Proceeding of Peach Summer School 2007, http://www.peachbit.org
In collaboration with Psychology Journal, http:// www.psychology.org

Peach Summer School 2007 Extended Abstract:

Temporal calibration between the visual, auditory, and tactile senses: A psychophysical approach

Tonja Machulla, Massimiliano Di Luca, Marc O. Ernst
Max Planck Institute for Biological Cybernetics, Tuebingen, Germany

***Abstract* – Temporal miscalibration between different feedback devices in multimodal virtual environments may decrease observers' sense of immersion and presence. However, subjective perception of synchrony may vary from physical synchrony. Using a psychophysical paradigm we assessed the perception of synchrony for three different modality pairs (audiovisual, audiotactile, visuotactile). For each pair we found that one of the two modalities has to lead considerably in order for the two to be perceived as synchronous. Further, observers were most sensitive to asynchrony in the visuotactile modality combination. Taking these findings into account we offer general suggestions for the temporal calibration of multimodal virtual environments.**

***Keywords* – Temporal Calibration, Synchrony Perception**

I. INTRODUCTION

Human observers acquire information about physical properties of the environment through different sensory modalities. For *natural events*, these sensory signals show a specific temporal, spatial and contextual configuration that aids the integration into a coherent multisensory percept. For *multimodal virtual environments*, however, signals have to be created and displayed separately for different modalities, which may result in a miscalibration of these signals. This, in turn, can greatly reduce the observer's sense of immersion and presence.

Using a psychophysical approach, we investigate how the human brain binds information from different senses into one coherent representation of the environment. Our aim is to use this knowledge to make suggestions for the calibration of multimodal virtual environments as well as develop a number of 'perceptual tricks' that decrease the observer's sensitivity to inconsistencies in the virtual reality setup.

In the present paper, we focus on the temporal aspect of multimodal binding processes. Watching somebody speak provides us with physically synchronous visual and auditory information. The understanding of the verbal message is partially influenced by the visual information, which is most evident in the well-know McGurk effect [1]. However, this effect depends on the perceived temporal alignment of signals [2].

There are two factors that influence the perception of synchrony and temporal order of two signals from different modalities. Firstly, physical transmission times differ greatly between the senses: light travels much faster to the eye than sound travels to the ear. Secondly, neural processing times also differ between the senses [3].

These factors may lead to physically synchronous information to be perceived as being asynchronous. In turn, stimuli that are presented with a temporal off-set that corresponds to the processing time difference between the involved modalities are perceived as being simultaneous.

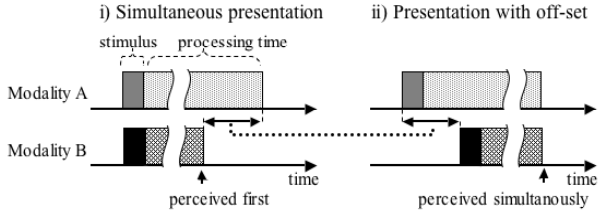

Figure1. Synchrony Perception across modalities

In the following, we examine the perception of subjective simultaneity of signals for several modality combinations. How does the observer align corresponding signals and how sensitive is he to asynchrony between different modalities? measured this temporal off-set for several modality combinations (audiovisual, audiotactile, visuotactile) as well as the sensitivity of the observer for deviations from this point of subjective simultaneity (PSS).

II. METHOD

Participants. Participants were students of the Karl Eberhards University of Tuebingen, recruited from the Max Planck Institute subject database. They reported normal hearing and normal or corrected-to-normal vision. All participants were naive to the purpose of the study. Participation was voluntary; all participants provided written consent and were paid for their time. All

participants were run individually and they completed the experiment in a session lasting about 30 minutes.

Apparatus. The experiment was conducted on a PC; the software to control the experimental stimuli was written in MatLab. The stimuli were presented using a custom-built device designed to generate co-located sound, vibration, and light with high temporal accuracy. The sound was produced by two vertically aligned speakers. The distance between the centers of the speakers was 7.5 cm. A vibration device (an electro-magnetic shaker, Monacor Bass Rocker BR25) was situated between the speakers. It was mounted on a damping mass, and thus produced vibrations without audible noise. A seven by five array of red LEDs (1.6 cm x 1.3 cm) was mounted on top of the vibration device, serving as a vibrating surface as well as a light source. All devices (speakers, shaker, LEDs) were driven using the same type of amplifier. Stimuli were generated by a multi-channel sound card (M-audio 1010LT) using ASIO drivers capable of simultaneously playing and recording the signals. The background noise produced by the computer fan was 30** dB SPL.

Stimuli. Stimuli were generated using sinusoids with linearly ramped (1 ms) onsets and offsets to prevent DC artifacts (e.g., audible clicks from the speakers at the onset of the sound). The sampling frequency of the sinusoids was 44100 Hz and their duration, including the two ramps, was 100 ms. Sounds were produced using 1000 Hz sinusoids at 61 dB SPL intensity, vibrations were produced using one wavelength at 40 Hz which created the sensation of a light 'tap' on the finger, lights were produced by reversing the negative half of the sinusoid at 145 Hz frequency.

Design. We used a factorial design with two within-subjects factors: combination of modalities (audiovisual (AV), visuotactile (VT), audiotactile (AT)) and stimulus onset asynchronies (SOA) between the two stimuli (–240 ms to 240 ms in steps of 40 ms). A negative SOA value was arbitrarily assigned to cases were either visual stimuli preceding auditory, visual preceding tactile or tactile preceding auditory. Each combination of SOA and modality pair was randomly presented 10 times, for a total of 390 experimental trials.

Procedure. Participants sat at a table in a dark, sound-attenuated room. They placed there left index finger onto the vibrating surface. Since visual stimuli were also presented at that location participants were instructed to maintain fixation on their finger throughout the entire experiment. On any given trial two stimuli from different modalities were presented. Participants were required to judge which signal had been presented first. Responses were enter over one of three buttons using the right hand.

To acquaint subjects with the task a block of 30 training trials was presented. Feedback was given after the participant pressed a response key. Wrong button presses (e.g., answer ‘tactile’ on an AV trial) were indicated by a low-pitch tone, correct button presses were indicated by a high-pitch tone. Experimental trials were similar to training trials, except that no feedback was provided.

Table 1. Mean PSS and Mean JND in ms (Standard Errors in Parentheses) for each Modality Combination

	Modality Combination		
	Audiovisual	Audiotactile	Visuotactile
PSS (SE)	28 (11)	55 (10)	34 (10)
JND (SE)	127 (12)	119 (11)	80 (8)

III. RESULTS

Data from the practice trials and data with wrong button presses (e.g., response ‘tactile’ on an AV trial) were excluded from analysis. For each participant the proportion of sound-first responses were calculated for each SOA for the AV and the AT modality combinations and the proportion of tactile-first responses for each SOA for the VT modality combination. We then fitted these data with a cumulative Gaussian distribution using the pfit function in MatLab and 500 repetitions of the bootstrap procedure (www.bootstrapsoftware.org). From the fit, we obtained the point of subjective simultaneity (PSS) and the just noticeable difference (JND) at 75%. PSS indicates at which temporal off-set between the two stimuli the observer is maximally uncertain about the temporal order of presentation. The JND reflects the amount of deviation from subjective temporal alignment that will be detected in 50 percent of the cases.

Mean PSS and mean JND across subjects were computed for each modality pair. Results are illustrated in Table 1. Mean PSSs were each submitted to a one-sample t-test, revealing a significant deviation from 0 for all three modality pairs (audiovisual: $t(15) = -2.609$, $p = 0.02$; audiotactile: $t(15) = -5.770$, $p < 0.001$; visuotactile: $t(15) = -3.497$, $p < 0.01$). JND analysis showed that participants were more sensitive to subjective asynchrony for the VT modality pair than either the AV ($t(15) = 4.982$, $p < 0.001$) or the AT modality pair ($t(15) = 4.026$, $p = 0.001$).

IV. DISCUSSION

Subjective synchrony perception deviates substantially from physical synchrony for all three tested modality combinations. Since physical transmission differences in our setup are negligible for auditory and visual stimuli (< 1 ms) and not existent for tactile stimuli, PSSs most likely reflect differences in neural conduction time between the senses. At this point, it has to be mentioned that neural conduction times dependent on properties of the stimuli

such as intensity or length. Therefore, the absolute value of the PSS should not be overemphasized. However, our results agree with previous research concerning the order required for perception of synchrony. For instance, the overall finding for AV stimuli is that the visual modality has to be presented before the auditory [4, 5, 6].

We conclude that temporal delays between feedback devices in VRE are less likely to perturb user experience if the lagging channel is the one that requires more neural transmission time. In particular, observers are less sensitive to:

- auditory signals lagging behind visual,
- auditory signals lagging behind tactile,
- tactile signals lagging behind visual.

Further, since observers are most sensitive to asynchrony in the visuotactile modality pair most efforts in calibrating multimodal VRE should go towards increasing the temporal precision between the visual and the tactile feedback devices.

V. ACKNOWLEDGEMENTS

Supported by EU grant 27141 "ImmerSence", SFB 550-A11, and the Max Planck Society . Special thanks to Katja Mayer.

REFERENCES

[1] McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. Nature, 264(5588), 746-748.

[2] Munhall, K. G., Gribble, P., Sacco, L., & Ward, M. (1996). Temporal constraints on the Mcgurk effect. Perception & Psychophysics, 58(3), 351-362.

[3] King, A. J. (2005). Multisensory integration: Strategies for synchronization. Current Biology, 15(9), R339-341.

[4] Lewald, J., & Guski, R. (2003). Cross-modal perceptual integration of spatially and temporally disparate auditory and visual stimuli. Cognitive Brain Research, 16(3), 468-478.

[5] Slutsky, D. A., & Recanzone, G. H. (2001). Temporal and spatial dependency of the ventriloquism effect. Neuroreport, 12(1), 7-10.

[6] Zampini, M., Guest, S., Shore, D. I., & Spence, C. (2005b). Audio-visual simultaneity judgments. Perception & Psychophysics, 67(3), 531-544.