"SITTING TOO CLOSE TO THE SCREEN CAN BE BAD FOR YOUR EARS": A STUDY OF AUDIO-VISUAL LOCATION DISCREPANCY DETECTION UNDER DIFFERENT VISUAL PROJECTIONS

Ashley Walker and Stephen Brewster

University of Glasgow
Department of Computing Science
17 Lilybank Gardens,
Glasgow, G12 8QQ
http://www.dcs.gla.ac.uk/~stephen
stephen@dcs.gla.ac.uk

ABSTRACT

In this work, we look at the perception of event locality under conditions of disparate audio and visual cues. We address an aspect of the so called "ventriloquism effect" relevant for multi-media designers; namely, how auditory perception of event locality is influenced by the size and scale of the accompanying visual projection of those events. We observed that recalibration of the visual axes of an audio-visual animation (by resizing and zooming) exerts a recalibrating influence on the auditory space perception. In particular, sensitivity to audio-visual discrepancies (between a centrally located visual stimuli and laterally displaced audio cue) increases near the edge of the screen on which the visual cue is displayed. In other words, discrepancy detection thresholds are not fixed for a particular pair of stimuli, but are influenced by the size of the display space. Moreover, the discrepancy thresholds are influenced by scale as well as size. That is, the boundary of auditory space perception is not rigidly fixed on the boundaries of the screen; it also depends on the spatial relationship depicted. For example, the ventriloquism effect will break down within the boundaries of a large screen if zooming is used to exaggerate the proximity of the audience to the events. The latter effect appears to be much weaker than the former.

1. INTRODUCTION

Our natural environment provides us with sensory cues through several different sense modalities (vision, audition, touch, temperature, taste, etc,). Properties of these environmental stimuli – particularly spatial and temporal properties – are coupled due to the physical laws governing their generation. The speed of light and speed of sound, for example, have a fixed relationship, as do the propagation directionalities of light and sound, their effects at the surfaces of solid objects, etc. Numerous studies of multi-sensory perception have demonstrated that information received from the various sensory systems are not processed independently [1]. On the contrary, multi-sensory couplings are believed to be integral to our perceptual model of the world.

The integration of multi-sensory cues is clearly important in the synthesis of information unavailable from a single sense source. This information may be borne of a grouping process, but also may be understood as a general improvement in interaction comprehension. The everyday example of opening a door is a good case in point. Here, two cues – the audible click of a door's locking mechanism combined with a synchronous change in the resistance offered by the handle or knob – are perceptually integrated into a single successful "open" event.

The implication of this for the multimodal designer is that she cannot merely be a jack of all design trades (graphic, audio, haptic, etc.). She must also master an understanding of inter-modal effects. These include an awareness of perceptual sensitivities to temporal asynchronies between multi-sensory signals, the interaction between multi-sensory stimuli with discrepant spatial and temporal rate information, as well as cross-modal effects in attention. These factors affect the quality as well as the intelligibility of events.

In this work, we look at the perception of event locality under conditions of disparate audio and visual cues. In particular we examine how auditory perception of event locality is influenced by the size and scale of the visual window through which those events are viewed. This question confronts people involved with a variety of applications (e.g., film, TV, computer games and GUI's) and by people on either side of the production-consumption exchange. Each time a director calls for a zoom or wide-shot, for example, she opens up the possibility that similar egocentric changes will be needed in the multi-track audio. Likewise, when the proud owner of a home-theatre system decides to move the projector back for a bigger picture, he should worry about the placement of the speaker array.

In the rest of this section, we present a brief review of studies relevant to the topic of audio-visual location discrepancy detection. The literature spans a range from sensory and cognitive psychology to multimedia design.

1.1. Audio-visual space: a literature review

When perception in one modality depends on stimuli in another modality an inter-sensory bias is said to exist. A prime example is the influence that eye orientation exerts on the perceived direction of a sound source. When visually fixating a laterally displaced point, the apparent location of an auditory source is shifted in the direction opposite to the eccentric gaze orientation [2]. Another familiar example of an inter-sensory spatial bias is the so called "ventriloquism effect", wherein the perceived location of an auditory source is influenced by the presence of an associated visual object [3,4]. When there is a moderate mismatch between the location of a pair of associated visual and audio stimuli, the perceived location of the auditory stimuli is

shifted toward the actual location of the visual stimuli (and one perceives no discrepancy).

This phenomenon depends upon a number of factors, including the amount of discrepancy between the audio and visual sources, the temporal synchronization of the stimuli, and the reporting requirements. The visual capture strength also increases with the cognitive compellingness of the stimulus situation (e.g., the effect is strengthened if the participant has observed that the visual stimulus has produced the auditory stimulus in the past). Early studies of the ventriloquism effect using puppets, for example, showed that factors like removing immobile facial features from a puppet weakens capture [3]. Nevertheless, the effect can operate over quite a wide angle. It is not uncommon to hear reports of 30-40 degrees (in the median plane). Also, it is possible for frontal visual stimuli to capture sound sources located behind an observer, e.g., capture has been reported for 160 degrees but not for 140 degrees presumably because 160 exploited the strong frontal capture which occurs at 20 degrees. It appears that auditory localization in adults (with constant head sizes) is calibrated to visual space.

1.2. Audio-visual space: questions for multimedia designers

The ventriloquism effect, or "spatial magnetization" as it is called in the cinema, has been exploited by media designers for decades. Throughout the first four decades of the sound cinema, directors relied on these phenomena to give apparent spatial attributes to a monaural sound-track. When an actor walks across the screen, for example, the seen position of his footfalls determines their heard position. Irregardless of whether the sound-track plays from a single speaker behind the screen, a speaker array distributed around the cinema, or through a pair of headphones at the drive-in, the impression it makes is of many mini loudspeakers positioned behind the screen – each resounding events in their proper locale [5].

When multi-track sound recording and playback technology became available, it was natural for sound-track designers to want to wrap spatial sound cues around the audience as a kind of wide-angle sound shot. However, initial literal attempts to spatialize sound led to what came to be labeled as an undesirable "in the wings effect" – i.e., a disconcerting feeling that we are to believe audio-visual space is being extended past the boundaries of the screen and into the theatre [5]. Audiences are generally confused about how to interpret sound which (by design or speaker imbalance) emanates from above the EXIT sign or toilets. While some sound-track designers learned to achieve their ends by relying on carefully (and individually) crafted blends of real and psychologically spatialized sound, others learned to use spatial sound to replace (as oppose to accompany) visual events. Where the filmgoer is immersed in a super sound-scape – with crickets chirping in the wings and helicopters zooming overhead, the off-screen sounds often fill-in the world around the scene. In this regard, films with more lush, emmersive sound-tracks, e.g., *Blade Runner* [6], *Hair* [7] and *The Mission* [8], can focus in on visual details without many wide-angle shots.

By contrast to the subtle and sophisticated use of spatial sound in the cinema, the sound employed in games and computer interfaces typically succeeds with the simple and literal spatialization that failed in the cinema. In these applications, sounds which emanate from a particular direction feed-forward cues that direct the audiences' attention through space as well as time. Computer users, unlike cinema-goers, expect that there is much more story material in the box than on the screen and, consequentially, come to a game or GUI with an expectation of navigation.

As distinctions between the projector vs. monitor and between the computer vs. cinema fade, there arises a question as to where audio events should be located with respect to the mobile and flexible container of the screen. This paper attempts to contribute an empirically tested answer to this question. To this end, we measured qualitative changes in auditory localization ability (or, equally, audio-visual discrepancy angle detection) in the presence of two commonly encountered projection scenarios; namely, resizing and rescaling.

Our first hypothesis is that sensitivity to audio-visual location discrepancies will occur at the edges of the screen. That is, sound emanating from an angular position within the visual cone (defined by the angle between the head and the horizontal edges of the projection space) will be spatially magnetized by a visual event in the center of the screen. On the other hand, sound played from more peripheral angles will appear to be off-screen and produce an unpleasant "in the wings effect".

The "in the wings effect" intrigued us because it is difficult to come up with an ecological explanation. It may arise from the audience's naïve assumption that speakers are mounted coincident with the display apparatus. By contrast, we hypothesized the existence of another projection effect that seems to have a more compelling explanation. Our second hypothesis is that sensitivity to audio-visual location discrepancies will be influenced by an observer's apparent proximity to a scene. If this is true, it should be possible to increase discrepancy detection thresholds by zooming (on a fixed size screen). It seems ecologically valid to suspect that the auditory system becomes more sensitive to angular cues as the proximity of a sound source increases (even if "proximity" is only visually implied).

2. MATERIALS AND METHODS

EXPERIMENTAL SCENARIO. Sixteen people from University of Glasgow served as participants. (The participants ranged in age from 18-22, five women and eleven men.) The experiment was a counter-balanced within-groups design with display size and zoom as the independent variables. Each of the participants performed the auditory localization task described below using a (i) full-screen wide-angle view (Condition 1), (ii) half-screen wide-angle view (Condition 2), and (iii) full-screen zoom view (Condition 3). Trials within each condition were randomly ordered for presentation.

Participants watched and listened (over headphones) to the animation described below. Within each condition, the user watched the same visual event twenty times, but the spatialization of the audio cues accompanying the final/key event was laterally displaced to the left or right of center. A range of audio-visual discrepancy angles were used in order to determine the discrepancy threshold angle (or minimum audible discrepancy angle) associated with each visual condition.

Following the presentation of each trial, users were asked where they heard the stimuli (left, right or center), in addition to two

dummy questions. The latter tested participants' familiarity with the scene. (The requirement to answer these questions effectively forced participants to watch as well as listen.) Participants also gave informal feedback following the whole experiment.

ANIMATION SCENARIO. A pair of cartoon beach balls (Figure 1) were bounced back and forth across an uncluttered stretch of beach. The balls bounced in opposite directions (as if being tossed between two off-screen players) until one falls and comes to a rest. Once the ball game ends, the display was cleared and replaced by a questionnaire.

The animation panels were designed and built in Macromedia Flash [9] and exported as a series of GIF files. These, in turn, were used in a more flexible Java-based animation interface [10] that ran in a browser. The monitor occupied a visual angle of approximately 28 degrees. Two of the animations were presented on a screen encompassing this full width, while the third encompassed half of the width.

Each time a ball impacted the beach, a characteristic impact sound was played. The sampled impact sound had a duration of 25 ms, with a sharp attack and slow decay. All impact sounds were monaural except the final series of impact sounds associated with a ball dropping. These were spatialized using twenty different levels of inter-aural intensity disparities. This spatialization was generated using an HRTF model [11] and the twenty levels correspond roughly with those resulting from lateral displacement of a virtual source by 20 degrees to the left or right.

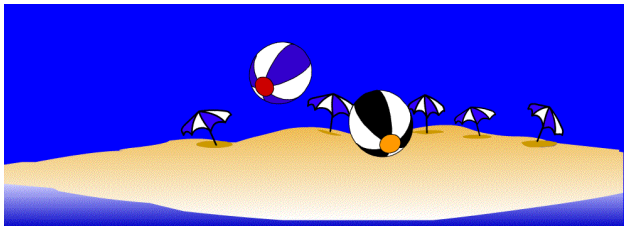

Figure 1. Visual stimuli.

3. RESULTS

Sensitivity to audio-visual angle discrepancies was affected by projection size and scale. Contrary to our hypothesis, the resizing effect was much stronger than the zooming effect.

The average minimum detected discrepancy angle in each of the conditions is summarized in the table below.

Condition	**Average angular threshold**
Large screen, wide angle	12.7 (degrees)
Small screen, wide-angle	7.8
Large screen, zoom	11.7

Table 1. Average audio-visual discrepancy detection thresholds.

A signal-factor ANOVA revealed a significant projection effect $F_{2,45} = 13.2$ and $P_{2,45} = 0.00003$. A Tukey HSD test was run to compare the individual effects of size and scale. The average detection thresholds associated with the large and small screens differed significantly ($q = 4.9$, $P_{2,45} = 0.01$). The difference between the average thresholds of the wide-angle and zoom conditions did not differ significantly ($q = 1.0$).

It is interesting to note that the width (center-to-edge) of the projection space in the small and large screen projections were approximately 7 and 14 degrees. These figures match up well with the angular sensitivities in the Table 1 above. However, it must be said that, due to the variation of inter-aural directional cues with participant head size (and other anthropomorphic details), we cannot be sure that angles actually heard by participants were exactly those generated by our HRTF model.

4. DISCUSSION

Our qualitative study into the effect of visual projection on audio-visual location discrepancy detection suggests that the ventriloquism effect breaks down near the boundaries of screen. In other words, what an observer takes as center of the perceptual field – both auditory and visual – is influenced by the width of the screen, with observers perceiving the left or right register of a stereo sound more acutely when observing correlated images in a small screen. Sensitivity to audio-visual location discrepancies is not only related to the absolute size of the display but also depends, to a much lesser degree, on the depicted proximity between observer and events. When a scene is zoomed to imply greater proximity, observers are appear to be slightly more sensitive to the left right register of sound.

Contrary to our hypotheses, the zooming effect is much weaker than the size effect. Although we can argue for the scaling effect on the basis of ecological validity, it is more difficult to provide an explanation for the strength of the screen size effect. Nevertheless, the size effect has been observed by other sources. A spatial disparity tolerance study, conducted within the framework of high definition television, yielded a comparable result [12]. This study positioned observers in front of a 72-inch set (which occupied a visual half-angle of approximately 15 degrees) while they listened to sound from one of ten speakers arranged in a semi-circle. The result was that novice participants perceived (and reported a mild annoyance as a result of) sound that emanated from an offset position of approximately equal to the screen width (just within 20 degrees). Moreover, as we cited earlier, film sound-track designers have filed anecdotal reports of this phenomena several decades ago.

In the case of the cinema, audiences might have some naïve assumption about the locus of cinematic technology being around the screen; however, it seems unlikely that such an assumption would be in operation in our study. Here, the boundary of the visual animation stimuli resided within the container of a browser that, in turn, resided within the container of a computer monitor. The audio hardware had a separate and obvious location of its own.

As a concluding remark on the study, we want to caution that the actual values reported in any of the conditions we observed are less important than the trends suggested by our results. The effect of scale, although less significant than size, is also an interesting result which deserves further study. Clearly one challenge in interpreting our results arises from our use of a single set of audio cues to denote near events (zoom condition)

as well as far events (wide-angle condition). Realistic cues, i.e., environmentally generated cues, arising from events at different ranges would contain distance information (e.g., intensity variations and reverberation effects) intermingled with the angular cues. Our failing to attenuate, for example, cues associated with the wide-angle scene may have biased the study. However, it is difficult to argue against the trend we observed by saying that attenuated sound could have made the spatial audio more detectable in the wide-angle case.

5. CONCLUSIONS

Multimedia design and authoring tools are evolving faster than our understanding of inter-sensory perceptual effects. Studies that empirically explore inter-sensory perceptual boundaries are badly needed. Toward that end, we looked at the perception of event locality under conditions of disparate audio and visual cues. Our results suggest that perception of event locality is influenced by the size and scale of the visual window through which those events are viewed.

This finding has relevance to people involved with both production and consumption of multimedia presentations. The consumer of a new home-theatre system, for example, should understand that installation of a projector strongly determines the placement of a speaker array. Alternatively, we might say this the other way around (i.e., a speaker installation determines the perceptually effective positions for the projector), as audio plays a crucial role in communicating perspective in a multimedia presentation. In fact, home theatre studies are confirming that audio and video are of equal importance in communicating the director/producer's intentions to the audience. For example, in one such study, the contribution of sound and picture to the appreciation of "space" was investigated using several screen sizes and audio reproduction technologies. Although the appreciation of space improved steadily with screen size, ratings of the importance of visual information only reached those of audio at the widest screen setting [13].

Ensuring that audio and video presentations elements cooperate to communicate perspective must start long before a presentation reaches the shelf. Scripting the right spatial relationship between audio and video story-telling elements is, of course, essentially the responsibility of the designer and producer. Let's look at a sophisticated example of this from the cinema. What Francis Coppola and Walter Murch accomplished in the mix of *Apocalypse Now* [15], for example, is increasingly being seen as a solution that began long before the sound-track reached the dubbing stage. Sound designer Randy Thom explains:

"… it began with the script, and with Coppola's inclination to give the characters in 'Apocalypse' the opportunity to listen to the world around them… The degree to which sound is eventually able to participate in storytelling is now recognized to be more determined by the use of time, space, and point of view [pov] in the story than by how often the script mentions actual sounds. Most of the great sound sequences in films are "pov" sequences. The photography, the blocking of actors, the production design, art direction, editing, and dialogue have been set up such that we, the audience, are experiencing the action more or less through the point of view of one, or more, of the characters in the sequence. Since what we see and hear is being filtered through their consciousness, what they hear can give us lots of information about who they are and what they are feeling" [15].

6. ACKNOWLEDGEMENTS

This work was funded by EPSRC GR/L 79212 and a grant from Microsoft Corporation.

7. REFERENCES

[1] A. Kohlrausch, and S. van de Par, "Audio-Visual Interaction: From Fundamental Research in Cognitive Psychology to (possible) Applications," *Human Vision and Electronic Imaging IV*, vol. 3644, pp. 34-44, 1993.

[2] J. Lewald, "The effect of gaze eccentricity on perceived sound direction and its relation to visual localization" *Hearing Research.* vol. 115, pp. 206-216, 1998.

[3] I. P. Howard and W. B. Templeton, *Human spatial orientation*, Wiley, New York, 1966.

[4] W. R. G. Thurlow and C. E. Jack, "Certain determinants of the 'ventriloquism effect'," *Perceptual and Motor Skills*, vol. 36, pp. 1171–1184, 1973.

[5] M. Chion, *Audio-Vision: Sound on Screen*, Columbia University Press, New York, 1990.

[6] R. Scott (director), *Blade Runner*.

[7] M. Forman (director), *Hair*.

[8] R. Joffe (director), *Mission*.

[9] Macromedia, *Flash v.4*, http://www.macromedia.com.

[10] Sun, Java v.1.2, http://java.sun.com.

[11] C. P. Brown, *Modeling the Elevation Characteristics of the Head-Related Impulse Response*, Thesis Report, San Jose State University, 1996.

[12] S. Komiyama, "Subjective evaluation of angular displacement between picture and sound for HDTV sound systems," *J. Aud Eng Soc*, vol. 37, pp. 210 – 124, 1989.

[13] S. Bech, V. Hansen, W. Wozczyk, "Interaction between audio and visual factors in a home theatre system: Experimental results", *Proc. 99th Convention Aud Eng Soc*, New York, preprint no. 4096 (K-7, 1999.

[14] F. Coppola, *Apocalypse Now*.

[15] R. Thom, "Designing a film for sound", http://filmsound.studienet.org