Detecting and using temporal correlations between random sequences of auditory and visual events

Armin Kohlrausch[1,2], Steven van de Par[2]

[1] *Technische Universiteit Eindhoven, Human-Technology Interaction, 5600 MB Eindhoven, The Netherlands*

[2] *Philips Research Laboratories, 5656 AA Eindhoven, The Netherlands, Email: armin.kohlrausch@philips.com*

Introduction

In this study we investigate the ability of human observers to detect the temporal coherence between auditory and visual patterns. For this purpose we used stimuli consisting of a white disk that was moving up and down on a computer screen, accompanied by an amplitude modulated tone. The disk movement was driven by a number of Gaussian pulses (25-ms standard deviation), each of which moved the disk upwards and downwards, back to its starting position. Each of these pulses can be considered to represent a single visual event. The tone was amplitude modulated with a number of Gaussian pulses (auditory events) in a similar way as the visual stimulus. Thus, only during the movement of the disk (the pulses) the tone was audible, at any other time point the tonal amplitude was zero.

In the first set of experiments we investigated to what extent subjects were sensitive to the synchronicity of the auditory and visual events. In a second set of experiments we investigated whether the synchrony between auditory and visual events could be used as a cue for binding auditory and visual stimuli.

Sensitivity to temporal jitter in AV patterns

In the first experiments subjects were presented with a 2-Interval Forced-Choice 2-down 1-up adaptive tracking procedure for measuring the thresholds for detecting temporal jitter between the auditory-visual stimulus events as a function of the number of pulses (AV events). Subjects had to indicate which of the two intervals contained the synchronous stimulus. After each trial feedback was given about the correctness of the response. We have chosen to let subjects select the synchronous interval instead of the asynchronous interval because it substantiates the question of whether subjects can perceptually bind auditory and visual stimulus patterns. This question will be addressed more explicitely in the second set of selection experiments that will be described further on.

Each visual stimulus was presented during 2000 ms. During the interval ranging from 200 ms after the start of the visual stimulus presentation to 200 ms before the end of the stimulus presentation, a number of events was generated, each at a random moment within this interval as is indicated by the vertical lines in the upper part of Fig. 1. The probability distribution of the event timings was uniform. For each of these events, the visible disk moved up and down following a Gaussian pulse trajectory as a function of time with a standard deviation of 25 ms. The Gaussian pulse trajectories were simply added, thus two pulses that were very close resulted in a larger movement than each separate pulse could have caused.

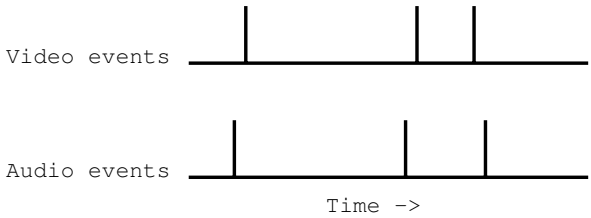

Figure 1: *Schematic representation of a reference interval of the first experiment with jitter between each auditory and visual event. Each vertical line represents a visual or auditory event.*

The visual events were used to generate a number of tonal pulses of 500 Hz with the same Gaussian envelope (auditory events). These tonal pulses were added in phase to avoid cancellation effects to constitute an auditory stimulus. This was done in two different ways for the two intervals of each trial:

- The auditory and visual events were synchronous (target interval).
- The auditory and visual events were jittered, thus each AV event randomly had either an audio or video lag (reference interval) as is shown in Fig. 1 by the vertical lines.

In Fig. 2 the thresholds for detecting AV jitter are shown for four subjects as a function of the number of AV pulses. As can be seen, some of the subjects are able to already detect 20-ms AV jitter. These low threshold values are especially found when 4 to 10 pulses were presented, while fewer pulses consistently led to larger thresholds, and 20 pulses led to larger thresholds for some of the subjects.

The rather low threshold values for about 5 pulses are remarkable in two respects. First of all, the threshold values of 20 ms are small compared to data reported in literature which are based on constant AV delays instead of jittered delays. Dixon and Spitz [1] measured thresholds of 180 ms and 75 ms for their most critical stimulus when audio or video were delayed, respectively. Secondly, there is a reduction in threshold for multiple pulses as compared to the threshold of a single pulse. This seems to indicate that interactions between successive pulses may play a role. A hypothesis is that subjects are us-

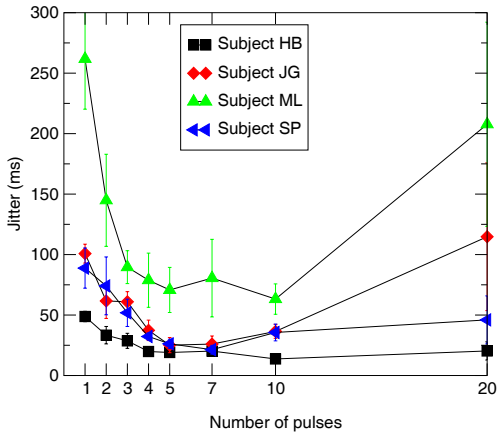

Figure 2: *Threshold values for detection AV jitter are shown as a function of the number of AV pulses for four subjects.*

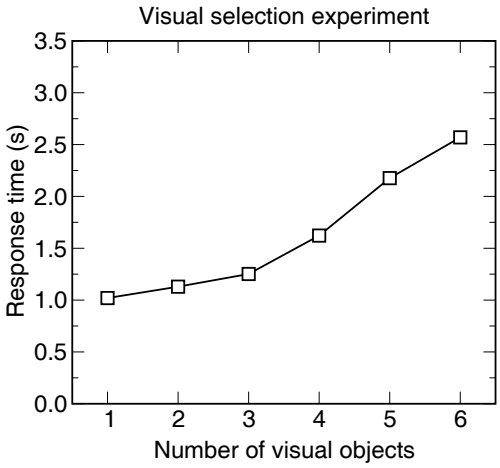

Figure 3: *Response times for selecting the visual object as a function of the number of visual objects. For the datapoint with one visual object, subjects had to indicate whether the auditory and visual patterns were coherent or not.*

ing an additional cue, besides the asynchrony cue, when more than one AV pulse is presented. The cue that might be used could be a comparison of the temporal intervals between successive visual events and successive auditory events. This pattern of intervals is identical in the synchronous target interval, but deviates across the auditory and visual modality in the asynchronous reference interval.

To test whether such a cue could have been used, the previous experiment was repeated with five pulses. A constant AV delay was applied to both the reference and the target interval. As a consequence, none of the auditory-visual intervals was synchronous and therefore synchronicity per se could not not be a cue. However, the pattern of intervals in the auditory and visual domain could still be compared in this case. Thus if this cue were used, jitter thresholds should not be affected by the constant delay that was introduced. The results show that jitter thresholds stay nearly constant for overall delays between -100 and + 100 ms. Despite this lack of physical synchrony, subjects were able to discriminate between the intervals with and without jitter which suggests that some kind of cross modal pattern comparison must have been the basis for performing the task.

Selection of visual objects based on cross modal coherence

In these selection experiments, we were interested to see whether subjects could use the cross-modal coherence of AV patterns to select one visual object from multiple concurrent visual objects based on its coherence with an auditory cueing stimulus. A number of simultaneous visual objects were presented, horizontally separated by about 4 cm at a viewing distance of about 50 cm. Each object moved according to an independent random pattern of Gaussian pulses (3.5 pulses per second, 25 ms standard deviation). The pulse pattern of one of the visual objects was used to generate an audio pattern in a similar way as in the first experiments. Subjects had to select as fast as possible the visual object which had the same pattern of pulses as the auditory stimulus, with the restriction that they should score as accurate as possible. Evaluation of the results obtained in this experiment showed that subjects scored at least 80 % correct on this task. A variable number of visual objects was used as an independent variable to measure the response times of subjects for selecting the visual object. In one condition a single visual object was presented, for which subjects had to indicate whether the auditory and visual patterns were coherent or not.

As can be seen in Fig. 3, response times increase considerably with the number of visual objects. Specifically going from 3 to 6 objects, the response time increases nearly linearly. Such a result is consistent with the hypothesis that subjects are following a serial search strategy where they have to observe each visual object for some time in order to decide whether it is coherent with the auditory pattern or not. There seems to be a constant offset in response times which may be caused by the time that is needed for the planning and the actual entering of the response on the computer keyboard. The serial search hypothesis is confirmed by reports of the subjects about the strategy they used to perform the task. In addition, a similar experiment, in which *one* visual pattern was used to select the one coherent out of several auditory patterns basically showed the same dependence of response time on the number of presented auditory objects.

These results with artificial stimuli show that AV temporal processing provides basic mechanisms that facilitate the segregation and ordering of objects both in the visual and auditory domain.

References

[1] N. F. Dixon and L. Spitz. The detection of audiovisual desynchrony. Perception **9** (1980), 719–721.