Supplementary Material: Accelerating Saccadic Response through Spatial and Temporal Cross-Modal Misalignments

Daniel Jiménez-Navarro
djimenez@mpi-inf.mpg.de
Max Planck Institute for Informatics
Saarbrücken, Germany

Xi Peng
xp2011@nyu.edu
New York University
Brooklyn, USA

Yunxiang Zhang
yunxiang.zhang@nyu.edu
New York University
Brooklyn, USA

Karol Myszkowski
karol@mpi-inf.mpg.de
Max Planck Institute for Informatics
Saarbrücken, Germany

Hans-Peter Seidel
hpseidel@mpi-inf.mpg.de
Max Planck Institute for Informatics
Saarbrücken, Germany

Qi Sun
qisun@nyu.edu
New York University
Brooklyn, USA

Ana Serrano
anase@unizar.es
Universidad de Zaragoza
Zaragoza, Spain

CCS CONCEPTS

• **Computing methodologies → Perception**; **Virtual reality**; **Mixed / augmented reality**.

KEYWORDS

Audiovisual integration, cross-modal interactions, multisensory perception, saccadic latency, virtual reality

ACM Reference Format:
Daniel Jiménez-Navarro, Xi Peng, Yunxiang Zhang, Karol Myszkowski, Hans-Peter Seidel, Qi Sun, and Ana Serrano. 2024. Supplementary Material: Accelerating Saccadic Response through Spatial and Temporal Cross-Modal Misalignments. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers '24 (SIGGRAPH Conference Papers '24), July 27-August 1, 2024, Denver, CO, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3641519.3657432

This Supplementary Material document contains additional details for the following sections:

- (S1) Main experiment: Additional details on stimuli. Beep and Flash paradigm (S1.1) and temporal simultaneity in audiovisual perception (S1.2)
- (S2) Gaze data processing
- (S3) Hardware details
- (S4) Details on model fitting (S4.1), statistical analysis for both the main experiment and the basketball application (S4.2), and results of the interactive farm game (S4.3)
- (S5) Main user study surveys: Demographic (S5.1) and Sickness (S5.2) questionnaires

S1. MAIN EXPERIMENT: ADDITIONAL DETAILS ON STIMULI

In this section, we include further details and discussions for the beep and flash paradigm used during the main experiment and the temporal limits for audiovisual simultaneity perception.

S1.1 Beep and Flash paradigm

Simple stimuli are widely used to establish foundational understandings of sensory processing. They are highly controllable and are less likely to be influenced by higher cognitive processes than more complex stimuli like speech reading, which engage specific language processing areas and require the integration of more complex, meaningful information.

Simple stimuli are thought to engage basic sensory integration mechanisms that are more universally present across different individuals and contexts, making findings more generalizable across a wider range of situations and populations [11]. The *beep and flash* setup is thought to tap into fundamental neurophysiological processes related to temporal resolution and integration in the brain, which are relevant across various types of more complex tasks [10]. Simple auditory and visual stimuli can be precisely controlled and manipulated in virtual settings, allowing for a clearer interpretation of results and the underlying processes. *Beep and flash* was selected due to its potential applicability to the audiovisual cues prevalent in interactive gaming environments. These environments often rely on the rapid perception of simple, meaningless synchronized auditory and visual signals to inform user actions and reactions. Such stimuli are analogous to the *beep and flash*, providing an experimentally sound basis that is highly relevant to the temporal processing demands encountered in gaming applications. This approach allows us to explore the fundamental neurophysiological mechanisms of multisensory integration with direct implications for enhancing user experience in virtual settings.

Nevertheless, further research using more complex, naturalistic tasks is necessary to fully understand how these principles extend

---

SIGGRAPH Conference Papers '24, July 27-August 1, 2024, Denver, CO, USA

ACM ISBN 979-8-4007-0525-0/24/07
https://doi.org/10.1145/3641519.3657432

to real-world multisensory experiences. Our study can be positioned as a crucial step in building toward a more comprehensive understanding of temporal integration across a spectrum of stimuli complexities. In Section 4 in the main document, where the application cases are proposed, a more realistic basketball scene is tested to overcome *beep and flash* limitations.

S1.2 Audiovisual Perception: Temporal Simultaneity

According to literature, audiovisual cues are perceived as temporally simultaneous as long as the temporal shift (ΔT) between their onset times is located within the limits of the Temporal Binding Window (TBW) [12]. This TBW has been reported to vary between individuals [13], time instants due to recalibration procedures and training [8] [14], and the task asked to the participants [5].

This temporal shift tolerated between visual and auditory cues such that the stimuli are recognized as a single event has been measured for different tasks under diverse conditions. Several works have addressed temporal audiovisual integration in several speech-related tasks and object-based videos [15], reporting an asymmetric TBW when this audiovisual integration keeps taking place. It is presented also how this TBW is not centered at the zero delay point (ΔT=0), but it is a bit shifted towards the visual leading part instead. However, it is not difficult to tell how these results can be biased by our prior knowledge. It is expected in normal conditions, and even more in human speech, to perceive the visual cue before the auditory effect: our experience tells us that it is not common to hear the human voice before seeing the lips' motion.

In VR, tolerance limits to audiovisual temporal binding are studied and its effect on presence and immersion feelings [7]. The average minimum temporal shift noticed by participants in their setup was reported to be 320 ms. A similar limit value is reported in [6] as the temporal threshold from which the temporal shift starts to be perceivable, thus the audiovisual musical cues were not perceived as simultaneous anymore. Based on these studies, we set our maximum ΔT to 400 ms.

S2. GAZE DATA PROCESSING

In this section, we provide the methodology to detect visual saccades and the procedure followed to discard spurious cases. We also show in Fig. 1 a sample of all recorded saccades for one of our participants to illustrate the saccadic profiles obtained during our experiment.

Saccades are obtained for all sessions using the algorithm explained in [1] (Sec. 3.2). Velocity thresholds used are both theoretically and empirically tuned. Saccades are detected using a threshold of $120°/s$, which is enough to detect the typically surpassed value of $100°/s$ [1]. To detect the visual saccade starting point (anchor threshold) and the saccade end (fixation threshold), a conservative value of $60°/s$ is selected so fixation movements are not included [3]. Once all the saccades are detected, the first detected saccade after each target onset is identified and checked to be a correct user response by comparing the target's real eccentricity location and the saccade landing point. Saccades are discarded if they were performed too early, the starting or ending points could not be detected in the specified range, or gaze tracking was lost during the saccade. Following previous literature [9] [4] [16] we filter out saccade latencies lower than 100 ms, associated with predicting cases, and higher than 500 ms, corresponding to clear error cases, for further analysis. Saccades detected during one experiment session can be found in Fig. 1. Undershoot saccades are observed for closer eccentricities [2] while eye-tracker limitations can also be observed for higher eccentricity values.

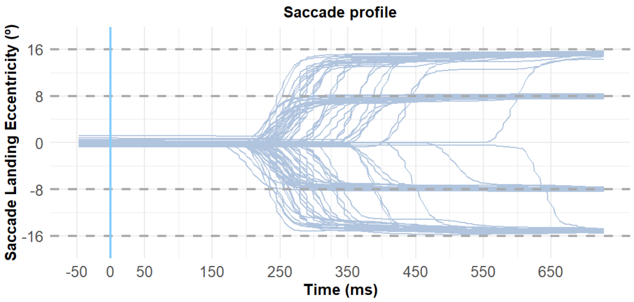

Figure 1: All saccades detected and measured in a single experiment session. At t=0 the visual target spawns, starting the latency measurement. Different saccade onset times are appreciated thus varying the final measured latency. Different landing eccentricities for each E condition are also observed, finding natural undershoot for a closer horizontal distance and also eye-tracker limitations for larger eccentricities.

S3. HARDWARE DETAILS

In our pipeline, two main error sources exist: visual stimulus latency (max 11.11 ms at 90 Hz) due to Unity's render to display delay, and eye-tracker logging latency (6-7 ms), from eye-tracker sampling (max 5 ms at 200 Hz) plus Unity processing. Combined, these contribute to a potential 18.11 ms error. Wired headphones introduce 1-10 ms latency, potentially creating a 10 ms visual-auditory mismatch. However, maximum delays are rare across trials, reducing their overall impact despite occasional occurrences. Highly specialized equipment like high-speed cameras and oscilloscopes would be needed for precisely tracking this error. Our GLMM analysis effectively manages random variations across subjects and trials. Although this specific error isn't explicitly modeled as a factor, it's included within the trial variability we account for, thus expected to have minimal impact on systematic bias in our results. These values were also confirmed by the headset manufacturer.

Table 1 shows the details of the HMD used for the experiment while Table 2 contains the details of the headphones.

Table 1: Specifications of the Varjo Aero HMD.

Display Resolution	2880 × 2720 pixels per eye
Display Refresh Rate	90 Hz
Eye Tracking Frequency	200 Hz
Eye Tracking Accuracy	Sub-Degree

Table 2: Specifications of the Beyerdynamic DT 770M headphones.

Frequency Response	5 Hz - 30000 Hz
Noise Cancellation Technology	Passive
Nominal Impedance	80 Ohm
Sensitivity	105 dB/mW

S4. ADDITIONAL DETAILS: MODEL, ANALYSIS AND APPLICATIONS

S4.1 Model Equation

The model equation can be found in Eq.1, with details on each factor meaning below.

$$Latency = (0.008) * E + (-0.174) * \Delta S + (-0.206) * \Delta T + (0.198) * E^2$$
$$+(-0.004) * E * \Delta S + (-0.006) * E * \Delta T + (-0.009) * \Delta S^2 + 249.197 \quad (1)$$

(1) Squared Terms
  (a) E^2: The positive coefficient (0.198) indicates that as Eccentricity increases, its effect on Latency increases in a quadratic manner.
  (b) ΔS^2 and ΔT^2: These terms have very small or zero coefficients, suggesting that their quadratic relationships with Latency are negligible.

(2) Linear Terms
  (a) E: The coefficient 0.008 suggests that a one-unit increase in Eccentricity is associated with an increase of 0.008 units in Latency, all else being equal.
  (b) ΔS: The coefficient -0.174 indicates that a one-unit increase in ΔS is associated with a decrease of 0.174 units in Latency.
  (c) ΔT: Similarly, the -0.206 coefficient for ΔT suggests that increasing ΔT by one unit decreases Latency by 0.206 units.

(3) Interaction Terms
  (a) $E - \Delta S$, $E - \Delta T$ and $\Delta S - \Delta T$: These coefficients represent how the interaction between these variables affects Latency. For example, a coefficient of -0.004 for Eccentricity Disparity implies a combined effect on Latency when both Eccentricity and Disparity vary together. However, the small magnitude of these coefficients suggests that these interactions have minimal impact on Latency.

(4) Intercept (249.197): This is the predicted value of Latency when all independent variables are 0.

In summary, the model suggests that Latency is primarily influenced by a significantly strong effect of Eccentricity (E^2), followed by linear terms of Eccentricity, Spatial shift (ΔS), and Temporal shift (ΔT).

Table 3: Statistical analysis of the results of the main experiment. Effect of the analyzed factors on saccadic latency.

Factor	t-value	p-value
ΔT	-4.032	5.64e-05
E	13.721	<2e-16
ΔS	-1.252	0.210479
E:ΔS	0.766	0.443796
ΔT:E	-3.891	0.000102
ΔT:ΔS	0.891	0.372795
ΔT:E:ΔS	-0.908	0.363684

Table 4: Statistical analysis of the results of the main experiment. Effect of the analyzed factors on saccade landing accuracy.

Factor	t-value	p-value
ΔT	-0.513	0.6082
E	8.198	3.27e-16
ΔS	-0.655	0.51224
E:ΔS	0.675	0.49957
ΔT:E	1.917	0.05525
ΔT:ΔS	2.150	0.05163
ΔT:E:ΔS	-2.121	0.05402

S4.2 Statistical Analysis

In this section, we include the tables corresponding to all pairwise comparisons in the statistical analysis for both the main experiment and the basketball application. In Tables 3 and 4 we show the results of the main experiment (latency and accuracy respectively). Following the results in these tables, we run post-hocs analysis for statistically significant factors. In Table 5 we show pairwise comparisons (post-hocs) for each temporal shift considered. In Table 6 we show pairwise comparisons (post-hocs) for the interaction of ΔT and eccentricity. We do not run post-hocs for the eccentricity factor, since it only has two levels. Finally, in Tables 7 and 8 we show the results of the basketball application case (latency and accuracy respectively), and in Table 9 the post-hocs analysis for the effect of ΔT on saccadic latency.

Table 5: Statistical analysis of the results of the main experiment. Post-hocs for the ΔT factor on saccadic latency.

Comparison	z-ratio	p-value
ΔT_0 - ΔT_100	8.826	<.0001
ΔT_0 - ΔT_200	15.725	<.0001
ΔT_0 - ΔT_300	19.665	<.0001
ΔT_0 - ΔT_400	21.027	<.0001
ΔT_100 - ΔT_200	6.654	<.0001
ΔT_100 - ΔT_300	10.507	<.0001
ΔT_100 - ΔT_400	12.007	<.0001
ΔT_200 - ΔT_300	3.897	0.0900
ΔT_200 - ΔT_400	5.524	0.0601
ΔT_300 - ΔT_400	1.692	0.4388

Table 6: Statistical analysis of the results of the main experiment. Post-hocs for the ΔT*Eccentricity interaction on saccadic latency.

Comparison	z-ratio	p-value
ΔT_0 E_8 - ΔT_100 E_8	5.658	<.0001
ΔT_0 E_8 - ΔT_200 E_8	9.569	<.0001
ΔT_0 E_8 - ΔT_300 E_8	12.014	<.0001
ΔT_0 E_8 - ΔT_400 E_8	12.689	<.0001
ΔT_100 E_8 - ΔT_200 E_8	3.850	0.046
ΔT_100 E_8 - ΔT_300 E_8	6.211	<.0001
ΔT_100 E_8 - ΔT_400 E_8	7.027	<.0001
ΔT_200 E_8 - ΔT_300 E_8	2.333	0.3675
ΔT_200 E_8 - ΔT_400 E_8	3.244	0.0691
ΔT_300 E_8 - ΔT_400 E_8	0.974	0.9937
ΔT_0 E_16 - ΔT_100 E_16	6.709	<.0001
ΔT_0 E_16 - ΔT_200 E_16	12.659	<.0001
ΔT_0 E_16 - ΔT_300 E_16	15.751	<.0001
ΔT_0 E_16 - ΔT_400 E_16	17.035	<.0001
ΔT_100 E_16 - ΔT_200 E_16	5.541	0.046
ΔT_100 E_16 - ΔT_300 E_16	8.605	<.0001
ΔT_100 E_16 - ΔT_400 E_16	9.924	<.0001
ΔT_200 E_16 - ΔT_300 E_16	3.169	0.0492
ΔT_200 E_16 - ΔT_400 E_16	4.565	0.0002
ΔT_300 E_16 - ΔT_400 E_16	1.415	0.9230

Table 7: Statistical analysis of the results of the basketball application. Effect of the analyzed factors on saccadic latency.

Factor	t-value	p-value
ΔT	-12.694	<2e-16
ΔS	-1.745	0.0812
ΔT:ΔS	0.182	0.8558

Table 8: Statistical analysis of the results of the basketball application. Effect of the analyzed factors on saccade landing accuracy.

Factor	t-value	p-value
ΔT	-0.107	0.9149
ΔS	-1.081	0.2799
ΔT:ΔS	0.761	0.4469

S4.3 Farm Game Results

Results obtained for each condition considered in the interactive farm game can be found in Table 10.

S5. USER STUDY SURVEY

In this section, questions and explanations presented to the participants are shown as well as the response options provided for both the demographic survey fulfilled before the experiment session and the sickness survey filled out twice, before and after doing the experiment.

Table 9: Statistical analysis of the results of the basketball application. Post-hocs for the ΔT factor on saccadic latency.

Factor	t-value	p-value
ΔT_0 - ΔT_100	5.680	<.0001
ΔT_0 - ΔT_200	10.181	<.0001
ΔT_0 - ΔT_300	13.104	<.0001
ΔT_0 - ΔT_400	13.737	<.0001
ΔT_100 - ΔT_200	4.112	0.0004
ΔT_100 - ΔT_300	7.102	<.0001
ΔT_100 - ΔT_400	8.027	<.0001
ΔT_200 - ΔT_300	3.177	0.0132
ΔT_200 - ΔT_400	4.300	0.0602
ΔT_300 - ΔT_400	1.241	0.7272

Table 10: Results of our farm game application. For each eccentricity, we test two target saccadic latency accelerations (no acceleration vs. 20, 40, 60 ms). Based on the desired saccadic latency, we estimate with our model the ΔT that needs to be applied. Then, we show the predicted latency by our model for those conditions and the measured latency (mean ± standard deviation) in the study. Predicted latency by our model is very close to that measured in all cases. The average error across conditions is 8 ms.

Eccentricity (°)	Target acceleration (ms)	Applied ΔT (ms)	Predicted latency (ms)	Measured latency (ms)
10	0	0	269	268 ± 10
10	20	75	249	242 ± 11
12	0	0	277	270 ± 9
12	40	145	237	222 ± 7
14	0	0	288	281 ± 9
14	60	135	228	214 ± 4

S5.1 Demographic Questionnaire

E1. How does the test work?

Once the trial starts, audiovisual stimuli will be presented to you under different conditions. A simple flash (white circle shape) is used as a visual target while the auditory cue is provided by a beep of 880Hz. Your task always starts by fixating on the center of the screen where another flash is visible, we refer to it as the fixation point. Once you are looking at it, you need to press the space bar on the keyboard to start the next case (it has to be pressed also to generate the first case). After pressing it, the audiovisual stimuli will be generated and the fixation point disappears. As soon as you detect or perceive the visual target, you have to look at it, so perform the saccade from the fixation point to the visual target. After a while, the visual target disappears and the fixation point spawns again, returning to the initial situation. After a certain number of cases, the session ends. The experiment finishes after 4 sessions. If you have any questions, please ask the experimenter now.

☐ *Understood*

E2. Consent for participation in the study

I agree to participate in the research study. I understand the purpose and nature of this study and I am participating voluntarily. I understand that I can withdraw from the study at any time, without any penalty or consequences. I grant permission for the data generated from this questionnaire to be used in the researcher's publications on this topic. The generated data will be stored anonymously under a randomly generated unique ID. Any information that is obtained in connection with this study and that may be identified with you will remain confidential and will be disclosed only with your permission.

☐ *I agree*

Q1. Subject anonymous ID

____

Q2. Age

____

Q3. Gender

☐ *Male* ☐ *Female* ☐ *Rather not to say* ☐ *Other*

Q4. Home Country

____

Q5. Do you have any visual impairments

☐ *Yes* ☐ *No*

Q6. If you answered "Yes" to the previous question, please specify your condition (e.g. poor distance vision):

____

Q7. If you have any visual impairments, do you have it corrected?

☐ *Yes* ☐ *No*

Q8. If you answered "Yes" to the previous question, please specify how you have it corrected (e.g. glasses):

____

Q9. Do you have any auditory impairments

☐ *Yes* ☐ *No*

Q10. If you answered "Yes" to the previous question, please specify your condition (e.g. age-related hearing loss):

____

Q11. If you have any auditory impairments, do you have it corrected?

☐ *Yes* ☐ *No*

Q12. If you answered "Yes" to the previous question, please specify how you have it corrected (e.g. ear-mounted device):

____

Q13. Do you play videogames

☐ *No* ☐ *Yes, Sporadically* ☐ *Yes, often* ☐ *Yes, everyday*

Q14. Specify your experience with Virtual Reality

☐ *None, I have never used a virtual reality device*
☐ *Basic, I have used virtual reality devices less than 5 times*
☐ *Experienced, I have used virtual reality devices several times*
☐ *Professional, I use virtual reality devices on a daily basis*

Q15. If you have already tried virtual reality, please specify those that apply:

☐ *I have tried desktop-based devices like Oculus, HTC Vive, or Play Station VR*
☐ *I have tried smartphone-based devices like Google Cardboard*
☐ *I use virtual reality devices everyday*
☐ *I suffered fatigue, dizziness or eyestrain when using virtual reality devices*

S5.2 Sickness Questionnaire

Q1. Subject anonymous ID

____

Q2. Session

☐ *Before* ☐ *After*

Q3. According to your current condition, indicate the degree of the following symptoms:

Symptoms	None	Mild	Moderate	Severe
General Discomfort	☐	☐	☐	☐
Headache	☐	☐	☐	☐
Eyestrain	☐	☐	☐	☐
Dizziness or nausea	☐	☐	☐	☐

REFERENCES

[1] Elena Arabadzhiyska, Okan Tarhan Tursun, Karol Myszkowski, Hans-Peter Seidel, and Piotr Didyk. 2017. Saccade landing position prediction for gaze-contingent rendering. *ACM Transactions on Graphics (TOG)* 36, 4 (2017), 1–12.

[2] Wolfgang Becker and Albert F Fuchs. 1969. Further properties of the human saccadic system: eye movements and correction saccades with and without visual fixation points. *Vision research* 9, 10 (1969), 1247–1258.

[3] D Boghen, BT Troost, RB Daroff, LF Dell'Osso, and JE Birkett. 1974. Velocity characteristics of normal human saccades. *Investigative Ophthalmology & Visual Science* 13, 8 (1974), 619–623.

[4] Maarten A Frens, A John Van Opstal, and Robert F Van der Willigen. 1995. Spatial and temporal factors determine auditory-visual interactions in human saccadic eye movements. *Perception & psychophysics* 57 (1995), 802–816.

[5] Miguel A García-Pérez and Rocío Alcalá-Quintana. 2015. Converging evidence that common timing processes underlie temporal-order and simultaneity judgments: A model-based analysis. *Attention, Perception, & Psychophysics* 77 (2015), 1750–1766.

[6] Torin Hopkins, Suibi Che-Chuan Weng, Rishi Vanukuru, Emma Wenzel, Amy Banic, Mark D Gross, and Ellen Yi-Luen Do. 2022. Studying the Effects of Network Latency on Audio-Visual Perception During an AR Musical Task. In *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 26–34.

[7] Hayeon Kim and In-Kwon Lee. 2022. Studying the effects of congruence of auditory and visual stimuli on virtual reality experiences. *IEEE Transactions on Visualization and Computer Graphics* 28, 5 (2022), 2080–2090.

[8] Albert R Powers, Andrea R Hillock, and Mark T Wallace. 2009. Perceptual training narrows the temporal window of multisensory binding. *Journal of Neuroscience* 29, 39 (2009), 12265–12274.

[9] Susan M Ross and Leonard E Ross. 1981. Saccade latency and warning signals: effects of auditory and visual stimulus onset and offset. *Perception & Psychophysics* 29 (1981), 429–437.

[10] Ladan Shams, Yukiyasu Kamitani, and Shinsuke Shimojo. 2000. What you see is what you hear. *Nature* 408, 6814 (2000), 788–788.

[11] Barry E Stein and Terrence R Stanford. 2008. Multisensory integration: current issues from the perspective of the single neuron. *Nature reviews neuroscience* 9, 4 (2008), 255–266.

[12] Ryan A Stevenson and Mark T Wallace. 2013. Multisensory temporal integration: task and stimulus dependencies. *Experimental brain research* 227 (2013), 249–261.

[13] Ryan A Stevenson, Raquel K Zemtsov, and Mark T Wallace. 2012. Individual differences in the multisensory temporal binding window predict susceptibility to audiovisual illusions. *Journal of Experimental Psychology: Human Perception and Performance* 38, 6 (2012), 1517.

[14] Erik Van der Burg, David Alais, and John Cass. 2013. Rapid recalibration to audiovisual asynchrony. *Journal of neuroscience* 33, 37 (2013), 14633–14637.

[15] Argiro Vatakis and Charles Spence. 2006. Audiovisual synchrony perception for music, speech, and object actions. *Brain research* 1111, 1 (2006), 134–142.

[16] Manuel Vidal and Françoise Vitu. 2022. Multisensory temporal binding induces an illusory gap/overlap that reduces the expected audiovisual interactions on saccades but not manual responses. *Plos one* 17, 4 (2022), e0266468.