Filling-in visual motion with sounds

Aleksander Väljamäe¹ and Salvador Soto-Faraco²

¹ Division of Applied Acoustics, Chalmers University of Technology, Gothenburg, Sweden
²ICREA & Parc Científic de Barcelona, Barcelona, Spain
{aleksandervaljamae@chalmers.se

Abstract

Information about the motion of objects is extracted by multiple sensory modalities and integrated during perception. Often the flow of such information is rather discontinuous, (e.g. seeing and hearing a cat quickly moving through the furniture in a cluttered room). The present study addressed audio-visual interactions in the perception of time-sampled object motion in depth. Consistent with the results of previous studies using continuous motion, we found significant auditory motion aftereffects both within and across modalities, that is, following adaptation to discrete (sampled at 18.8 Hz) auditory and visual motion. The visually-induced (crossmodal) auditory motion aftereffect was eliminated when adapting to visual motion stimuli flashing at half the rate (9.4 Hz). Remarkably, the addition of the high-rate acoustic flutter to this sparsely time-sampled visual motion restored the auditory aftereffect to a strength comparable to what it seen with high-rate bimodal adaptors (flashes and beeps). The results suggest that the observed aftereffect resulted from the occurrence of sound-induced illusory flashes. The occurred auditory filling-in of time-sampled visual motion supports feasibility of reduced frame rate in multisensory broadcasting and virtual reality applications.

1. Introduction

Creating the illusion of reality has been a driving force in the development of new media forms. Current Virtual Reality (VR) research often stresses the importance of perceptual optimization of mediating technologies, that is, finding the minimal set of external cues that can lead to an illusion of presence in a virtual environment (e.g. Sanchez-Vives & Slater, 2005, and references therein). This challenging technological task is impossible without a good deal of knowledge about the mechanisms of human perception and cognition. Indeed, perceptual processes taking place in our brain often implement a filling-in strategy whereby the illusion of sensory information helps to complete missing external cues, e.g. the shape of visual objects that are partly occluded from view (see Komatsu, 2006 for a recent review). However, despite the widely accepted fact that human perception is multisensory in its nature and the importance that cross-modal interaction studies have for new media applications, there is barely any systematic research addressing the role of multisensory interactions in the process of perceptual filling-in.
In traditional media (i.e., cinema and television) sound has been successfully used for creating an illusion of visual action continuity. In classic example from the George Lucas film “The Empire Strikes Back” (1980), an opening of an automatic spaceship door consists of two successive stills (a door closed and a door opened) plus a “swapping” sound effect, altogether creating a visual illusion of the door movement (Chion, 1994). Recent studies focusing on cross-modal interactions confirm that sound can induce visual illusions. For example, Shams et al. 2000 (see also, Shams et al., 2002) reported an illusion of multiple visual flashes that was produced when one brief visual stimulus was coupled with multiple auditory beeps. In these experiments participants were asked to count the number of times a flickering white disk had flashed while presented with one or more task-irrelevant short sounds. The number of flashes reported by observers increased with the number of beeps. In a later study, measurements of visually-evoked potentials revealed a modulation of visual cortex activity that was time locked to the presentation of acoustics beeps (Shams et al., 2001). Interestingly, the electrophysiological activity corresponding to the illusory flashes was found to be very similar to the activity produced when an actual flash was physically presented.

Given that motion perception has been repeatedly shown to be subject to strong multisensory interactions (e.g., Meyer & Wuerger, 2001; Kitagawa & Ichihara, 2002; Soto-Faraco et al., 2002, 2004; See Soto-Faraco et al., 2003 for a review), we addressed the effects of sound-induced visual flashes when perceiving time-sampled object motion. Such discontinuous motion is represented by approaching or receding stimuli which flash (or beep) on and off. We chose this test because adaptation to a continuous object motion in depth is known to produce consistent motion aftereffects (MAE) both unimodally as well as cross-modally (Kitagawa & Ichihara, 2002). For example, after exposure to a looming visual object the viewer will perceive a steady visual stimulus as if it was receding. In addition, as was shown by Kitagawa and Ichihara, visual or audio-visual continuous object motion in depth can also result in an auditory aftereffect consisting of changing-loudness (which indicates acoustic movement in depth).

The present experiment closely followed the methodology used by (Kitagawa & Ichihara, 2002) where auditory changing-loudness MAE was measured after the adaptation to motion in depth stimuli. In particular, we addressed the potential interactions between time-sampled visual and acoustic motion in depth by measuring its potential to produce auditory aftereffects. Although critical flicker-fusion frequency is typically cited to be around 60 Hz (Brown, 1965), the fusion of individual flashes into a perceptually continuous event might occur at a lower stimulus repetition rates of approximately 20 Hz (Jewett et al., 2006 and references therein). Studies in VR show drastic decrease in task performance and subjective experience of presence for the frame rates below 15 Hz (Barfield & Hendrix, 1995; Reddy, 1997).

Our first hypothesis was that MAE (visual and acoustic) will depend on the frequency of the flashes representing discrete motion of the adapting visual object. Sparsely time-sampled visual motion should produce lower MAE compared to higher rate (i.e., perceptually fused) moving stimuli. Our second hypothesis was that combination of a
slow train of flashes with a rapid train of beeps (flutter) might lead to the sound-induced illusory flash illusion (Shams et al., 2002) which would help fill-in visual object motion. If this is the case, such audio-visual combination should produce MAE comparable to MAE induced by a real stimulus containing synchronized flashes and beeps at a high rate.

2 Method

Participants

Fifteen PhD and undergraduate students at the University of Barcelona (mean age of 28.5 years, 10 females) took part in the experiment voluntarily. All reported to have normal or corrected to normal vision and no hearing problems. They were paid 1 cinema ticket for their participation. The current study has been conducted under approval of the local ethics committee at the University of Barcelona.

Stimuli

The auditory and visual adaptation stimuli were synthesized and rendered using the psychophysics toolbox in Matlab (Brainard, 1997; Pelli, 1997). The factorial design was 2 (direction of adapting stimulus) by 5 (stimulus type). Direction could be approaching or receding motion. Stimulus type defined the adapting motion: Ah (high-rate flutter at 18.8 beeps/s), Vh (high-rate flicker, at 18.8 flashes/s), Vl (low-rate flicker, 9.4 flashes/s), AhVh (synchronized high-rate flicker and flutter), and AhVl (high-rate flutter combined with low-rate visual flicker).

The adapting visual stimulus (2 s duration) consisted of a uniform white disk expanding (from 0 to 9 degrees of visual angle) or contracting (from 9 to 0 degrees of visual angle) from the center of the screen at ±4.5 deg/s velocity. Flash duration was 27 ms on and 27 ms off in the high-rate flicker condition, and 27 ms on 80 ms off for low-rate flicker. The adapting sound stimulus (2 s duration) consisted of a 400 Hz triangular waveform generated using Cool Edit Pro software at 48 kHz sampling rate (Syntrillium Software Corp., www.syntrillium.com). Similar tonal stimuli have been successfully used in previous auditory looming experiments (e.g., Maier et al., 2004). The auditory adaptor rose (from 40 to 80 dB) or fell (from 80 to 40 dB) in sound pressure level (SPL) at ±20 dB/s. To create the high-rate flutter, auditory stimulus was windowed by intensity envelopes of 27 ms (8 ms ramps of Hann half-window) separated by 27-ms silence. The auditory MAE was measured using a 1.5 s test sound of the same frequency as the adapting sound stimulus. The test sound pressure level was 60 dB at the onset, and it increased or decreased at a velocity calculated according to a psychophysical staircase procedure (see next section).

Apparatus and procedure
Participants sat in a dark, sound-attenuated, testing room 57 cm away from a computer monitor (Mitsubishi, mod. N0701, 75 Hz). Sound was rendered via two multimedia loudspeakers attached to either side of the monitor (thus, the auditory image was effectively located by the center of the screen).

The auditory changing loudness aftereffect occurs when a tone that is falling or rising at a certain rate (in dB/s) is perceived as if it is steady (e.g., Kitagawa & Ichihara, 2002). The point of this subjective steadiness was assessed using a double-staircase method (Cornsweet, 1962) where the staircases were governed according to a PEST (Parameter Estimation by Sequential Testing) procedure (Taylor & Creelman, 1967). The adaptive staircases started with the test tone rising or falling at ±1.33 dB/s velocity and they were terminated when the adaptation step-size of the staircase was smaller than ±0.33 dB/s (from the initial ±2.66 dB/s).

Participants were presented with 10 blocks of adapting stimuli. There were two blocks (one approaching and one receding) of object motion for each of the 5 modality combinations. Presentation order of these modality combinations followed a 5-item Latin square design. Each block had the following structure: pre-adaptation measurement of the point of subjective auditory steadiness; 60 repetitions of adaptation stimulus separated by 200 ms; post-adaptation measurement of the point of subjective auditory steadiness. In the pre- and post-adapting blocks, the auditory steadiness was measured using the adaptive double-staircase described above. In the post-adaptation phase, each test stimulus presentation was preceded by 5 adaptation stimuli to preserve the after effect.

3 Results and discussion

Before running the main experiment, we tested our methodology using continuous audio-visual stimuli similar to ones used by Kitagawa and Ichihara (2002). Data from two subjects showed comparable results, with congruent audio-visual stimuli pairs producing auditory changing-loudness MAE in the range of 4 to 7 dB/s. The results from the main experiment using time-sampled adaptors (see Figure 1) show that auditory MAEs are approximately twice smaller than for continuous adapting stimuli. In fact, data from 2 participants in the main experiment had to be replaced (by testing 2 new participants) because no MAE was registered in more than half of the experimental conditions.

The magnitude of the motion aftereffect values for each modality combination was obtained as a difference between the averaged points of subjective steadiness from the two pre-adaptation staircases and the two post-adaptation ones. Individual MAE values were submitted to a within-subjects 2 (direction) x 5 (stimulus type) ANOVA with Greenhouse-Geisser correction for unequal variances. In line with Kitagawa and Ichihara’s (2002) results, neither the effect of direction of the adapting stimulus (F(1,14) = 2.19, p = 0.16) nor the interaction between direction and stimulus type (F(3,42) = 0.70, p = 0.56) reached significance. These results differ from Maier et al. (2004) study which argued that approaching audio-visual objects have a higher biological salience and thus should produce a perceptual bias. This disparity might be accounted due to the use of
discrete stimuli or to different methodology used, although a more dedicated study should address this question in the future.

Critically, the main effect of the stimulus type resulted significant $F(3,44) = 11.11, p < 0.001$, indicating significant variation in the magnitude of the MAE as a function of the particular combination of modalities used for adaptation. We used pair-wise non-adjusted (LSD) comparisons for further planned analyses of the observed effects of stimulus type, based on our two initial hypotheses.

First, we tested if time-sampled audio-visual object motion in depth produces results coherent with data from (Kitagawa and Ichihara, 2002). There was both unimodal and crossmodal MAE, with no significant difference ($p = 0.82$) between the aftereffect produced by high-rate flutter (Ah) and high-rate flicker (Vh). Furthermore, the combination of auditory and visual information in the adapting stimulus resulted in a significant increase of the aftereffect compared to single modality conditions. Namely, high-rate flicker synchronized with high-rate flutter (VhAh) produced larger MAE than the one by high-rate flutter (Ah) alone ($p < 0.05$) or by high-rate flicker (Vh) alone ($p < 0.01$).

Figure 1 Magnitude of the auditory aftereffect (in dB/s) after adaptation to time-sampled approaching (+) or receding (-) motion in depth of the following types: auditory, visual and audio-visual stimuli.
Next, we found that sparsely time-sampled visual stimuli produced less auditory MAE than visual motion sampled at a higher rate, as indicated by the significant difference (p < 0.05) between high-rate flicker (Vh) and low-rate flicker (Vl) conditions. In fact, low-rate flicker did not produce any significant after-effect (as compared with zero; t(14) = 0.036, p = 0.97, for approaching; and t(14) = 0.202, p = 0.84, for receding). Crucially, the addition of high-rate beeps to the low-rate flashes restored the MAE to the same levels as obtained with the high-rate flicker-flutter conditions. There was a significant difference (p < 0.001) between the MAE produced by low-rate flicker (Vl) alone and the MAE obtained for low-rate flicker synchronized with high-rate flutter (VlAh). More importantly, the MAE produced by VlAh condition was even larger (p < 0.05) than the MAE of the high-rate flutter alone (Ah). This suggests that apart from direct influence of sound on the MAE in VlAh condition, low-rate flicker now also had a significant adaptation effect not observed for such visual stimuli presented alone. Finally, no significant difference between both audio-visual conditions (VhAh vs. VlAh) was observed (p = 0.93), indicating that the size of their MAEs was comparable.

The above-mentioned results suggest that sparsely time-sampled visual motion may be filled-in by the addition of discrete auditory stimuli sampled at a higher rate. Interestingly, in a recent study Mastoropoulou et al. (2005) investigated the influence of sound effects on discrimination between video sequences (3 second walkthroughs in a virtual environment) sampled at 10, 12, 15, 20 and 24 frames per second (fps). Mastoropoulou et al. showed that naïve participants could reliably discriminate only between 10 vs. 24 fps pair in audio-visual conditions, however, without sound only two pairs were difficult to discriminate (10 vs. 12 fps and 12 vs. 15 fps). While the authors attributed their findings to divided attention effects, our results suggest that sound may exert a more direct impact, filling-in missing visual detail. The significant difference in the MAE between VlAh vs. Ah conditions suggest potential occurrence of illusory visual information. This effect may be based on the sound-induced visual flash illusion (Shams et al. 2000), which demonstrates that irrelevant sounds can induce the experience of extra visual flashes of a stationary flickering disc (see also Kamitani & Shimojo, 2001 for apparent motion experiment). On the other hand, our results might related to the “auditory driving” phenomenon (e.g. Welsh et al., 1986), where perceived flicker frequency can be increased or decreased by simultaneous presentation of auditory flutter. Further study should clarify this relationship measuring the auditory aftereffects produced by combinations of low-rate flutter with high-rate flicker.

4 Conclusions

The present findings reveal that discrete audio-visual motion stimuli follow the same cross-modal interaction patterns as continuous stimuli. More importantly, our results provide empirical evidence that sound can fill-in sparsely time-sampled visual motion, possibly arising from the occurrence of illusory visual events in the low-rate flicker-high-rate flutter conditions (VlAh). Similarly to Shams et al. (2000) findings, sound beeps might induce the illusory flashes which filled-in the low flash-rate stimulus. Such
perceptual “upgrading” explains the similarity between aftereffect sizes for V1Ah and VhAh conditions. Indeed, illusory visual stimuli can produce motion after-effects, as it has been shown in a study about perceptual filling-in of the blind spot area (Murakami, 1995).

The present results open an avenue for perceptual optimization of dynamic multisensory virtual environments and provide scientific support for limited animation techniques where rhythmic sound effects and music enhance the action smoothness (e.g. Walt Disney’s “Mickey Mousing” technique (Thomas & Johnston, 1981)). In order to further assess the extrapolation of this phenomenon to the creation of virtual environments, it will be important to investigate the role of higher order cognitive factors. Emotional context and attention might play role in influencing this auditory induced filling-in of dynamic visual information.

Acknowledgements

This work was carried out as part of the PRESENCIA project, an EU funded Integrated Project under the IST programme (Project Number 27731) and the Ministerio de Educación y Ciencia (TIN2004-04363-C03-02).

References


