horizontal motion was favoured and for which ratios vertical motion. This would give us a range of aspect ratios to use in our experiments. The results showed a huge bias for seeing vertical motion, if a ratio of 1 is taken as the theoretical point of optimal bistability. The average aspect ratio for optimal bistable apparent motion for our subjects was around 1.7. In order to evoke a theoretically less biased visual perception of the motion patterns we then used a 45° rotated version of the quartet in our experiments. Experiments with this quartet showed clearly smaller biases. The average aspect ratio for optimal bistability was now close to 1. The two motion patterns that were perceived are shown in figure 2. The pattern shown at the top we call the ‘45° motion pattern’, referring to the angle of the line along which the motion is perceived. For the same reason we call the pattern shown at the bottom the ‘315° motion pattern’.
Figure 2: The two black squares in the motion quartet are perceived as moving either in as the upper pattern (45° motion pattern) or as in the lower pattern (315° motion pattern).

2.1 Methods

Apparatus

A monitor screen (Vision Master Pro 454) was placed at a distance of 57 cm from a chin rest. One audio speaker was placed in the centre in front of the monitor, not obstructing the screen. The other speakers were placed on top of the monitor using a shelf in a way that they both made an angle of 45° with the bottom speaker (figure 3). The distance between the centres of the two top speakers was 100 cm. Stimuli were generated on a Macintosh computer running Matlab using the Psychophysics Toolbox extensions (Brainard, 1997; Pelli, 1997).

Stimuli

The motion quartets were presented as black squares on a grey background (22.3 cd/m²). The size of each square was 9 pixels, corresponding to a visual angle of 0.32°. The aspect ratios were 1:2, 2:5, 5:12, 12:25, 1:6, 6:17, 3:7 and 7:21. The angular distance between the centres of the squares on the longest side of the quartet was always 1.3°. In the centre of the screen a red fixation point (6.50 cd/m², 4 pixels, 0.14° visual angle) was present during the whole trial. The distance between the centre of the fixation point and the centre of the lowest possible square was 0.63°.

Motion quartets were shown for two cycles, one cycle being a two-frame presentation of both pairs of squares. The duration of one frame was 200 ms.
Figure 3: Experimental setup of monitor screen and audio speakers. The top speakers make an angle of 45° with the bottom speaker.

Between frames there was an interstimulus interval of 47 ms.

The auditory stimulus consisted of two (simultaneously presented) sine waves. The sine waves with a frequency of 220 Hz and a duration of 247 ms were created using the Audacity software. The Nyquist Panning plug-in was used to create a panning effect. When simultaneously played over two speakers, the waves combined to a sound that appeared to be moving between the two speakers (70±5 dB). The onset of the sound coincided with the onset of the pair of squares. The two waves were played through the bottom and top right speaker (‘45° sound’) or through the bottom and top left speaker (‘315° sound’). The wave with decreasing amplitude always started in the bottom speaker. In all of our experiments there is a ‘no sound’ condition, giving us the framework to which we compare the results of the sound conditions.

A mask, consisting of five frames, was presented between trials. The first frame showed two squares, one on the top and one on the left corner of an imaginary rectangle. In the second frame the two squares were shown on the bottom and right corner. In the third frame the squares were shown on the upper and right corner and the fourth frame showed the squares on the bottom and left corner. The final frame presented four squares, one on every corner. The visual mask was shown to get apparent movement on both diagonal lines before the next trial was presented. Frame duration and interstimulus interval were equal to the ones used for the motion quartets. The shape and size of the imaginary rectangle and squares were the same as those of the preceding motion quartet.

Simultaneously with the visual mask an auditory mask was presented. During the first four frames the sine wave with the increasing amplitude was played over the central speaker. The fifth frame this wave was played simultaneously over all three speakers to evoke an auditory perception of a sound in the centre between the speakers. The auditory mask was presented to get a centred auditory focus before the next trial started.

Subjects

Four subjects participated in this experiment, three of which were students at Utrecht University and which were naïve with respect to the purpose of the
experiment. The author was the fourth subject. All subjects had normal or corrected-to-normal hearing and vision.

**Procedure**

Subjects were instructed to fixate the red square in the centre of a grey screen (holding for 2000 ms) after which a motion quartet (with or without one of the two sounds) was presented for 941 ms (see figure 4). When the motion quartet appeared, the subjects had to focus attention on the square that appeared right above the fixation point. After the motion quartet disappeared, the subjects had to press one of two keys on the computer keyboard to indicate whether the focused square moved right up (45° motion pattern) or left up (315° motion pattern). Preceding the next motion quartet, a mask was shown for 1188 ms followed by the fixation point without stimulus. All combinations of aspect ratio and sound (45° sound, 315° sound, no sound) were presented equally often, in randomized order in blocks of 135 trials. There were four blocks.

![Figure 4](image)

Figure 4: Schematic of a trial (without mask) over time in the synchronous grouping experiment. In the upper part the visual frames are shown (fixation point [2000 ms], motion quartet [200 ms] and interstimulus interval [47 ms]). In the lower part an approximation of the relative volume levels of the top (T, left or right) and bottom (B) speaker are shown.

**2.2 Results**

Figure 5 shows the results for each subject. The aspect ratio is the length of the 315° side of the imaginary rectangle divided by the length of the 45° side. Clearly, for larger ratios the 45° motion pattern is perceived more often as the distance between the squares on this line is relatively small for these ratios. For three out of four subjects the point of 50% for the ‘no sound’ condition appears between the aspect ratios of \(\frac{5}{6}\) and \(\frac{1}{4}\). Using Matlab we plotted the data and fitted a normal cumulative distribution function for each sound condition (see figure 6), supposing this function describes the correlation between the variables the most accurate. With a 95% confidence interval no significant differences were found between the three functions. This means no effect of sound on the number of 45° motion pattern percepts was found.
3 Asynchronous grouping

Studies on the mentioned bounce-inducing effect (Watanabe, 2001; Watanabe and Shimojo, 2001; Shimojo and Shams, 2001) suggest there is a time window in which the transient signal shows the effect. The effect can be obtained within a certain range of image-sound asynchrony. Visual, tactile and auditory transient signals can have a slightly increased effect when presented right before (with an interval of around 100 ms) the point of coincidence (Shimojo and Shams, 2001). In our second experiment we presented the auditory cue 118 ms before onset of the pair of squares.

3.1 Methods

Apparatus, stimuli and procedure were almost the same as in the synchronous grouping experiment. The only difference was that the interstimulus interval between frames was increased to 129 ms to make an onset asynchrony possible. The onset of the auditory stimulus was now 118 ms before the onset of the pair of squares. All four subjects that participated in the first experiment took part in this one.

3.2 Results

Figure 7 shows the results of the second experiment for each subject. The larger interstimulus interval did not prevent the subjects from having a clear motion perception as gained in the first experiment. Using the same analysing methods as in the synchronous grouping experiment, we found no significant difference between the number of 45° motion pattern percepts in the no sound conditions
Figure 6: Data plot of the synchronous grouping experiment. A normal cumulative distribution function is fitted for each sound condition.

of both grouping experiments. Moreover, we found no effect of sound on the number of 45° motion pattern percepts, despite the small asynchrony (see figure 8).

4 Priming

As the visual motion perception was not affected by sound in a condition in which the visual and auditory stimulus were presented to be grouped, we took a different approach to test whether the auditory signal provides clues for the visual motion perception.

Motion priming can be described as biasing the perception of a certain motion by preceding the ambiguous motion stimulus with a ‘prime stimulus’. Evidence for priming of visual motion perception was given by Ishimura and Shimojo (1994), Ishimura (1995) and Wohlschläger (2000). They showed that the perceived motion direction of an ambiguous motion stimulus was biased by hand movements of the observer. In our research we used the auditory stimulus as a prime.

4.1 Methods

Apparatus and stimuli were the same as in the grouping experiments. In the priming experiment the auditory stimulus was presented before the motion quartet was shown (figure 9). The first pair of squares appeared right after the last sound had stopped. Only motion quartets of ratio $\frac{5}{6}$, $\frac{1}{1}$ and $\frac{5}{5}$ were used in this experiment.

Subjects

Four subjects participated in this experiment, three of which were students at Utrecht University and which were naïve with respect to the purpose of the experiment. None of them took part in the grouping experiments. The author
Figure 7: Results from the asynchronous grouping experiment. Sound is presented 118 ms before onset of the pair of squares.

was the fourth subject. All subjects had normal or corrected-to-normal hearing and vision.

**Procedure**

Subjects were instructed to fixate the red square in the centre of a grey screen (holding for 2000 ms) after which the auditory stimulus was presented for 988 ms (the auditory stimulus could also be ‘no sound’). Right after the last sound stopped the motion quartet was presented for 941 ms. When the motion quartet appeared the subjects had to focus attention on the square that appeared right above the fixation point. After the motion quartet disappeared the subjects had to press one of two keys on the computer keyboard to indicate whether the focused square moved right up (45° motion pattern) or left up (315° motion pattern). Preceding the next motion quartet, a mask was shown for 1188 ms followed by the fixation point without stimulus. All combinations of the three aspect ratios and sound (45° sound, 315° sound, no sound) were presented equally often, in randomized order in blocks of 84 trials. There were three blocks.

**4.2 Results**

The results for each subject are shown in figure 10. As in the priming experiment only ratios of $\frac{5}{6}$, $\frac{1}{3}$ and $\frac{6}{5}$ were used, no sigmoid curve could be fitted on the data. We analysed the data using GLM repeated measures. No effect of sound on the number of 45° motion pattern percepts was found ($F_{[2,6]} = 1.624$, $p = .273$).
5 Discussion

Although Sekuler et al. (1997) showed a convincing effect of sound on the perception of a bistable visual motion stimulus in a grouping experiment, our grouping experiments do not show any effect of sound on the visual perception of the motion quartet. From what our subjects reported, the motion patterns in the quartets were easily perceived and the sounds were clearly distinguishable from one another and were perceived as ‘sort of moving’. We presented the sounds exactly at the onset of the pair of squares in one experiment and created a little asynchrony (118 ms) in another. Based on our findings and previous studies there are several possible explanations for why the effect failed to appear.

The mentioned bounce-inducing effect has an intuitively clear ecological explanation: most collision events in the natural environment yield synchronized cross-modal signals (Shimojo and Shams, 2001). In most cases when two objects collide, the collision produces perceivable sound. Although it seems intuitively plausible to accompany the motion quartet with sounds that seem to be moving under the same angle and in the same direction as one of the possible percepts, a possible effect would lack a clear ecological explanation. Moving objects don’t...
Figure 10: Results from the priming experiment.

have to produce perceivable sound. An idea could be to use only transient sounds in between frames of squares (during the interstimulus interval). Instead of presenting two sine waves that sum up to an apparent moving sound, short signals are played over the bottom and one of the top speakers. The squares in the motion quartet can be perceived as bouncing back and forth as the transient signals can have the bounce-inducing effect. The problem, however, is that these sounds combined with the motion quartet create a multi-modal stimulus that is still ambiguous. The sounds could be perceived as created by the bouncing of one square moving on one side of the quartet or as created by two squares bouncing on one side of the quartet. The two different interpretations of the auditory stimulus correspond to the two different motion percepts, hence the ambiguity.

A second possible explanation for why no effect occurred in the grouping experiments is the use of auditory and visual motion illusions. The stimuli in our experiments were not moving physical sources. Meyer et al. (2005) claimed that for facilitation of motion detection to occur high quality auditory spatial signals are crucial. In an experiment by Alais and Burr (2004) no facilitation was shown when they presented a discrete moving visual stimulus together with auditory motion signals created by varying inter-aural time differences among other things. Our auditory stimulus was simply created by cross-fading two sine waves. Although these methods provide an illusion of motion, the created
signals do not contain all cues of a moving sound source that are normally available to the listener (e.g. shoulder echo and pinna response, inter-aural time and level difference). Although the cross-fading had to evoke the perception of a moving sound source, it is possible to perceive the auditory stimulus the way it is presented, as two sound sources changing in amplitude.

Thirdly, there is an evident difference between the apparent motion of the visual and the auditory stimulus we used in our experiments. The motion quartet is created by squares flashing on and off on four clearly separated locations. The perception of visual motion in this quartet is solely build upon these discrete signals. The apparent auditory motion is created by a smooth cross-fading of audio signals. For cross-modal interaction to occur in this case, a discrete visual signal has to be grouped with a continuous auditory signal. This could be impossible, although there is a perception of visual continuity.

Another hypothesis for why no effect occurred in the grouping experiment comes from neurophysiological studies. Studies by Meredith and Stein (1983, 1985) and Stein et al. (1988) with cats have documented the existence of multisensory neurons in deep layers of the superior colliculus. These neurons respond to stimuli in different modalities, for example, vision and audition. The integration of multi-modal inputs by the multisensory neurons depends on both the spatial and temporal proximity of the stimuli. Only temporally and spatially coincident stimuli from different modalities are integrated. Stimuli that are temporally or spatially disparate produce less or no interaction at the single cell level. These neurophysiological findings are supported by several behavioural studies (e.g. Stein et al., 1988; Meyer et al., 2005; Bolognini et al., 2005). In our grouping experiments the visual and auditory stimuli were temporally coincident or within a certain time window, but spatially the stimuli were disparate. We aligned the direction of apparent movement by the visual and auditory stimulus, but the visual angle between the visual stimulus and the centre of both top speakers was approximately 56°. Moreover, the distance between the bottom speaker and each of the top speakers was 71°, at least 54 times larger than the distance over which motion of the squares was perceived. Possibly, due to these large spatial discrepancies the stimuli could not produce interaction at the level of multisensory neurons and were processed as separate events. Also, the accuracy with which human listeners can localize sound, i.e. the minimum audible angle (at least 3° for 100% accuracy; Hartmann and Rakerd, 1989), does not correspond with the visual angle on which the motion quartet is presented (0.7-1.3°). Enlarging the motion quartet, on the other hand, makes the motion patterns harder or even impossible to perceive.

Also the priming experiment shows no effect of sound on the number of 45° motion pattern percepts. Observing the test results per subject, it appears that the results of the author (subject D) indicate the possibility of an effect. In the grouping experiments no essential difference is found between the results of the author and the other subjects. One reason for the author’s results in the priming experiment could be that as an ‘experienced observer’ of motion quartets it is more easy to perceive the less favoured motion pattern if primed (or decided) in front. Kolher et al. (2008) reported that the observers’ ability to influence their movement percept of motion quartets was substantial. Although in our priming experiment the task was not to have intentions on which motion pattern to perceive, the prime could have a bigger effect for ‘experienced observers’.
If a priming effect could be found (for ‘experienced observers’) we could say that however the signals (when temporally coincident) are not grouped as coming from the same source, the auditory signal does provide one or more clues for the visual motion perception.

Cross-modal integration of auditory and visual apparent motion signals does not seem to be a very robust process. Our perceptual system might not be fooled easily if a combination of signals does not correspond ecologically. Signal structure as well as temporal and spatial coincidence seem to be decisive in whether two signals are processed as coming from one and the same source.

References


