The Perception of Auditory-visual Looming in Film

Sonia Wilkie and Tony Stockman

Queen Mary University of London

sonia.wilkie@eecs.qmul.ac.uk
tony.stockman@eecs.qmul.ac.uk

Abstract. Auditory-visual looming (the presentation of objects moving in depth towards the viewer) is a technique used in film (particularly those in 3D) to assist in drawing the viewer into the created world. The capacity of a viewer to perceptually immerse within the multidimensional world and interact with moving objects, can be affected by the sounds (audio cues) that accompany these looming objects. However the extent to which sound parameters should be manipulated remains unclear. For example, the amplitude, spectral components, reverb and spatialisation can all be altered, but the degree of their alteration and the resulting perception generated, need greater investigation. Building on a previous study analysing the physical properties of the sounds, we analyse peoples responses to the complex sounds which use multiple audio cues for film looming scenes, reporting which conditions elicited a faster response to contact time, causing the greatest amount of underestimation.

Keywords: Auditory-Visual Looming; Sound Design; Psychoacoustics

1 Introduction

A feature of film and gaming is interacting with objects that move in space, particularly objects that move in depth towards the viewer. Examples can be seen in 3-D presentations where objects appear to leap out of the screen towards the viewer; and in gaming where judgements are made to avoid or attack approaching objects.

The sound that accompanies these looming objects can affect the extent to which a viewer can perceptually immerse within the multidimensional world and interact with the moving objects. To accurately generate a dynamic and rich perception of the looming objects, the design of such complex sounds should be based on a firm scientific foundation that encompass' what we know about how we visually and aurally perceive events and interactions.

2 Previous Research and Practice

Previous research on auditory looming has revealed that people associate an approaching object with at least three attributes of sound, including interaural temporal differences, frequency change, and amplitude change [1].

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

In addition to finding that all three attributes of sound were associated with a looming object, they found that the change in amplitude elicited the fastest response to contact time, at the point in which the object passed, whilst the change in frequency prompted a response before the object had passed [1]. This underestimation of the contact time of a looming object, implies that the object is approaching at a faster rate and is anticipated to contact sooner.

Later studies on auditory looming showed that people overestimate the magnitude of intensity when presented with increasing stimuli [2],[3]. This implies that the increasing intensity of the approaching object is more dramatic than it physically is.

In an evolutionary context for both the physical and virtual worlds, these overestimations of magnitude and underestimation of contact time provide an advantage to the observer, giving them more time to prepare (an increased safety margin) for the objects arrival, and to initiate the appropriate response (being fight or flight), therefore increasing the chance of survival.

However, many of these previous auditory looming perception experiments [1],[2],[3],[4], have been conducted in extremely controlled conditions, with the aural stimuli consisting of simple tones (often a sine or triangle wave at 400 - 1000 Hz), and sound parameter manipulations such as an amplitude increase (between 10 - 30 dB), frequency change (using 804 Hz - 764.6 Hz, and 602.9 Hz - 572 Hz, which in musical terms equates to the tone and deviation of G5 $\pm$ 43 cents, and D5 $\pm$ 45 cents), and interaural temporal differences (a delay between the channels from 0.557 ms to 0.00 ms).

Limiting these variables used in experimental conditions compromises the ecological validity of the results, sound parameters manipulated, and real world application.

In contrast however, the film and gaming industries require sound designers to manipulate complex sounds, with the purpose of maximising the viewers experience, immersiveness, responsiveness to onscreen action, and overall perception of the virtual environment.

Examination of the sound manipulation techniques that sound designers and post production technicians use as cues for an approaching object in looming scenes provides a basis for a broader range of variables that can then be used in psychological studies on the perception of approaching objects.

Building upon our previous research [5] that examined the audio cues and techniques that sound designers use to generate the perception of an object moving in depth (looming), this research examines the percepts generated by the complex sounds.

3 Feature Analysis Studies

DSP analysis was previously conducted on the audio track of the 27 film looming scene samples used in this study, to understand which features the sound designers and post production technicians were using as cues for auditory looming,

how the features were manipulated, and the degree of the manipulation. Features that were analysed include: amplitude change; amplitude levels; amplitude slope; interaural amplitude differences; pan position; spectral centroid; spectral range; spectral spread; spectral flux; reverb; roll-off; and image motion tracking of the object.

In summary, our findings showed a number of similar techniques existed between the variety of samples. This includes:

- An average amplitude increase of 62.68 dB ($SD = 15.49$) on a linear / near-linear slope.
- The pan position centrally placed, and close to the image position, however fluctuates more than the image position. This fluctuation emphasises the spatial movement without having to hard pan to a single channel.
- An average spectral centroid increase of 1673.36 Hz.
- An average spectral flux increase of 167.0 Hz (with an average amount of flux of 13.8 Hz at the start of the sample, and 180.8 Hz at the peak).

In contrast to the previous auditory looming studies, the feature analysis of the film samples showed that they have:

- A greater range of variables used simultaneously to form complex looming stimuli (compared to the simple waves in the psychoacoustic studies).
- A greater increase in the levels that the variables were manipulated (ie 62.68 dB amplitude increase in the film samples, versus 10 - 30 dB in the psychoacoustic studies).

4 An Investigation of Responses to Complex Looming Sounds

This study is an extension of our previous research which examined the sound features in the looming samples, and will examine subjects responses to the looming stimuli that uses complex sounds produced by the sound designers and technicians.

4.1 Aim

The aim of this study is to determine if a subjects response to a looming object differs with the inclusion of complex designed sounds that use multiple audio cues, as opposed to looming scenes with no sound.

4.2 Hypothesis

It is hypothesised that the combination of the multimodal (auditory-visual) presentation (with the greater number of cues used, and the greater amount of stimuli change) will cause people to underestimate the contact time of the approaching object, thereby eliciting a faster response time than the looming scenes with no sound.

4.3 Method

4.3.1 Participants

A sample of 15 participants naive to the study purpose were recruited. They were Ph.D students and Postdoc. researchers from Queen Mary, University of London aged between 20 and 36 years ($\mu = 27.07$ years, $SD = 4.70$), with more male participants than female participants (11 male, 4 female).

4.3.2 Stimuli

The stimuli consisted of 27 film scenes that presented objects moving towards the viewer, and were comprised of both auditory and visual components. The scenes used are listed in *Table 1*. They were presented via computer with the visual stimulus presented on the monitor, and the auditory stimulus output through a pair of headphones.

The 27 scenes were presented in each of the three conditions - the multimodal (sound and image) condition, and the two unimodal conditions (sound only or image only). Each trial condition was presented once only (totaling 81 trial presentations) and in a randomised order.

4.3.3 Apparatus

Participants were located at a computer workstation with their head distanced approximately 40 cm from the computer monitor and eyes level with the centre of the monitor.

A Mac Pro 1.1 with a NEC MultiSync EA221WM (LCD) monitor was used. The screen size was 22 inches with the resolution set to 1680 x 1050 pixels and the display was calibrated to a refresh rate of 60 Hz.

The auditory stimulus was presented through Sennheiser HD515 headphones.

The program MAX / MSP / Jitter version 4.6 was used to construct the software application that presented the auditory and visual stimuli; presented the trials in a randomised and collected order, timed the participants responses using the computer's internal clock, and collected the participant responses in a text file.

4.3.4 Procedure

Participants sat at the computer workstation and were informed of the experiment procedure. They were given an information sheet summarising both the procedure and the ethics approval, signed a consent form, and completed a background questionnaire asking questions on gender, age, cinema experience and whether they have had corrections made to their vision or hearing.

Before commencing the experiment, the participants completed a practise test using 6 looming scenes (that were not additionally presented in the experiment). It was conducted as a supervised learning procedure to provide them with the opportunity to comprehend the experiment, the procedure, the micro time scale of the stimulus, and how to complete the task.

Participants were then instructed to start the experiment when ready.

#	Title	Year	Chapter, Time (min : sec)
1	The Matrix	1999	Chapter 1, 1:22 - 1:25
2	Star Wars (*Return of the Jedi*)	1983	Chapter 3, 0:20 - 0:24
3	Star Wars (*Revenge of the Sith*)	2005	Chapter 31, 3:08 - 3:09
4	X-men (*The Last Stand*)	2006	Chapter 15, 0:35 - 0:36
5	The Day After Tomorrow	2004	Chapter 12, 2:29 - 2:33
6	King Arthur	2004	Chapter 7, 10:46 - 10:48
7	Sherlock Holmes	2009	Chapter 22, 4:36 - 4:38
8	Van Helsing	2004	Chapter 17, 1:52 - 1:54
9	I Am Legend	2007	Chapter 17, 0:00 - 0:03
10	Troy	2007	Chapter 27, 2:22 - 2:24
11	Beowulf	2007	Chapter 2, 4:03 - 4:05
12	The Bourne Identity	2002	Chapter 12, 2:10 - 2:12
13	Charlie & the Chocolate Factory	2005	Chapter 15, 1:24 - 1:26
14	Mr and Mrs Smith	2005	Chapter 20, 0:40 - 0:44
15	Sin City	2005	Chapter 18, 1:06 - 1:07
16	28 Days Later	2002	Chapter 11, 0:01 - 0:04
17	Gattaca	1997	Chapter 21, 2:39 - 2:40
18	Alice in Wonderland	2010	Chapter 15, 0:19 - 0:20
19	Avatar	2009	Chapter 22, 1:42 - 1:45
20	Clash of the Titans	2010	Chapter 13, 4:11 - 4:13
21	Despicable Me	2010	Chapter 18, 2:23 - 2:24
22	Kill Bill vol2	2004	Chapter 6, 0:03 - 0:06
23	Mission Impossible 3	2006	Chapter 4, 1:06 - 1:08
24	Yogi Bear	2010	Chapter 1, 1:25 - 1:27
25	Final Destination	2009	Chapter 15, 0:06 - 0:07
26	Salt	2010	Chapter 9, 3:13 - 3:14
27	Saving Private Ryan	1998	Chapter 19, 3:17 - 3:21

Table 1. List of film scenes that were used in the experiment.

The task required participants to watch and/or listen to the scene of an approaching object, and to press the keyboard 'space bar' when they thought the object was closest to them.

Each trial lasted for a total duration of 0.5 - 4.0 seconds (depending on the looming scene presented) and a 6 second break was given between each trial. With a total of 81 trial presentations, the experiment lasted for approximately 25 minutes.

Participants were not given any information implying there might be correct, incorrect or preferred responses.

4.4 Results

Image motion tracking was previously performed on each scene to determine the approaching objects position and size, over time. For the purpose of this

study, the time (of the frame) in which the object was largest was considered the contact point and is called the 'peak'.

Participants responses to the stimuli (by pressing the keyboard 'space bar' when they thought the object was closest) was timed. This time was subtracted from the 'peak' time, to give the amount of time that was underestimated or overestimated, and for the purpose of this study is called the 'time to contact'.

Average time to contact (before and after peak time)
The condition which generated the least 'time to contact' (and was closest to the 'peak' time), was the *Image Only* condition (μ = 154.02 ms, SD = 681.05), followed by the *Audio-visual* condition (μ = 386.60 ms, SD = 548.87); and the *Audio Only* condition (μ = 443.59 ms, SD = 613.92).

Average time to contact (before peak time)
The condition that had the most number of trials in which the 'time to contact' was before the 'peak' time (therefore underestimating the contact time), was the *Audio Only* condition (with 25 trials, totaling 92.59% of the trials presented for that condition; weighted mean = 520.44 ms, weighted standard deviation = 391.87); and the *Audio-visual* condition, (with 24 trials, totaling 88.89% of the trials presented for that condition; weighted mean = 472.79 ms, weighted standard deviation = 249.79); followed by the *Image Only* condition (with 21 trials, totaling 77.78% of the trials presented for that condition; weighted mean = 324.94 ms, weighted standard deviation = 221.15).

Average time to contact (after peak time)
The condition that had the most number of trials in which the 'time to contact' was after the 'peak' time (therefore overestimating the contact time), was the *Image Only* condition (with 6 trials, totaling 22.22% of the trials presented for that condition; weighted mean = -170.93 ms, weighted standard deviation = 195.46); followed by the *Audio-visual* condition (with 3 trials, totaling 11.11% of the trials presented for that condition; weighted mean = -86.19 ms, weighted standard deviation = 94.92); and the *Audio Only* condition (with 2 trials, totaling 7.41% of the trials presented for that condition; weighted mean = -76.85 ms, weighted standard deviation = 35.30).

No trials had an average contact time during the image 'peak' (which had a duration of 41.67 ms, or one frame at 24 fps), with no individual participants indicating contact during this time.

4.5 Discussion

The results indicate that the *Image Only* condition had the slowest response to the contact time both before and after the peak time, with the least amount of underestimation before the 'peak' time and greatest amount of overestimation after the 'peak'.

However, the *Audio Only* condition, although still only providing unimodal information about the approaching object, prompted participants to have the fastest response to contact time overall, both before and after the peak image frame, with the greatest amount of underestimation before the peak time and least amount of overestimation after the peak. This suggests that the addition of sound and looming audio cues (in both the *Audio Only* condition and the *Audio-visual* condition) prompted people to underestimate the contact time more often, and with a greater time frame, than the scenes that had no sound.

5 Conclusion

Although the individual sound parameters that act as the audio cues for an approaching object could not be controlled and varied in this study, this investigation of the complex sounds in there original form as created by the sound designers has shown that the addition of sound, and the multiple techniques used to create audio cues, cause people to underestimate the contact time of an approaching object. This result suggests that further investigation is warranted, with future research on the complex stimulis' individual sound parameters, as independent variables, and the perception generated as a result.

References

1. Rosenblum, L, Carello, C, & Pastore, R. (1987). *Relative effectiveness of three stimulus variables for locating a moving sound source.* Perception, 16, 175186.
2. Neuhoff, J.G. (2001). *An adaptive bias in the perception of looming auditory motion.* Ecological Psychology, 13 (2), 87-110.
3. Neuhoff, J.G., & Heckel, T. (2004). *Sex differences in perceiving auditory looming produced by acoustic intensity change.* In Proceedings of ICAD 04-Tenth Meeting of the International Conference on Auditory Display, Sydney, Australia.
4. Cappe, C., Thut, G., Romei, V., & Murray, M. M. (2009). *Selective integration of auditory-visual looming cues by humans.* Neuropsychologia, 47, 1045-1052.
5. Wilkie, S., Stockman. T., & Reiss, J. D. (2012). *Amplitude Manipulation For Perceived Movement In Depth.* Audio Engineering Society, 132nd Convention, Budapest.
6. Ghazanfar, A.A., Neuhoff, J.G., & Logothetis, N.K. (2002). *Auditory looming perception in rhesus monkeys.* In Proceedings of the National Academy of Sciences, USA 99, 15755-15757.
7. Maier, J. X., Chandrasekaran, C., Ghazanfar, Asif A, Spemannstrasse, B. C., Germany, T., & Procedures, E. (2008). *Integration of Bimodal Looming Signals through Neuronal Coherence in the Temporal Lobe.* Current Biology, (18), 963-968.
8. Maier, J.X., & Ghazanfar. A.A. (2007). *Looming biases in monkey auditory cortex.* Journal of Neuroscience, 27 (15), 4093-4100.
9. Maier, J, Neuhoff, J, Logothetis, N, & Ghazanfar. A. (2004). *Multisensory integration of looming signals by rhesus monkeys.* Neuron, 43 (2), 177-181.
10. Neuhoff, John G. (2004). *Ecological psychoacoustics: introduction and history.* Ecological Psychoacoustics. Elsevier Academic Press, California.
11. Rosenblum, L, Wuestefeld, A., & Saldana, H. (1993). *Auditory looming perception: Influences on anticipatory judgements.* Perception, vol 22, 1467-1482.