DOI: 10.61782/fa.2023.1050

TEMPORAL MISMATCH EFFECTS IN SHORT-TERM MEMORY OF AUDIO-VISUALLY PRESENTED SPOKEN DIGITS

Lukas J. Vollmer[1]* **Cosima A. Ermert**[1] **Janina Fels**[1]

[1]Institute for Hearing Technology and Acoustics, RWTH Aachen University, Germany

ABSTRACT

In everyday life, humans are confronted with stimuli from multiple sensory modalities which must be associated with, or segregated from each other to form meaningful object-based percepts. A key physical attribute of these stimuli is their spatiotemporal coincidence, especially in the auditory and visual domains. For example, maintaining the synchrony of audio and video streams is one of the main requirements for online conferencing tools.

In a previous study, spoken digits (auditory input) were presented with their corresponding graphemes (visual input) at various stimulus onset asynchronies (SOAs). When SOAs were noticeable, the results indicated a tendency to ignore one of the input streams among participants. Consequently, it is assumed that the lack of temporal correspondence between these stimuli reduced the detectability of SOAs and that SOAs were easy to ignore, when detected.

Here, we investigate distracting effects of SOAs on verbal serial recall capability using audio-visual speech recordings. The temporal correspondence of these stimuli is increased by the natural cooccurrence of speech and lip movements. Key questions address whether potential stimulus onset asynchrony effects occur with audio-visual speech stimuli and if these effects reflect the perceptual asymmetry of simultaneity judgments.

Keywords: *multi-sensory integration, short-term memory, temporal mismatch, lip movement*

1. INTRODUCTION

Multisensory integration processes have received increasing attention over the last decades, the primary focus being perceptual processing of multisensory stimuli [1–3]. Only few studies have investigated these integration effects in the context of other cognitive processes such as working memory [4]. Although there is evidence for multisensory integration effects in working memory, Quak et al. [4] conclude in their review that our understanding of multisensory working memory processes is still incomplete.

In a previous study [5], results of a serial recall task using spoken digits and digit graphemes were evaluated to assess if working memory takes advantage of simple, semantically redundant multisensory stimuli. Stimulus onset asynchronies (SOAs) were introduced between the auditory and visual stimuli, to modulate multisensory integration [6]. No significant effect of SOAs on working memory performance could be found.

Based on an informal temporal order judgment task (for a detailed description of the task see [7]) and studies by Wassenhove et al. [6] and Bhat et al. [8], it was expected that all participants should notice at least the largest SOAs of ±400 ms. However, only 68 % of volunteers reported in a post-experiment questionnaire that they noticed the SOA. Consecutively, they reported to have attended to only one of the stimulus modalities. Therefore, it was hypothesized that a lack of temporal correspondence between the “fluctuating” auditory stimuli and the “static” visual stimuli made it easy to ignore the asynchrony. This hypothesis was tested in a similar experiment using a video of the corresponding lip movements instead of graphemes as the visual part of the stimuli.

**Corresponding author*: *lukas.vollmer@akustik.rwth-aachen.de.*

2. METHODS

Thirteen volunteers (aged 21 to 29 years, normal hearing, and normal or corrected-to-normal vision) participated in this pilot study which included an audio-visual serial recall experiment. In the paradigm, each trial started with a visual cue of $1.5\,\mathrm{s}$ followed by a cue-stimulus interval that was randomly sampled from a uniform distribution between 600 and $800\,\mathrm{ms}$. After the cue stimulus interval, digits one to nine were presented (once, without repetition) in pseudo-random order. Presentation could consist of the unimodal stimulus components in isolation (conditions A, V), or their audio-visual combination including SOAs in steps of $100\,\mathrm{ms}$ between -400 and $400\,\mathrm{ms}$ (conditions AV-400, $\ldots$, AV400), see Fig. 1. Positive SOAs indicate that the visual stimuli were presented first, reflecting the natural order in which components of an audio-visual event arrive at an observer. The retention interval between digit sequence presentation and sequence recall was $3\,\mathrm{s}$. The sequence recall interface showed all nine digits on a three-by-three grid in random order. Selected digits disappeared from the screen and error corrections were not possible.

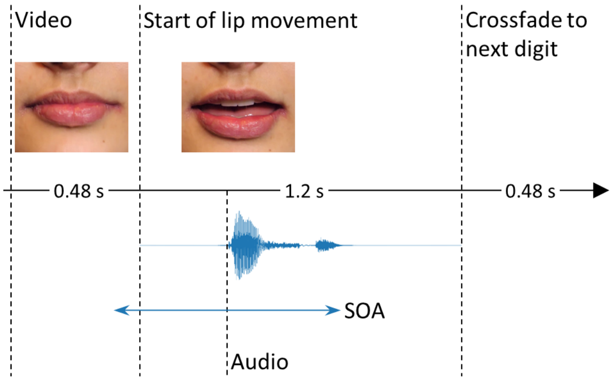

Figure 1. This figure shows how SOAs were presented. The videos included $480\,\mathrm{ms}$ still images of the first and last frame before and after the lip movement, respectively. Closed lips were defined as the start and end point of the articulation phase. The audio signal accounted for inaudible articulatory movements, such that SOAs could be implemented relative to the start of the lip movement instead of the start of audible articulation.

All presentation conditions were repeated eight times, which enabled the calculation of proportion-correct scores for each presentation condition and serial position within the digit sequence, independent of the actually presented digits.

Subjective evaluations of task difficulty, SOA notability, the SOA's effect on rehearsal strategy as well as modality specific focus were collected in a post-experiment questionnaire.

3. RESULTS

The proportion-correct scores were first evaluated by a one-way repeated-measures analysis of variance (ANOVA) to assess the effect of presentation modality between conditions A, V, and AV0. Secondly, a Friedman test (due to violated normality assumption) was carried out to assess the effect of SOA between conditions AV-400, $\ldots$, AV400. Fig. 2 shows boxplots of the proportion-correct scores obtained for modality (left) and SOA (right). The center graph depicts the mean proportion-correct scores grouped by modality (excluding $SOA \neq 0\,\mathrm{ms}$) for each serial position.

3.1 Effect of Presentation Modality

The one-way repeated-measures ANOVA revealed a significant effect of presentation modality on proportion-correct scores, $F(2,24) = 35.306$, $p = 7.098 \times 10^{-8}$. Bonferroni-corrected pairwise t-tests indicate significant differences between auditory and visual presentation, $t(12) = 7.269$, $p = 2.965 \times 10^{-5}$, as well as audio-visual and visual presentation , $t(12) = 7.055$, $p = 3.987 \times 10^{-5}$. There was, however, no significant difference between auditory and audio-visual presentation, $t(12) = 0.931$, $p = 0.37$.

3.2 Effect of Stimulus Onset Asychnrony

The Friedman test for an effect of SOA was not significant at the $\alpha = 0.05$ significance level, $\chi^2(8) = 8.343$, $p = 0.4$. Therefore, no further tests, were performed.

Qualitatively, the AV300 condition is conspicuous due to its compactness around the $60\,\%$ performance level. There is, however, no explanation for these results based on study design, implementation, or participant reports. It seems unlikely that an asynchrony effect should be limited to a narrow time window, especially in the more natural case of delayed auditory stimuli. Therefore, the qualitative difference between AV300 and all other asynchrony conditions is attributed to the small number of participants.

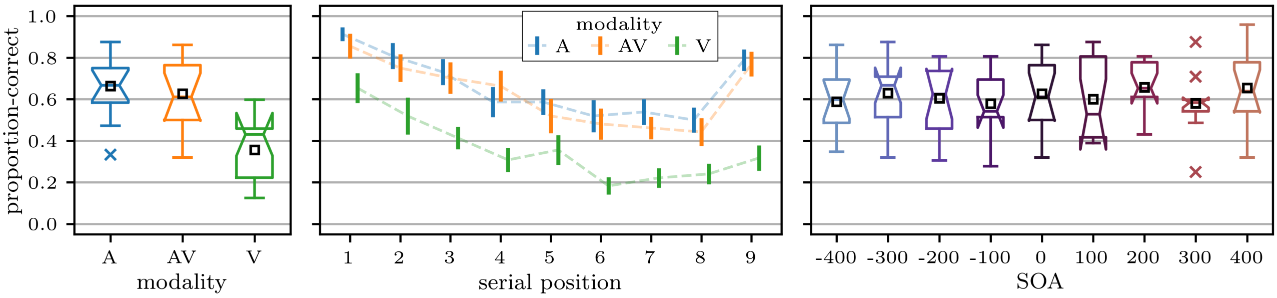

Figure 2. *Left*. Median boxplots of the proportion-correct scores for presentation conditions A, V, and AV0. The black squares indicate the mean score. Performance decreased from A over AV0 to V. *Middle*. Mean proportion-correct scores and standard error lines for each serial position grouped by the same presentation conditions as on the left. Performance in the visual condition is generally lower by about 20%, also the recency effect is less pronounced. *Right*. Median boxplots for the nine SOA conditions, including black squares to indicate mean performances.

3.3 Questionnaire Results

The task was evaluated as difficult ($\hat{\mu} = 5.385$, $\hat{\sigma} = 0.961$) on a seven-point Likert scale. All of the thirteen participants noticed the SOAs, however they mostly reported that SOAs did not strongly affect their ability to memorize the digits ($\hat{\mu} = 3.385$, $\hat{\sigma} = 1.805$). Further, all volunteers reported that they focused on the acoustic stimuli and mostly ignored the visual stimuli ($\hat{\mu} = 1.923$, $\hat{\sigma} = 1.115$, $1 \widehat{=}$ audio, $7 \widehat{=}$ video focus). Perceptual evaluations of the SOAs were mixed. Some participants reported that they only noticed negative SOAs (audio leading), others noticed SOAs in both directions. Among those who noticed both SOA directions, evaluations which direction was more "difficult" were inconclusive.

4. DISCUSSION

Although there was a significant effect of presentation modality, the audio-visual presentation did not lead to a significant improvement of recall performance. Rather, there is a tendency that auditory-only presentation leads to slightly better performance than audio-visual presentation (c.f., Fig. 2). This is consistent with the results reported in [5], where digits were presented as graphemes. In contrast to the previous study, however, recall performance was significantly better after presentations including auditory stimuli than after visual-only presentation. These results are most likely explained by the necessity to perform lip-reading of the visual stimuli. The clear tendency that participants focused on the auditory stimuli supports this explanation. Further, visual stimuli tend to exhibit no or a less pronounced recency effect [9–11] (c.f., center graph of Fig. 2 between serial positions eight and nine), which likely contributes to lower general performance, although to a lesser extent than lip-reading.

Another contrast to the previous study [5] is that all participants noticed the SOAs, although mostly in conditions where the auditory stimuli were presented first. This indicates that (1) the temporal correspondence between spoken digits and lip movements increased the perceptual sensitivity to SOAs, and (2) the perception of SOAs was asymmetric, which is in line with the temporal window of integration concept [6, 7].

5. ACKNOWLEDGMENTS

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 368482240/GRK2416 and SPP2236 - 444724862. The authors would like to thank Ignacio Merten for conducting this pilot study.

6. REFERENCES

[1] C. Spence and C. Frings, "Multisensory feature integration in (and out) of the focus of spatial attention," *Attention, Perception, & Psychophysics*, vol. 82, pp. 363–376, July 2019.

[2] L. Chen and J. Vroomen, “Intersensory binding across space and time: A tutorial review,” *Attention, Perception, & Psychophysics*, vol. 75, pp. 790–811, May 2013.

[3] C. Spence, “Audiovisual multisensory integration,” *Acoustical science and technology*, vol. 28, no. 2, pp. 61–70, 2007.

[4] M. Quak, R. E. London, and D. Talsma, “A multisensory perspective of working memory,” *Frontiers in Human Neuroscience*, vol. 9, Apr. 2015.

[5] L. J. Vollmer, J. Burger, C. A. Ermert, and J. Fels, “Stimulus onset asynchronies and audio-visual serial recall performance,” in *Tagungsband - DAGA 2022 : 48. Jahrestagung für Akustik*, 2022.

[6] V. van Wassenhove, K. W. Grant, and D. Poeppel, “Temporal window of integration in auditory-visual speech perception,” *Neuropsychologia*, vol. 45, pp. 598–607, Jan. 2007.

[7] J. Vroomen and M. Keetels, “Perception of intersensory synchrony: A tutorial review,” *Attention, Perception, & Psychophysics*, vol. 72, pp. 871–884, May 2010.

[8] J. Bhat, L. M. Miller, M. A. Pitt, and A. J. Shahin, “Putative mechanisms mediating tolerance for audiovisual stimulus onset asynchrony,” *Journal of Neurophysiology*, vol. 113, pp. 1437–1450, Mar. 2015.

[9] M. W. Battacchi, G. M. Pelamatti, and C. Umiltà, “Is there a modality effect? Evidence for visual recency and suffix effects,” *Memory & Cognition*, vol. 18, pp. 651–658, Nov. 1990.

[10] S. J. Schlittmeier, J. Hellbrück, and M. Klatte, “Does irrelevant music cause an irrelevant sound effect for auditory items?,” *European Journal of Cognitive Psychology*, vol. 20, pp. 252–271, Mar. 2008.

[11] M. Hurlstone, “Serial recall,” in *The Oxford Handbook of Human Memory* (M. J. Kahana and A. D. Wagner, eds.), Oxford University Press, 2022.