THE EFFECT OF AUDIOVISUAL CONGRUENCY ON SHORT-TERM MEMORY OF SERIAL SPATIAL STIMULI: A PILOT TEST

Benjamin Stahl

University of Music and Performing Arts
Institute of Electronic Music and Acoustics
Leonhardstraße 15, 8010 Graz, Austria
benjamin-cosimo.stahl@student.kug.ac.at

Katharina Vogt

University of Music and Performing Arts
Institute of Electronic Music and Acoustics
Leonhardstraße 15, 8010 Graz, Austria
vogt@iem.at

ABSTRACT

This paper is motivated from the question if the use of spatial sounds enhances learning in multi-modal teaching aids. In a basic pilot study the consolidation of serial stimuli was tested for unimodal conditions (visual; auditory) and a bimodal condition (spatially congruent, audiovisual). In contrary to our hypothesis, the audiovisual condition did not show better results than the visual one. In this particular test the auditory display was clearly inferior to the two other ones.

1. INTRODUCTION

One possible field of applications for multimedia displays that present spatially congruent audiovisual events realized by state-of-the-art technologies for holophonics such as speaker arrays and ambisonics are learning and teaching aids like animated illustrations of learning contents.

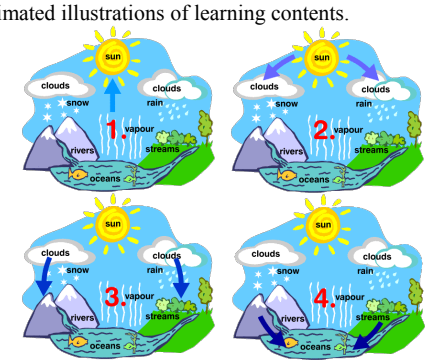

Figure 1: *Example of an animated illustration of a simple learning content (original GIF image: http://i268.photobucket.com/albums/jj11/heffa44/watercycle.gif)*

Figure 1 shows an image series taken from a typical GIF animation approaching a simple topic. Such animations convey information as a series of events at different locations on the image. When learning a topic in such a way, one of the most important aspects is how well the series of spatial events is consolidated into the student's short-term memory.

As part of a seminar on sonification, we came up with the idea of doing a basic test on the effect of adding congruent auditory stimuli to simple visual sequences concerning the memory of such sequences.

2. RELATED WORK

Concerning virtual environments (VEs), Larsson et al. (2001) showed in a study, that an audiovisual display was superior to a purely visual one with respect to the participants' performance in simple searching tasks. Moreover, the participants found the audiovisual display more pleasant and sensed a higher degree of presence [1].

Wenzel et al. (2014) showed a positive effect of bimodal VEs compared to unimodal ones regarding the navigation in such environments under certain conditions [2].

From a low-level neuroscientific and psychological point of view, Spence (2007) points out that the key factors deciding whether a pair of auditory and visual stimuli are bound by an observer are temporal and spatial congruency, correlation in temporal patterns and semantic congruency [3].

Teder-Sälejärvi et al. (2005) conducted a study on the reaction time and event-related potentials (ERPs) measured in the participants' EEG comparing auditory, visual, congruent and non-congruent audiovisual stimuli. He found out that reaction times to bimodal stimuli were shorter than to unimodal ones, but didn't differ comparing congruent and non-congruent stimuli; however, differences in the ERPs could be observed in that comparison [4].

3. EXPERIMENT

To test our hypothesis, we designed an experiment that compared the participants' ability of reconstructing a series of spatial stimuli in a visual (V), an audiovisual (AV) and an auditory (A) modality. The V stimuli were circles that lit up with a random (gray-scale) noise pattern inside, the A stimuli were white noise bursts from the particular directions, and the AV stimuli were a combination of both.

We recruited 6 unpaid healthy[1] students (age 22 – 28) to participate in our experiment, of which 5 were male and 1 was female.

The experiment was realized by putting a screen in front of the participant (distance: 3 m) and projecting six white circle outlines on the screen at eye level (1.22 m). The centers of the circles were 0.325 m apart from each other, which -from the participant's point of view- resulted in azimuth angles of -15.2°, -9.2°, -3.1°, 3.1°, 9.2° and 15.2°. The circle diameters were 0.12 m. Behind the acoustically transparent screen we set up six loudspeakers, one for each circle. The SPL at the participant's position was 57.5±1 dB(A), when an auditory stimulus occurred. Graphic rendering was realized with *Processing*, sound synthesis with *Pure Data*. *OSC* was used to synchronize the events.

In the visual modality, the insides of the circles lit up with a random pattern for 1.8 s per stimulus and an interstimulus interval of 0.7 s in a sequence of either five or seven stimuli; the order of the stimuli's positions was random, and it could also occur that a stimulus was presented

[1] no hearing impairment, no or corrected visual impairment, no known learning or concentration impairment

more than once in a row at the same position. Figure 2 shows an example of a visual series containing 5 stimuli.

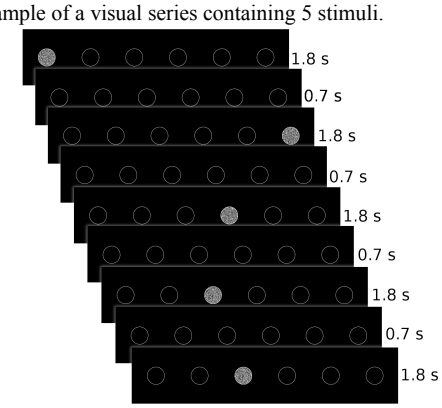

Figure 2: *Temporal succession of a visual sequence*

In the audiovisual modality, a noise burst was played back by the respective loudspeaker when a circle lit up. The noise bursts were temporally and spatially congruent to the visual stimuli.

In the auditory modality, the circles didn't light up, the spatial sequence was only encoded by the positions of the noise bursts.

At the end of a sequence, the participant had to re-construct it by clicking into the circles in the observed order.

At the beginning of the experiment, the participants did short learning sessions in all 3 modalities.

The learning sessions were followed by 4 test sessions, that each contained 6 sequences (2 V, 2 AV and 2 A) of 5 stimuli and 6 sequences of 7 stimuli. The order of these 12 sequences in a session was random. The presented sequences as well as the participant's answers were written to file. When a session was finished, the participant was asked to take a 2-minute break before starting a new session.

At the end of the test, the participants were asked in which modality the series were easiest to remember and in which they were hardest.

4. RESULTS AND DISCUSSION

To determine the correctness of the participants' answers, the average error was calculated out of each sequence. The circles on the screen were indexed with the numbers 0 to 5 from left to right, and in each sequence, the presented order and the response order were represented as progressions p[n] and r[n], where the progression value at n is equal to the index of the presented/responded stimulus. Equation (1) shows how the average error is calculated in a progression containing N stimuli.

$$AvE = \frac{1}{N}\sum_{n=1}^{N}|r[n] - p[n]| \quad (1)$$

We collected 48 data series (6 participants · 4 sessions · 2 sequences) in each of the 6 categories (5 V stimuli; 5 AV stimuli; 5 A stimuli; 7 V stimuli; 7 AV stimuli; 7 A stimuli) and calculated the means and standard deviations of the average errors. For the experiment is a pilot test, a consideration of the statistical significance was spared.

The results in Table 1 show that we couldn't find a clear difference between the visual and the audiovisual display. In the auditory display, the average error was clearly higher both for series with 5 and with 7 stimuli.

sequence length	*five stimuli*			*seven stimuli*		
modality	*V*	*AV*	*A*	*V*	*AV*	*A*
μ	*0.063*	*0.175*	*0.667*	*0.351*	*0.321*	*0.911*
σ	*0.261*	*0.508*	*0.548*	*0.582*	*0.533*	*0.495*

Table 1: *Means and standard deviations of the average reconstruction errors in the six testing categories*

This finding was also confirmed by the participant's subjective impression: all of them considered the series hardest to remember, when they were presented using an auditory display. 4 of 6 participants found the series easiest to remember, when they were presented audiovisually, 1 found the audiovisual display equal to the visual one and 1 preferred the visual display. The relatively great average error in the auditory display is probably a result of the difficulty of localizing the sound source in a short time and then encoding the stimulus to a spatial map. The average error both in the audiovisual and in the visual display was relatively small, which indicates that the participant's cognitive load was rather low, some of the participants reported it was pretty easy for them to remember the sequences using number progressions.

5. OUTLOOK

We would like to retest the hypothesis that spatially congruent audiovisual stimuli are easier to remember than unimodal visual stimuli in a new experiment with more participants and a new test design.

Giving the participants more time to learn to localize the spatial sounds could improve their performance in remembering the auditory and audiovisual sequences.

The cognitive load could be enhanced by either doing a dual-task test or increasing the length of the sequences.

6. ACKNOWLEDGMENT

We would like to acknowledge the contribution of Johanna Reichert from the Department of Psychology at the University of Graz to the literature research.

Furthermore we'd like to say thank you to Georgios Marentakis, Franz Zotter and Marian Weger from the Institute of Electronic Music and Acoustics at the University of Music and Performing Arts Graz for their scientific and technical support.

7. REFERENCES

[1] P. Larsson, D. Västfjäll and M. Kleiner, "Ecological Acoustics and the Multi-Modal Perception of Rooms: Real and Unreal Experiences of Auditory-Visual Virtual Environments," in *Proceedings of the 2001 International Conference on Auditory Display,* Espoo, 2001.

[2] E. M. Wenzel, M. Godfroy-Cooper and J. D. Miller, "Spatial Auditory Displays: Substitution and Complementarity to Visual Displays," in *Proceedings of the 20th International Conference on Auditory Display*, New York, 2014.

[3] C. Spence, "Audiovisual multisensory integration," *Acoustical Science and Technology*, vol. 28, no. 2, pp. 61-70, 2007.

[4] W. A. Teder-Sälejärvi et al., "Effects of Spatial Congruity on Audio-Visual Multimodal Integration," *Journal of Cognitive Neuroscience*, vol. 17, no. 9, pp. 196-1409, 2005.