Perceived Asynchrony of Rhythmic Stimuli Affects Pupil Diameter and Smooth Pursuit Eye Movements

Lina Klass
lina.klass@uni-weimar.de
Bauhaus-Universität Weimar
Weimar, Thuringia, Germany

Anton Benjamin Lammert
anton.benjamin.lammert@uni-weimar.de
Bauhaus-Universität Weimar
Weimar, Thuringia, Germany

Laura Simon
laura.simon@uni-weimar.de
Bauhaus-Universität Weimar
Weimar, Thuringia, Germany

Eva Hornecker
eva.hornecker@uni-weimar.de
Bauhaus-Universität Weimar
Weimar, Thuringia, Germany

Bernd Froehlich
bernd.froehlich@uni-weimar.de
Bauhaus-Universität Weimar
Weimar, Thuringia, Germany

Jan Ehlers
jan.ehlers@uni-weimar.de
Bauhaus-Universität Weimar
Weimar, Thuringia, Germany

Abstract

In networked applications, latency can disrupt the sense of synchrony by causing offsets e.g. between local speech and remote visual response. We investigate the influence of frequency and Stimulus Onset Asynchrony (SOA) on synchrony perception during rhythmic audiovisual experiences. Our results show that the Point of Subjective Synchrony (PSS) is influenced by frequency, whereas the Window of Subjective Synchrony (WSS) is not. Variations in SOA induce adaptive gaze behavior in response to audiovisual latencies, while pupil diameter increases with increasing SOA, suggesting a higher cognitive load for successive unisensory rather than integrated events. This has practical implications for the design of computer-mediated applications that promote a sense of community through rhythmic interaction. Eye tracking data may indicate perceived (a)synchrony in audiovisual integration. In addition, the choice of frequencies may help to mask latencies, enhance the experience of synchrony and thus support feelings of closeness and intimacy in virtual interaction.

CCS Concepts

• **Human-centered computing** → *Empirical studies in HCI*; • **Applied computing** → **Psychology**.

Keywords

Eye movement, eye tracking, smooth pursuit, gaze, pupillometry, synchrony

ACM Reference Format:
Lina Klass, Anton Benjamin Lammert, Laura Simon, Eva Hornecker, Bernd Froehlich, and Jan Ehlers. 2025. Perceived Asynchrony of Rhythmic Stimuli Affects Pupil Diameter and Smooth Pursuit Eye Movements. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26-May 1, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3706598.3713152

CHI '25, Yokohama, Japan
© 2025 Copyright held by the owner/author(s).
This is the author's version of the work. It is posted here for your personal use. Not for redistribution. The definitive Version of Record was published in *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26-May 1, 2025, Yokohama, Japan*, https://doi.org/10.1145/3706598.3713152.

1 Introduction

Latencies are inevitable in computer-mediated interactions, arising from factors like network transmission delays, processing times, and hardware limitations. They can become particularly problematic when remote users attempt to rhythmically synchronize with each other. Even seemingly simple tasks like synchronous clapping during a video conference often result in disjointed, asynchronous experiences due to latency. Similar challenges have been noted in remote music collaborations, with musicians emphasizing that "latency is an enormous issue" [8]. Likewise, many virtual reality (VR) applications like multi-user virtual dancing, rely on precise synchronization of full-body movements, often leading participants to adjust their timing to compensate for delays [52]. This work investigates how latency impacts synchrony perception in rhythmic audiovisual stimuli, linking basic mechanisms of audiovisual integration to the individual experience of synchrony.

Our experiment explores the effect of stimulus onset asynchrony (SOA) on the point of subjective synchrony (PSS) and the window of subjective synchrony (WSS) in a multisensory decision-making task. We vary the frequency and temporal proximity of rhythmic auditory and visual events to correlate synchrony judgements with implicit measures of cognitive load (pupil diameter) and visual processing accuracy (smooth pursuit eye movements). The aim is to understand how large a temporal offset can be, before we experience unisensory rather than integrated events and how this perception is modulated by different rhythms. In addition, we are investigating whether it is possible to identify eye-based markers that allow automated detection of synchrony violations and thus breaks in co-presence.

This investigation into the perceptual mechanisms underlying feelings of synchrony is motivated by the critical role that synchronized behavior plays in fostering a sense of community [21, 23, 30, 33, 44]. High levels of asynchrony disrupt the smooth flow of events and the emergence of intimacy, closeness, and group cohesion [21, 44]. In contrast, synchrony has been shown to promote likeability [33], trust [1], and collective connection [17].

There are various technical approaches to reduce or mask network latencies, such as reducing image resolution or frame rates [34], employing compression algorithms [29], or client-side prediction [38]. An essential research question relates to the threshold at which synchrony perception is disrupted. Understanding this is crucial in order to apply such techniques dynamically and avoid unnecessary quality trade-offs.

Previous work has shown that the perception of synchrony can vary significantly, depending on the context and type of stimuli involved [12]. However, there is almost no research on *rhythmic* audiovisual stimuli, despite their relevance in remote interactions such as collaborative music performances, virtual dance rehearsals, online fitness classes and multiplayer rhythm games. This raises the question of which factors specifically affect the synchrony perception of rhythmic audiovisual stimuli.

Our results show that both the frequency of the rhythmic stimulus and the order of latency-affected stimuli (audio-first vs. video-first) significantly influence the point of subjective synchrony (PSS). Additionally, eye tracking measures indicate that stimulus frequency and latency impact cognitive load. These findings have practical implications for the design of rhythmic networked interactions: eye tracking data can serve as physiological indicators of perceived (a)synchrony, while a careful choice of frequencies could help to mask latencies, enhance the experience of synchrony and thus promote a sense of closeness in remote computer-mediated interactions.

2 Related Work

In the following we discuss related work on the impact of latency and synchrony in computer-mediated interactions, synchrony judgement of non-rhythmic and rhythmic audiovisual stimuli, as well as on eye movements and synchrony judgements.

2.1 Impact of Latency and Synchrony in Computer-Mediated Interactions

Synchronized activities, such as singing and dancing, can foster a sense of connection and belonging through *entrainment*, a process where individuals naturally and unconsciously align their rhythms [44]. Previous research has shown that such activities can promote intimacy, closeness, and empathy [23, 30, 33]. Pro-social effects of interpersonal motor synchrony (IMS) have recently been explored in the field of Human-Computer Interaction (HCI) across various contexts [1, 26, 33, 47, 48, 60, 67]. For instance, Robinson et al. present *In the Same Boat* – a canoeing computer game – and demonstrate that synchrony as a game mechanic enhances the sense of closeness between players, especially if physiological or embodied controls such as breath rate or facial expressions are used instead of standard keyboard controls [60]. A similar but more immersive experience is the VR installation *JeL*, which fosters connection with others and nature through synchronized breathing, visualized as biofeedback-driven animations of jellyfish and corals [67]. Other examples include *Yamove!*, a real-time in-person dance battle game [26], *NeuralDrum*, an XR drumming experience providing feedback on brain synchrony [47], or *ExerSync*, a system that uses audio-visual cues to synchronize the users' movements in diverse exercises such as rope jumping, cycling or running [48].

Despite this growing interest, further research is needed to fully understand IMS, and HCI applications that leverage synchronized motor activities between (remote) users should build upon these insights. As Rinott et al. emphasize, "IMS can be designed, and should be designed: the knowledge created in the laboratories can serve to increase the prevalence and sophistication of such experiences created deliberately." [59]. In line with this perspective, the goal of this paper is to deepen the understanding of IMS and its prerequisites through a controlled laboratory experiment.

The first step toward understanding synchronization is timing: Achieving interpersonal motor synchronization requires aligning one's own actions with the predicted movements of another user, which demands a high degree of temporal precision [64]. However, in the context of long-distance interactions, achieving such accuracy is challenging due to latency caused by network transmission delays, processing times, and hardware limitations. This is vividly illustrated by the examples above: games specifically designed to foster connection between remote users consistently highlight latency as a significant problem [48, 60], e.g. "Most participants [. . .] mentioned the delay in the game due to network latency was challenging" [60].

Of course, transmitted networked data (e.g. audio and video data) can be buffered so that it can be presented in a synchronous manner to a remote client. Nevertheless, interactions can still be perceived as asynchronous, e.g. a delayed visual or acoustic response from a remote user to a stimulus from a local user. In mediated audiovisual interactions, latency can result in mismatches where the audio lags behind the video, e.g. the sound of a remote orchestra lags behind the movements of a local conductor [61]. It is also possible for the video to lag behind the audio, for example the movements of a remote dancer to lag behind the performance of a local musician or behind a globally synchronized music playback. Another practical example of such latency challenges is the *ExerSync* system mentioned earlier, which uses audio-visual cues derived from the rhythm of the leading user to promote interpersonal synchrony. In this system, latency between the leader's movements and the audio-visual cues perceived by remote users is inevitable due to network transmission delays [48].

Typical latencies vary depending on the application and the infrastructure. In VR, network latencies below 15 ms are ideal for collaboration [13], while for videoconferencing, latencies below 150 ms have been considered acceptable [79]. Various attempts have been made to develop recommendations for acceptable thresholds in HCI applications. However, it has been shown that the perception of latency can vary significantly depending on the task [65], the sensory modalities [28], or the content type (e.g. music vs. speech) [12]. Understanding these different user perceptions is crucial, as latency has a direct impact on the overall user experience. As van Damme et al. stress: "The existing literature [...] has mostly focused on network requirements from a system point-of-view, where the key performance parameters are only provided in the form of Quality-of-Service (QoS) parameters (such as end-to-end latency). However, the translation of these network impairments to the end-user experience is often omitted." [72].

Latency in computer-mediated interactions can disrupt the communication flow, cause confusion, lead to frustration, hinder decision making, and reduce spontaneity [46, 62]. Over time, it can erode trust and rapport, making interactions feel disjointed [24]. Nonverbal cues such as gestures and facial expressions are particularly affected, as latency disrupts synchrony, leading to misunderstandings and reduced emotional engagement [24, 46, 62]. Latency also significantly impacts collaborative activities such as music-making and dance, disrupting synchronization in joint actions [9], for instance with negative effects on online music education [61] and in performance arts [25]. The referenced studies investigate real-world conditions; therefore, we conclude that the observed negative effects occur across typical latency levels in various contexts (as outlined in the previous paragraph) and are not confined to extreme delays, such as those associated with poor internet connections.

The occurrence of latency cannot be fully prevented (as it is unavoidable in networked environments). While radical strategies to minimize latency could involve reducing frame rates or image resolution during latency peaks, the application of methodologies designed to enhance the perception of synchrony in asynchronous scenarios may present a potential approach to *mask* latency effects [59]. However, especially for methods that involve trade-offs in streaming quality, it is crucial to understand the thresholds of perceived synchrony – specifically, what levels of latency are still considered acceptable in certain scenarios [23, 57].

2.2 Synchrony Judgements of Non-Rhythmic Stimuli

Synchrony judgements (SJ), where participants rate events as synchronous or asynchronous, are widely used to study synchrony perception [12, 16, 74, 78]. In this context synchronous stimuli are often offset in time and the influence of this Stimulus Onset Asynchrony (SOA) is used to identify when participants perceive different stimuli to be in sync [5, 68, 73]. The Point of Subjective Synchrony (PSS), the SOA perceived as most synchronous, varies across individuals but was found to remain stable for each individual [68]. It has been shown that for audiovisual stimuli it also depends on the event type and duration [12] and is generally shifted toward audio lags [39, 74]. In addition, temporal recalibration plays a significant role in SJ, as prior exposure to audiovisual delays influences perception, shifting the PSS toward prior experienced offsets. Research shows this shift can be as large as 30 ms, effectively recalibrating synchrony perception based on previous audiovisual experiences [16, 75]. While the PSS refers to a specific SOA, the Window of Subjective Synchrony (WSS) or Temporal Integration Window (TIW) is often used to analyse the range of SOAs within which participants perceive synchrony [39, 78].

2.3 Synchrony Judgements of Rhythmic Stimuli

Research on the emergence of intimacy during *rhythmic* (multisensory) processing in computer-mediated interactions remains scarce, with the interplay between stimulus frequency and temporal asynchrony still not fully understood. Understanding this relationship is essential, as many applications rely on rhythmic contexts. Examples include music education [61], remote orchestras [8], virtual conducting [9], online dance collaborations [25, 52], and rhythm-based virtual reality games [31], all of which depend on precise temporal coordination. Identifying specific frequencies that are particularly sensitive to synchrony would enable designers to adapt their systems accordingly, for example by avoiding certain frequency ranges or by prioritizing network latency optimization in scenarios where such frequencies are prevalent.

As previously suggested by Hopkins et al. [22], it can be hypothesized that for rhythmic stimuli, as soon as latency exceeds the frequency's period, users will not be able to distinguish which cycle the stimulus corresponds to. For instance, in a 2 Hz rhythmic interaction (with a period of 500 ms), a shift of 600 ms, might be perceived identically to a 100 ms shift because of the rhythmic nature of the stimulus.

For rhythmic audiovisual stimuli, matching stimuli becomes unreliable once the frequency exceeds 4 Hz [5]. This suggests that in the context of networked experiences, frequencies below 4 Hz may be advantageous to enable synchronous experiences. In contrast, Wojtczak et al. [77] found that audio stimulus' pitch does not influence synchrony perception. For moving visual stimuli, it was found that SJ is primarily influenced by the peak velocity within the trajectory, even if participants are instructed to focus on position cues [58, 69]. Furthermore, Heins et al. [19] found that intentional sound production, such as in tap dancing, creates a broader TIW compared to incidental sound production, and that higher event density increases synchrony ratings, even when audiovisual signals are not perfectly aligned. These findings indicate that higher frequencies might facilitate synchronous experiences in audiovisual mediated interactions. However, to our knowledge, no systematic investigation has yet addressed the influence of specific frequencies on perception of synchrony.

2.4 Pupillometry and SPEM

Pupillometry serves as a non-invasive, sensitive, and reliable indicator of various psychological, cognitive, and physiological states. In general, the pupil takes about 2 seconds to dilate and return to baseline in response to a single, brief stimulus [20]. Changes in pupil diameter can reflect underlying cognitive, emotional, and physiological processes, making pupillometry a valuable tool for gaining insights into brain function and psychological states [3, 18]. This is particularly beneficial for states that may not even be consciously perceived or where explicit reporting is not possible for the participant [14].

For example, pupil dilation is frequently utilized as an indicator of mental effort in cognitive research [71], but it can also reflect related psychological states, such as surprise, cue uncertainty, or violation of expectations [4, 36, 54]. The pupil also indicates entrainment to rhythmic patterns [76], and demonstrates dilation in response to increasing speed, as shown in studies by Fink et. al [14]. This suggests that the pupil not only synchronizes with rhythmic stimuli, but also adjusts its dilation dynamically to accommodate variations in stimulus speed, reflecting its sensitivity to changes in the temporal structure of sensory input.

Dynamic Attending Theory (DAT) can provide further insight into the behavior of the pupil in relation to synchrony detection. DAT, as proposed by Large and Jones [32], assumes that attention is aligned with the rhythmic patterns of external stimuli, such as music or speech. In terms of synchrony judgement, our brain uses internal oscillations to predict when events will happen, allowing us to synchronize with them. If two sounds, e.g. drums and bass, are slightly out of sync (due to microtiming asynchronies), the brain detects these deviations by widening its attentional focus to encompass both sounds [11]. This adjustment helps maintain synchronization [27] but requires additional cognitive effort to manage timing differences, as shown by Skaansar et al. [66]: They examine asynchronies between bass and drum, demonstrating that due to higher cognitive load the pupil size increases during higher SOAs. While they concentrate on the perception of music, in this paper we explore the influence of audio-visual stimuli, which are usually central in computer-mediated interactions. Skaansar et al. also do not investigate the perception of synchrony and instead focus on "groove", a construct describing the feeling of wanting to move along to music. Further, their study does not examine the effects of different frequencies on synchrony judgements or on eye movement data and does not address smooth pursuit eye movements as potential indicators in this context.

Smooth Pursuit Eye Movements (SPEM) allow us to smoothly track small moving objects. Early findings by Lisberger et al. [37] show that for higher frequencies the lag of eye position behind the target increases but that this lag is "independent of the amplitude of the movement" [37]. In the context of synchrony judgement, SPEM play a role in following and predicting rhythmic patterns, as discussed by Masson and Stone [41]. SPEM helps align visual attention with external stimuli, aiding the brain in anticipating when events, like beats in music, will occur. This anticipatory mechanism, as outlined by Schütz et al. [63], supports the brain's ability to assess synchrony, such as between visual and auditory cues, and determine if they are in sync.

2.5 Summary

While synchrony perception in the context of non-rhythmic audiovisual stimuli is well studied, research on rhythmic audiovisual stimuli is limited so far. In particular, there is, to our knowledge, no research on the relation between stimulus frequency, synchrony judgements and pupillometry. Our investigation bridges this gap, providing insights on how frequency influences synchrony judgements and how both actual and perceived asynchronies influence pupillometry, enabling a better understanding of synchronous rhythmic experiences. Our findings contribute to research on enhancing user experiences in HCI, especially in scenarios where precise timing and synchrony are critical and different latencies can be present, such as in collaborative virtual environments and real-time communication tools.

3 Method

We conducted a controlled laboratory experiment to simulate latency in rhythmic audiovisual stimuli, as commonly encountered in computer-mediated contexts. This was achieved by systematically manipulating the stimulus onset asynchrony (SOA). We investigate 1) how the perceptual integration of rhythmic stimuli is influenced by SOA and frequency, 2) how this is related to individual perception of synchrony and 3) whether eye tracking data can serve as an indicator of perceived audiovisual (a)synchrony.

3.1 Experimental Procedure

The experiment took place in a room with an ambient illumination of approx. 200 lux. Participants were seated upright at a table with a computer monitor approximately 65 cm away. A chin rest was used to minimize head movements and was adjusted to a comfortable height that also allowed valid calibration of the eye tracker. Prior to the start of the experiment, participants completed a short questionnaire collecting demographic information and details of any visual or hearing impairments.

The experiment consisted of two blocks (corresponding to two visual compositions, cf **Section 3.2**), each lasting approx. 15 to 20 minutes. There was a 10-minute break in between. Prior to the start of each block, participants completed a five-point eye tracker calibration. To ensure accurate control of the audiovisual stimulation, the system was calibrated using the synchronisation tool SyncOne2[1]. Before the start of the experiment, a short test (ten trials) was administered to ensure that the instructions were understood.

During task processing, participants followed a 15 x 15 mm fixation cross (approx. 0.9 degrees of visual angle) with their eyes as it moved horizontally between two vertical boundary lines, cf. **Figure 1**. Each time the cross reached one of the boundaries, a tone was played. The auditory stimulus was a 250 Hz square wave (cf. [68]), played for 33 milliseconds (cf. [39]) at 60 dB. The tone was reproduced using wired over-ear headphones to avoid the effects of ambient noise or variations in speaker placement. Depending on the experimental condition, the tone either slightly preceded or followed the cross's contact with the line; in other trials, both events occurred simultaneously.

Participants were instructed to rate the synchrony of the audiovisual event by entering their ratings on a four-point scale using the keys 1, 2, 3 and 4 on the keyboard. All four keys were color coded red, light red, light green, and green respectively, and additionally marked with - -, -, + and ++, representing the rating labels "very asynchronous", "rather asynchronous", "rather synchronous", and "very synchronous". The keys next to the designated response keys had been removed to prevent accidental key presses. The synchrony judgements could be made at any time, but there was a minimum processing time of four seconds to ensure that a sufficient amount of smooth pursuit eye tracking data was collected in each trial. In case of earlier key presses, the fixation cross turned blue to indicate that the entry was recorded. After the individual judgement was made and the minimum duration had elapsed, an intertrial screen was displayed for two seconds. The complete experiment took approximately one hour.

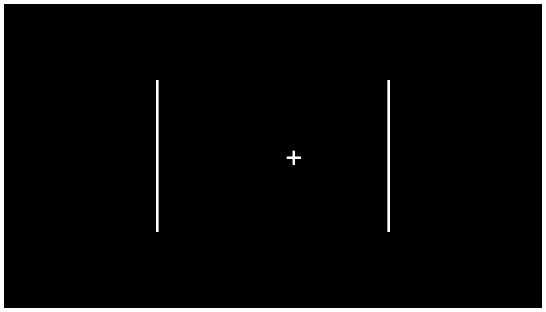


Figure 1: Visual layout of the audiovisual task during the condition of 0.8 units distance between the boundary lines.

Table 1: Speed and distance according to the visual composition during all experimental conditions. Note, that values are given in normalised window units (-1 to 1) (per second) and degrees (per second).

	0.5 Hz	1 Hz	2 Hz
constant distance	velocity: 0.4 (9 deg/s) distance: 0.8 (18 deg)	velocity: 0.8 (18 deg/s) distance: 0.8 (18 deg)	velocity: 1.6 (36 deg/s) distance: 0.8 (18 deg)
constant speed	velocity: 0.8 (18 deg/s) distance: 1.6 (36 deg)	velocity: 0.8 (18 deg/s) distance: 0.8 (18 deg)	velocity: 0.8 (18 deg/s) distance: 0.4 (9 deg)

3.2 Experimental Design

We varied stimulus onset, stimulus frequency and the visual composition in a simple audiovisual judgement task. Explicit (subjective reports) and implicit (pupil size changes, SPEM) measures were applied to quantify the individual experiences. Selection of SOAs was based on previous findings (cf. [39, 68]) and covered a range from -250 ms to +250 ms. According to Benjamins et al. [5], synchrony perception becomes unreliable for audiovisual stimulus processing above 4 Hz. We included 2 Hz since it corresponds to common rhythms in music [40], and added 1 Hz and 0.5 Hz as lower subdivisions. However, in the 2 Hz condition, an SOA of ±250 ms would play the tone exactly when the cross was in the middle of the boundary lines and it would be no longer possible to distinguish between +250 ms and -250 ms asynchrony. As a result, this combination of frequency and SOA was excluded.

Fink et al. [14] report pupil diameter to be affected by stimulus velocity. Accordingly, we incorporated two different concepts for the visual composition: Constant speed (movement distance is varied) and constant distance (speed depends on the frequency) for which values are shown in **Table 1**. Movement acceleration remained stable across all conditions as peak velocity has been shown to influence synchrony judgements [69].

The variation of SOA, frequency and visual composition resulted in an 11 x 3 x 2 within-subject design. The factor SOA included the subdivisions ±250, ±200, ±150, ±100, ±50, and 0 ms (perfect synchronization). Negative SOAs indicate that the visual event (cross meets boundary line) was preceded by the tone.

SOA values simulate the total round trip time required for the first presented stimulus to be transmitted to the remote partner and the response stimulus to be transmitted back. The simulated latency can therefore be interpreted as SOA / 2, which results in a simulated latency range of 0 ms to 125 ms. This range corresponds to typically occurring or recommended latencies, as described in **Section 2.1**. By accounting for both negative and positive SOA values, this encompasses scenarios where either the video precedes the audio (e.g. audio response from a remote orchestra to the visual movement of a local conductor) or vice versa (e.g. a remote partner dancing to a song created by a local partner).

As depicted above, we covered the frequencies 0.5, 1 and 2 HZ, indicating the time the cross moves between the two boundaries. As the visual composition was identical for 1 Hz in both the constant distance and constant speed condition (cf. **Table 1**), only one configuration was used.

To avoid sequence effects, the variation in frequency and visual composition was counterbalanced across participants using a Latin square design [6], while the SOA was randomized and presented five times per condition. The exclusion of one participant resulted in a slightly unbalanced design. As dependent variables we recorded behavioral data, i.e. synchrony judgements on a four-step single item rating scale and the time until synchrony judgement. To determine cognitive workload and visual processing accuracy, we used cognitive pupillometry and recorded smooth pursuit eye movements.

3.3 Participants

24 volunteers (13 female, 9 male, 2 divers / not disclosed; mean age: 26 years (SD: 3)), all (international) students or researchers at Bauhaus-Universität Weimar participated in the experiment. Hearing impairments were not reported. For optimum eye tracking results, visual impairments were corrected to normal via contact lenses instead of glasses. Information on regular medication was not collected. All participants received a 10€ compensation. One participant was excluded from further analysis due to difficulties during task processing. This resulted in $N = 23$ participants. Written informed consent was obtained prior to the start. All measurements were performed in accordance with the Declaration of Helsinki and were approved by the ethics committee of Technische Universität Ilmenau.

3.4 Apparatus & Software

We used a 360 Hz, 27” (2560 x 1440) monitor (ASUS ROG Swift PG27AQN) connected to a computer with no unnecessary programs running in the background to avoid any overhead. A frame rate of 160 fps was chosen; this allowed all SOAs to be achieved accurately, as their values are multiples of the resulting frame duration of 6.25 milliseconds. The audio stimulus was presented through wired headphones (RAZER Kraken V3 X over-ear). A Tobii Pro Nano eye tracker with a sampling rate of 60 Hz was used to record pupil size changes and smooth pursuit eye movements. The experimental software PsychoPy [49] was used to set up the experiment as it meets the requirements of providing adequate audio-video synchronization with a sync variance of approximately 0.93 milliseconds [7].

A low latency jitter is particularly important for the current experiment as –unlike a constant latency– it cannot be corrected by calibration. Triple buffering was disabled on the experiment PC because Psychopy expects a double-buffered rendering pipeline. For sound presentation, we used the ptb library which is the recommended setting for high-precision audio-timing [50]. Audio latency

[1] https://sync-one2.harkwood.co.uk/

priority was set to critical. The current PsychoPy implementation is close to the one described by Bridges et al. [7], but adapted to our stimuli. For example, we uploaded a custom wave file to be able to use a 250 Hz square wave audio (cf. [68]), but buffered it to ensure low latency. Unlike Bridges et al., we implemented the audio presentation using a code component rather than the GUI element, otherwise the duration of the audio signal would not be precise enough, resulting in glitches or short audios (about 70 ms) sometimes not being played at all. As recommended by PsychoPy, we used a frame-based implementation to achieve precise timing. [51].

3.5 Data Processing

The current subsection outlines our analysis process, including information on the pre-processing of the eye tracking data and on the statistical analysis.

3.5.1 Pre-Processing of Eye Tracking Data. Since the presentation of the audio stimulus was latency corrected (cf. **Section 3.4**), we applied the same correction during the pre-processing of the eye data to determine the exact time at which the tone was played.

Pupil data was blink-reconstructed (using the python datamatrix library with default values [43]) and baseline corrected (subtractive baseline correction). The baseline period was defined as the time frame before the onset of the first audio stimulus plus the minimum latency of the pupil response (200 ms) [42]. This period was excluded from the pupil data series. Trials in which baselines could not be calculated due to invalid values (n = 27, 0.46%) were excluded, as were trials with extreme baselines, i.e. with absolute z-scores greater than 2.0 at participant level [42] (n = 213, 3.63%). If a trial had no valid eye tracking values, both the mean pupil size and the mean gaze stimulus distance were also marked as invalid (n = 6, 0.10%). In total, 246 trials (4.19%) were excluded from further analysis of the eye tracking data, leaving 5619 valid trials.

Gaze position data were trimmed to exclude recordings prior to the first audio stimulus because, similar to pupil responses, SPEM cannot be attributed to perceived asynchrony before the asynchrony has occurred. We calculated the distance between the mapped gaze samples and the position of the moving stimulus, with negative values indicating that the gaze was behind the stimulus and positive values indicating that it was ahead of the stimulus. Outliers were excluded using the median absolute deviation (MAD) with a rather conservative threshold of 3 MAD [35]. This resulted in 117 trials that were additionally excluded from the SPEM data.

3.5.2 Statistical analysis. Except stated otherwise, we ran linear mixed models (LMMs) in R (version 2024.04.2) [55] using the lme4 package (version 1.1-35.5) [2] for the main analysis. In all models, we added participants as a random effect. Maximum Likelihood Estimation (ML) was used for model fitting and comparison, and a restricted Maximum Likelihood approach (REML) for final model estimations and to calculate inter-class correlation (ICC). To test the assumptions of linearity, homoscedasticity, as well as independence of residuals we used residual plots and ACF plots. The normal distribution of synchrony judgements and eye tracking data was evaluated graphically using Q-Q plots and histograms. It has been shown that ordinal data can reliably approximate a continuous variable [45, 70]. Therefore, we treated synchrony judgements as continuous for running LMMs, despite the limitation that it was measured on a 4-step scale. A significance level of 0.05 was used to determine the statistical significance of the following results.

4 Results

The current section presents the results on subjective reports (PSS, WSS), behavioral responses (decision time) and oculomotor responses (pupil diameter, SPEM).

4.1 Synchrony Judgement, Pupil Diameter and Gaze-Target Distance

Synchrony judgements were based on a 4-point single-item Likert scale. The distance in the x-direction between mapped gaze sample and the target position was applied to quantify the ocular pursuit (degrees of visual angle, negative values indicate gaze lag). Decision times (time between task onset and motor response) were recorded for each trial. **Figure 2** shows the mean values of the dependent variables; SOAs are depicted separately for each frequency. Note that, as mentioned above, the ±250 ms SOA condition is not available at a frequency of 2 Hz.

We used separate LMMs to analyze each of the three primary dependent variables. Starting with the null model, we first added the frequency condition as a predictor. Given that visual inspection of our data suggests a turning point around 0 ms SOA for all three dependent variables, we introduced soa_negative (if TRUE, the audio stimulus was played before the fixation cross reached the border line) and absolute SOA as separate predictors. The third independent variable 'visual composition' was added as 'distance', specifying the distance between the two border lines (cf. **Table 1**). We chose to add the independent variables in this form as this seemed meaningful for later interpretation.

To improve interpretability of the resulting coefficients, predictors were neither centered nor standardized. Thus, we prioritized preserving the coefficients' original units rather than identifying the variable with the most significant relative effect, as our focus was on understanding each variable's individual influence (e.g. what change in pupil dilation can be expected in response to strongly perceived asynchrony) rather than its relative importance.

We inspected whether incrementally adding each of the predictors improved the models in terms of AIC (Akaike Information Criterion). This was the case for all three models. After establishing the model with main effects we experimented with including interaction terms among these predictors: We tested whether adding an interaction between soa_negative and absolute_SOA improved the fit of the synchrony judgement model. Given that previous research and our data indicate that the PSS is slightly shifted towards positive asynchronies [68, 74], incorporating this interaction might better capture how the SOA effect varies with the order of presentation. We did not include any other interaction terms in our models, as this would likely lead to overfitting. Additionally, AIC and BIC (Bayesian Information Criterion) indicated that including additional interaction terms would not have improved our models much while adding unnecessary complexity. For example, including all potential interaction terms increased the BIC of the synchrony

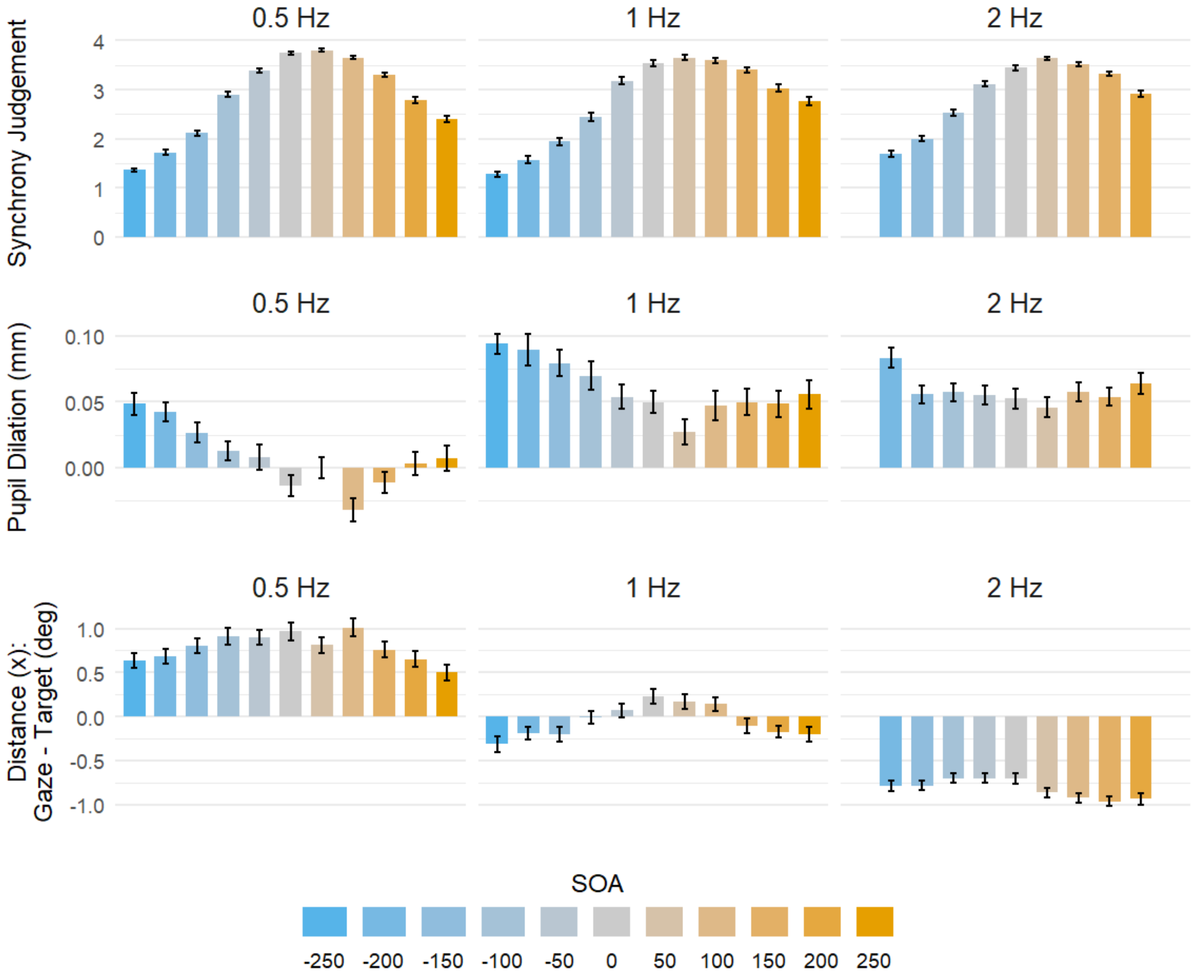


Figure 2: Mean values for synchrony judgement (upper row), baseline corrected pupil diameter (middle row) and distance in the x-direction between gaze position and fixation cross (lower row) in all frequency - SOA combinations. Error bars indicate standard errors, adjusted for repeated measures (n = 23).

judgement model to 13602.08, while the model with only one interaction between soa_negative and absolute_SOA resulted in a BIC of 13592.21. The higher BIC values indicate that the additional complexity from these interaction terms was not justified, as it did not provide substantial improvement in model fit.

Each model was validated as described in **Section 3.5.2** and estimates are presented in **Table 2**. It can be observed that all predictors are statistically significant, except for the distance predictor in the synchrony judgement model. The interaction between negative SOA (audio first) and absolute SOA also showed a significant negative effect in the judgement model. In the pupil dilation model, the estimate of the predictor absolute SOA corresponds to the rounded value 1.724e-04 indicating that for every millisecond increase in absolute SOA, pupil dilation increases by approximately 0.0001724 mm, which for example corresponds to an increase of 0.04 mm in case of the 250 ms SOA condition.

Investigating estimated marginal means plots (not shown) of the interaction in the judgement model (SOA negative × absolute SOA) reveals that the slope for negative SOAs is steeper than the slope for positive SOA conditions, indicating that as the absolute SOA value increases, the decrease in synchrony judgements is more pronounced when the audio is presented first. This can also be obtained from the descriptive statistics in **Figure 2**. ICC for the random effect participant, obtained from the null model was 0.09 for the judgement model, 0.11 for the pupil dilation model and 0.17 for the gaze model.

4.2 Influence of Frequency on Synchrony Judgement

To further investigate the influence of stimulus frequency on synchrony judgement, we combined the two visual compositions due to their lack of significant influence on subjective synchrony (cf.

Table 2: Combined results of the LMM models for the three dependent variables. Note, that estimates are unstandardized regression coefficients; t-tests use Satterthwaite's method; significant predictors are shown with the following codes: $p < 0.001$ ***, $p < 0.01$ **, $p < 0.05$ *.

Model	Estimates	95% *CI*	*df*	*t*	*p*
Synchrony Judgements					
(Intercept)	3.993	3.82 – 4.16	55.80	47.16	< .001 ***
distance	-0.060	-0.13 – 0.00	5837	-1.72	.068
frequency	-0.092	-0.13 – -0.05	5837	-4.02	< .001 ***
SOA negative / audio first	-0.174	-0.26 – -0.09	5837	-43.01	< .001 ***
absolute SOA	-0.004	-0.00 – -0.00	5837	-46.08	< .001 ***
SOA negative × absolute SOA	-0.005	-0.01 – -0.00	5837	-46.08	< .001 ***
Mean Pupil Dilation					
(Intercept)	0.004	-0.02 – -0.03	71.56	0.37	.716
distance	-0.022	-0.03 – -0.01	5592	-4.35	< .001 ***
frequency	0.025	0.02 – 0.03	5592	8.04	< .001 ***
SOA negative / audio first	0.022	-0.03 – -0.02	5592	-6.98	< .001 ***
absolute SOA	0.000	0.00 – 0.00	5592	5.95	< .001 ***
Mean Gaze-Target Distance					
(Intercept)	1.406	1.20 – 1.62	40.50	13.21	< .001 ***
distance	-0.393	-0.46 – -0.32	5246	-11.00	< .001 ***
frequency	-0.943	-0.98 – -0.90	5245	-46.00	< .001 ***
SOA negative / audio first	0.045	0.00 – 0.09	5245	2.13	.033 *
absolute SOA	-0.001	-0.00 – -0.00	5245	-8.34	< .001 ***

Table 2). While Gaussian probability density functions are often used to fit subjective synchrony judgements [10, 12, 39], we found that fitted functions did not describe our data well. To examine the influence of stimulus frequency on the point of subjective synchrony (PSS) and the window of subjective synchrony (WSS), we thus employed the At-A-Glance model [78], which is more agnostic, has been shown to produce good fits, and can account for temporal re-calibration. The At-A-Glance model was fit for each participant and frequency using the dichtomised synchrony ratings.

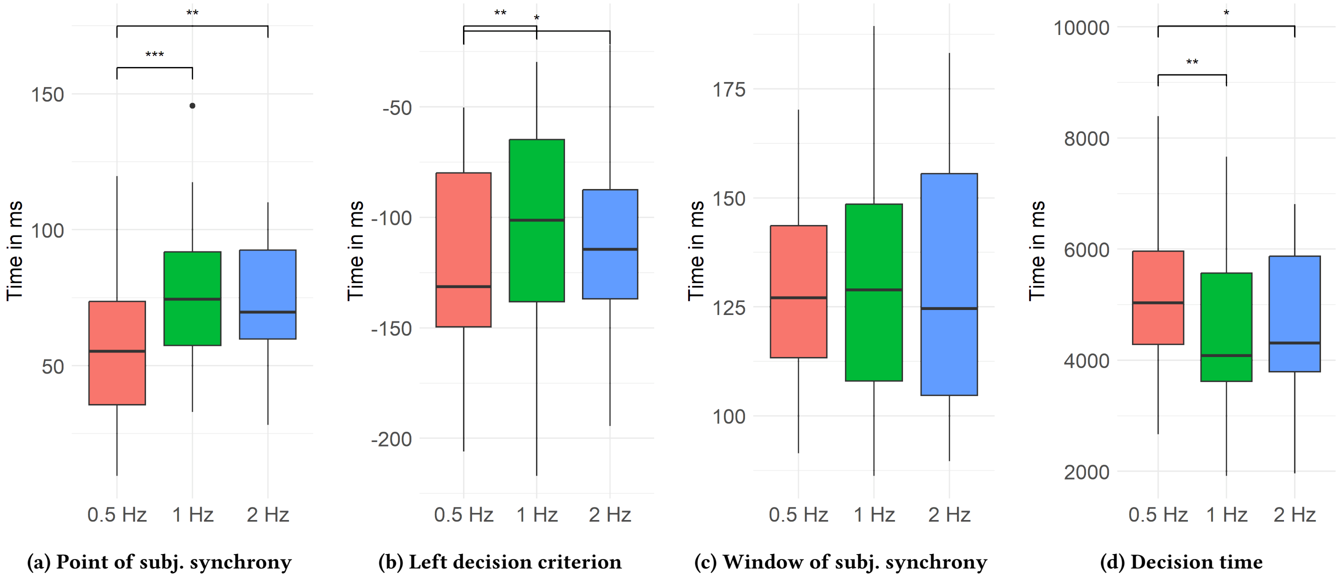
The fitting was done in R using the open-source code from Yarrow et al. [78]; the quality of the fit was evaluated using Leave-One-Out cross-validation (LOO) and Pareto k diagnostics (96.8% of the observations classified as good with $k < 0.7$). To verify the fit, R^2 scores were computed for the fitted functions and participants with R^2 score below 0.5 for at least one frequency condition were excluded from the analysis. Excluding participants based on this criterion was motivated by related work on the investigation of synchrony and temporal order judgements, as it is an indicator that participants were unable to achieve a task [39, 69]. This resulted in six participants being excluded from the synchrony judgement analysis.

The At-A-Glance model is structured around two cumulative normal distributions, each defined by a decision criterion and the associated variability. To quantify the judgement sensitivity, the decision range was computed by determining the temporal distance between the left decision criterion (LDC) and the right decision criterion (RDC). In addition, the WSS was calculated as the time range centered around the PSS for which the area under the function covers 68.26% of the estimated total area under the fitted At-A-Glance function. This computation is inspired by the computation of the Temporal Integration Window (TIW) in related work [39], which is derived through the standard deviation and refers to the time range centered around the PSS which covers 68.26% of the total area underneath the Gaussian curve.

In addition to the PSS, the decision range, the WSS, the LDC, and the RDC, we included the mean decision time (DT) in the analysis (cf. **Figure 3**). Q-Q plots were used to assess the normality of parameters across different frequency conditions (cf. **Section 3.5.2**). Most parameters, including the PSS, WSS, LDC, DT, and the decision range were normally distributed across all frequencies (0.5 Hz, 1 Hz, 2 Hz). Only RDC showed slight deviations from normality at 0.5 Hz and 1 Hz.

To investigate the influence of the frequency condition on the normally distributed parameters, different LMMs were used. Again, participants were chosen to be the random effect and PSS, WSS, LDC, DT and decision range were respectively entered as a dependent variable.

Here, the 0.5 Hz frequency condition was used as the reference level in the LMMs. The analysis reveals that frequency significantly influences both the PSS and the LDC. For the PSS, the intercept



(a) Point of subj. synchrony

(b) Left decision criterion

(c) Window of subj. synchrony

(d) Decision time

Figure 3: Parameters of the fitted At-A-Glance functions and the decision times. Note, that $p < 0.001$ ***, $p < 0.01$ **, $p < 0.05$ *

for the 0.5 Hz condition was 56.07. Significant effects were found at both 1 Hz ($\beta = 20.99$, $t = 3.81$, $p < .001$) and 2 Hz ($\beta = 15.57$, $t = 2.82$, $p < .01$). For the LDC, the intercept at 0.5 Hz was −123.4, with significant effects at 1 Hz ($\beta = 19.41$, $t = 2.87$, $p < .01$) and 2 Hz ($\beta = 14.05$, $t = 2.08$, $p < .05$). In addition, a significant effect on the DT was observed. The intercept was 5112.18 with significant effects for both the 1 Hz ($\beta = -693.05$, $t = -2.91$, $p < .01$) and the 2 Hz condition ($\beta = -589.50$, $t = -2.47$, $p < .05$). In contrast, the null hypothesis could not be rejected for the decision range (intercept = 359.0, with no significant effects at 1 Hz or 2 Hz) and the WSS (intercept = 127.6, with no significant effects at 1 Hz or 2 Hz), indicating no significant influence of frequency on these parameters.

4.3 Influence of Synchrony Perception on Oculomotor Responses

Finally, we aimed to investigate how *perceived* (a)synchrony affects pupil dilation and gaze-target distance. When selecting the LMMs, we took into account the criteria described described in **Section 4.1** and chose synchrony judgement, as well as distance and frequency as predictors. Frequency was added as an interaction term based on a preliminary visual inspection of our data.

The pupil dilation model revealed significant negative effects of distance ($\beta = -0.02, t = -4.60, p < 0.001$) and synchrony judgement ($\beta = -0.02, t = -8.07, p < 0.001$) on pupil dilation. The main effect of frequency was not significant ($\beta = 0.01, t = 1.090, p = 0.276$), however, the interaction between frequency and synchrony judgement was ($\beta = 0.01, t = -8.08, p = 0.012$), indicating that the effect of synchrony judgement on pupil diameter varied for different frequencies. The SPEM model showed a significant main effect of distance ($\beta = 0.26, t = 5.35, p < 0.001$), frequency $\beta = -0.64, t = -9.91, p < 0.001$), and synchrony judgement ($\beta = 0.20, t = 7.43, p < 0.001$). The interaction effect was also significant ($\beta = -0.11, t = -5.34, p < 0.001$), demonstrating that synchrony judgements affect SPEM differently depending on the frequency.

The Estimated Marginal Means (EMM) plots shown in **Figure 4** illustrate the interaction effects. For the predicted pupil diameter, it can be seen that the slope is slightly steeper at lower frequencies, illustrating that the effect of the synchrony judgement on pupil dilation decreases with increasing frequency. However, the difference is not as prominent as for the gaze-target distance. Here, for 2 Hz there is almost no effect of synchrony judgement visible, while 0.5 and 1 Hz show a clear upward trend with higher EMMs for trials that were rated as more synchronous. Additionally the negative EMM values for 2 Hz indicate that for this frequency the gaze was lagging behind the target, while for 0.5 and 1 Hz gaze was preceding the target.

The relationship between perceived asynchrony with pupil dilation as well as with SPEM might be affected by the fact that all three variables were influenced by frequency. To further investigate potential multicollinearity, we calculated Generalized Variance Inflation Factors (GVIF) and, to make them comparable across dimensions, $\text{GVIF}^{1/(2^*\text{Df})}$, as suggested by [15]. Since all $\text{GVIF}^{1/(2^*\text{Df})}$ values were <2 we conclude that multicollinearity is not a severe issue in our case. We also evaluated whether frequency might act as a confounding variable. Our analysis shows that removing frequency as a predictor does not drastically alter the influence of synchrony judgement on pupil diameter or gaze-target distance. Further, in both cases the full models as described in **Section 4.1** fit better than the reduced models. Therefore, frequency does not appear to be a significant confounder, and no serious issues of multicollinearity arise.

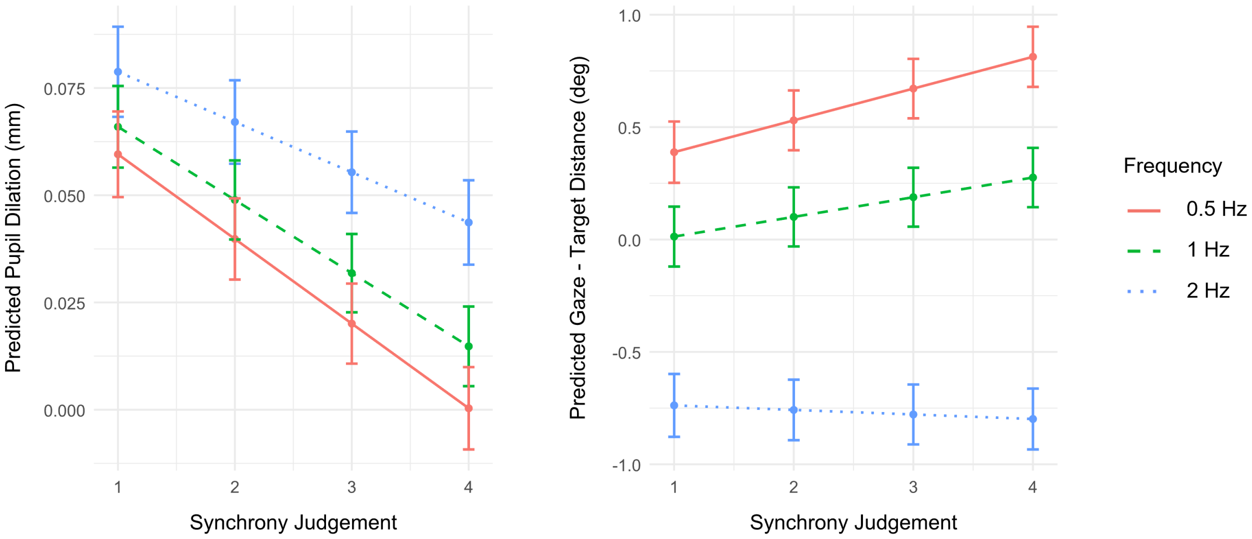


Figure 4: Estimated Marginal Means (EMM) for predicted pupil dilation (mm) and predicted gaze - target distance (deg) as a function of synchrony judgement.

5 Discussion

This section discusses the results presented in **Section 4** and reflects on factors that influence the processing of rhythmic audiovisual stimuli. We address the limitations of our work and discuss future work (**Section 5.4**).

5.1 Synchrony Judgements

Our results show that higher frequencies are perceived as slightly more asynchronous (cf. **Table 2**), suggesting that the task of perceiving synchrony becomes more difficult at higher frequencies [5]. As anticipated, stimuli with greater temporal offset are perceived as more asynchronous, confirming participants' ability to detect SOAs between audio and visual stimuli up to a certain limit [68]. Additionally, our model shows that when audio is presented first (negative SOA), participants perceive the stimuli as less synchronous. This is in line with previous findings that show that mean PSS are usually found to be video-leading [68, 73], and therefore by our definition, positive. A common explanation is that we are accustomed to sound preceding visual cues from the same source given that sound travels slower than light [73]. Further, the significant interaction between soa_negative and absolute_soa and our investigation of estimated marginal means suggest that the decline in synchrony judgements as SOA increases is more pronounced when the audio precedes the visual stimulus. Distance covered by the visual stimulus was found to not significantly influence synchrony ratings, indicating that participants primarily relied on the timing between the auditory and visual stimuli rather than the extent of the visual movement to judge synchrony.

Our analysis of the synchrony judgement parameters (cf. **Section 4.2**) reveals that while the WSS and the decision range were not significantly influenced by stimulus frequency, the PSS was. Specifically, a shift of approximately 20 ms was observed, indicating that at higher frequencies, participants tended to perceive synchrony when the sound slightly lagged behind the visual stimulus. This may be because at higher frequencies, the integration of audiovisual inputs becomes more demanding, and the brain might adjust by tolerating a slight audio lag to perceive synchrony. Furthermore, decision times were significantly longer for the 0.5 Hz condition. This is, however, likely because more time was required for observing a given number of stimulus occurrences compared to the 1 Hz and 2 Hz conditions.

Overall, our results suggest that participants are more forgiving of audiovisual asynchronies at higher frequencies, particularly when the audio lags behind the visual stimulus. This can have practical implications for scenarios where network latency leads to such asynchronies, as higher frequencies may more effectively mask these discrepancies. However, there is a general trend towards lower synchrony ratings at higher frequencies, suggesting that although the brain can adapt to slight delays, synchrony perception still becomes more difficult as the pace of the stimuli increases.

5.2 Mean Pupil Dilation

The analysis of mean pupil dilation revealed that absolute SOA significantly affects pupil diameter, even though the estimate is close to zero since we decided against standardizing variables for the sake of better interpretability of the estimates. Specifically, a non-rounded coefficient of 1.724e-04 suggests that, assuming other predictors remain constant, pupil diameter increases by approximately 0.0001724 mm per 1 ms SOA. As SOA increases by 100 ms, pupil dilation grows by around 0.02 mm, which falls within a plausible range for such changes. This effect aligns with our results on subjective synchrony perception, where a one-point decrease in the synchrony rating on a four-step Likert scale corresponds to a 0.02 mm increase in pupil diameter. These results confirm that the primary observations by Skaansar et al. [66] regarding pupil dilation in response to musical asynchronies extend to audiovisual stimuli as well. Both actual and subjective asynchronies are associated with increased pupil dilation, which Skaansar et al. attribute to the heightened cognitive workload caused by asynchronies that demand a broader attentional focus. The dilation may also be related to a violation of expectations, as observed when audio is perceived too early or delayed – a pattern supported by previous research on pupil dilation [4, 36, 54].

Interestingly, SOAs where the audio plays before the visual stimulus are linked to smaller pupil diameters than those where the audio is delayed, reflecting a close relationship between pupil dilation and *perceived* asynchrony, as synchrony judgements show the corresponding effect of steeper slopes for negative SOAs (cf. **Figure 2** and **Table 2**) Additionally, higher frequencies were found to increase pupil dilation, likely due to the corresponding increase in stimulus speed, which has been shown to cause pupil dilation [14]. The interaction effect between synchrony judgement and frequency suggests that as frequency increases, its influence on pupil dilation diminishes. This might be because frequency-induced dilation is bound by physiological limits. Lastly, the results also indicate that greater distances are associated with smaller pupil diameters, a surprising outcome that may reflect complex attentional dynamics.

These results suggest that pupil dilation might be used as a non-invasive indicator of perceived audiovisual asynchrony in interactive systems such as virtual reality, remote communication or collaborative systems. Tracking pupil responses could enable real-time adjustments to increase perceived synchrony and user experience. The relationship between frequency and pupil dilation also suggests that higher frequency interactions place greater cognitive demands on users.

5.3 Mean Gaze-Target Distance

The results (cf. **Table 2** and **Figure 4**) indicate that higher SOAs and correspondingly lower subjective synchrony ratings, lead to an increased lag in Smooth Pursuit Eye Movements (SPEM). Specifically, a decrease in the synchrony rating by one step on a four-step Likert scale is associated with a 0.20 degree increase in SPEM lag, potentially reflecting a compensatory adjustment in gaze to match the delayed stimulus. This underlines earlier findings that SPEM play a central role in the processing of synchrony and anticipation of stimuli [41, 63].

Our models further suggests that SPEM tends to lag behind the fixation cross for larger distances and higher frequencies, while at the slower frequency of 0.5 Hz, SPEM tends to precede the target. This is expected since both higher frequencies and larger distances are linked to increased velocity, likely making it harder for the gaze to keep up with the stimulus. This increased lag of SPEM for higher frequencies was already shown in early work on eye tracking [37]. However, in contrast to our results, they found no significant influence of distance. Additionally, the interaction between synchrony ratings and frequency mirrors the findings for pupil dilation: at higher frequencies like 2 Hz, the accuracy of eye movements is constrained, even when the stimuli are perceived as synchronous.

5.4 Limitations and Future Work

Initially, our study was designed to be more application-focused, incorporating user interaction and stimuli that more closely mirror typical HCI environments. However, we quickly realized that foundational research was necessary first to better understand the perception of audio-visual latency and determine whether implicit eye-tracking measures even have the potential to be an indicator of perceived stimulus asynchrony. Consequently, we shifted to a controlled laboratory approach to build a robust foundation for future application-oriented studies.

It has to be acknowledged that 87.65% of our pupil baseline measurements were smaller than the standard range of 3-7 mm, likely due to the brightness of the experimental environment. Although measures were taken to ensure uniform lighting conditions, the overall brightness may have been too high, possibly affecting the pupil dilation results. However, our results are consistent with previous research on pupil dilation and audio synchrony [66]. In addition, no major deviations from the identified trends were observed, indicating the robustness of our results.

Due to the exclusion of one participant who rated all stimuli as synchronous, our study was not fully crossed. Although this slightly unbalanced design may have led to minor variations, we believe that it did not significantly affect the overall results as the randomisation and counterbalancing of conditions ensured that the key factors of interest were still appropriately tested across all participants. The robustness of the observed effects suggests that the exclusion did not affect the analysis results or the conclusions drawn from the study.

As discussed in **Section 2.1**, latency perception varies significantly depending on the task and stimuli, raising the need for future work to assess the applicability of our results to more realistic scenarios. Nevertheless, the proposed implicit eye-tracking measures could potentially provide real-time insights into synchrony perception across diverse scenarios, reducing the need for exhaustive prior testing of every variant.

Although our work shows that there is a link between the perception of asynchrony, pupil dilation and SPEM, further research is needed before these measures can be used as a reliable indicator of perceived latency in systems.

It is important to note that pupil dilation is a nonspecific measure that cannot distinguish between cognitive load caused by task difficulty, other external factors, or perceived latency itself. Therefore, this measure can only be reliably applied in scenarios where cognitive load from non-latency-related sources remains relatively stable.

Pupil diameter is also known to be influenced by numerous other factors, like pupil light reflex or physical movement. However, prior work on rhythmic audio stimuli suggests that pupil responses continue to reflect timing differences even when users are actively engaging in rhythmic actions [66], indicating that these physiological markers hold potential for application in more complex, interactive contexts. Additionally, recent research demonstrates that it is possible to account for the influence of light by subtracting its effects [56]. In VR applications, this is particularly straightforward as the lighting of the environment is fully controlled at all times [53]. Similarly, artefacts caused by physical movements could also be mitigated using VR's body-tracking capabilities, which provide detailed information on the user's activity.

Nonetheless, it remains to be shown that our results extend beyond the controlled laboratory environment, especially in more dynamic and interactive contexts like social VR, or when using more realistic stimuli, such as songs with varying beats per minute (bpm). Future work should examine the impact of user interaction and movement on latency perception, for instance, by having participants wave in rhythm with an avatar. This would help identify potential changes in latency thresholds and assess how physical activity influences eye-tracking measurements as an implicit indicator of perceived asynchrony in rhythmic stimuli. We are currently planning a follow-up study to address these questions and to investigate further how we can enhance feelings of closeness in HCI applications, despite the unavoidable impact of latency.

6 Conclusion

In summary, this study simulated latency to investigate fundamental principles of synchrony perception in rhythmic audiovisual stimuli within computer-mediated contexts. We demonstrate that both the SOA (simulated latency) and frequency have a significant influence on synchrony judgements, pupil dilation and smooth pursuit eye movements.

Our results confirm findings from related work that audio which lags behind a video is perceived as more synchronous compared to the opposite case (video lagging behind audio). In addition, we found that frequency significantly affects the point of subjective synchrony, with higher frequencies shifting the PSS by up to 19.4 ms in the direction of audio lagging video. Building on our findings, we can derive recommendations for practical applications that leverage IMS by tailoring the synchronization rhythm to specific use cases. Medium to higher frequencies (1 Hz and 2 Hz, corresponding to 60 bpm and 120 bpm songs) may be more suitable for scenarios where the audio lags behind the video, for instance, a remote orchestra responding to local movements of a conductor or when remote musicians play music dynamically based on the movement of dancers. In contrast, lower frequencies (0.5 Hz, corresponding to 30 bpm) may be better suited for settings where video follows audio, such as dancing or work-outs to a globally synchronized beat.

We found that both actual and perceived asynchronies were associated with increased pupil dilation and a greater lag in eye movements. Our results suggest that eye-tracking data, such as pupil dilation and SPEM, can serve as physiological indicators of perceived asynchrony. This holds potential to improve real-time interactions in HCI systems, particularly in remote collaboration and communication, where precise timing and synchrony are important. For example, physiological measurements could indicate when users no longer perceive an interaction as synchronous, which is context-dependent and influenced by the virtual environment, the nature of the interaction, or interactivity levels. Relying solely on fixed latency thresholds may not be sufficient, as synchrony perception varies in different settings and individuals. In adaptive systems, these indicators could enable dynamic adjustments, such as reducing network traffic by lowering video resolution or compressing data, or applying latency-masking techniques to maintain the sense of synchrony.

As mentioned in **Section 5.4**, it remains a challenge to distinguish pupil changes caused by asynchrony from those due to task difficulty or other sources of cognitive load. This also means that when using pupil diameter as a measure of task load in computer-mediated, latency-affected contexts, it is important to consider that observed pupil dilation could also originate from perceived asynchrony, as demonstrated in our study. Nevertheless, pupillometry offers significant advantages: it is highly non-intrusive, readily available in HCI contexts through screen-based eye trackers, and allows for real-time monitoring, making it a potentially valuable tool to detect perceived latency if it proves effective in real-world applications.

In summary, this study suggests that eye-tracking data can serve as an implicit measure of perceived asynchrony and emphasizes the need for frequency-specific design in computer-mediated interactions, with medium to higher frequencies favoring audio-lagging-video contexts and lower frequencies suiting video-lagging-audio scenarios. Future research should study our results in more realistic and immersive scenarios to assess their effect on entrainment for increasing the sense of community and closeness.

References

[1] Wieslaw Bartkowski, Andrzej Nowak, Filip Ignacy Czajkowski, Albrecht Schmidt, and Florian Müller. 2023. In Sync: Exploring Synchronization to Increase Trust Between Humans and Non-humanoid Robots. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 367, 14 pages. doi:10.1145/3544548.3581193

[2] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67, 1 (2015), 1–48. doi:10.18637/jss.v067.i01

[3] Jackson Beatty and Brennis Lucero-Wagoner. 2000. The pupillary system. In *Handbook of psychophysiology*, J. T. Cacioppo, L. G. Tassinary, and G. G. Berntson (Eds.). Cambridge University Press, Cambridge, 142–162.

[4] Janika Becker, Marvin Viertler, Christoph W Korn, and Helen Blank. 2024. The pupil dilation response as an indicator of visual cue uncertainty and auditory outcome surprise. *European Journal of Neuroscience* 59, 10 (2024), 2686–2701.

[5] Jeroen S. Benjamins, Maarten J. van der Smagt, and Frans A. J. Verstraten. 2008. Matching auditory and visual signals: is sensory modality just another feature? *Perception* 37, 6 (2008), 848–858. doi:10.1068/p5783

[6] James V Bradley. 1958. Complete counterbalancing of immediate sequential effects in a Latin square design. *J. Amer. Statist. Assoc.* 53, 282 (1958), 525–528.

[7] David Bridges, Alain Pitiot, Michael R MacAskill, and Jonathan W Peirce. 2020. The timing mega-study: Comparing a range of experiment generators, both lab-based and online. *PeerJ* 8 (2020), e9414.

[8] Carrie J Cai, Michelle Carney, Nida Zada, and Michael Terry. 2021. Breakdowns and Breakthroughs: Observing Musicians' Responses to the COVID-19 Pandemic. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 571, 13 pages. doi:10.1145/3411764.3445192

[9] Rachel Carlson and Scot Hanna-Weir. 2021. Conducting During COVID: What is possible and how has the role of the conductor changed? *The Choral Journal* 61(9) (2021), 65–73. https://www.jstor.org/stable/27035082.

[10] John Cass, Diane Oake, and Erik van der Burg. 2015. Stretching time: Relativistic lag-induced shifts in perceived audiovisual synchrony using cluttered displays. *Journal of vision* 15, 11 (2015), 9. doi:10.1167/15.11.9

[11] Anne Danielsen, Mari Romarheim Haugen, and Alexander Refsum Jensenius. 2005. Moving to the beat: Studying entrainment to micro-rhythmic changes in pulse by motion capture. *Timing and Time Perception* 3 (2005), 133–154. doi:10.1163/22134468-00002043

[12] Ragnhild Eg and Dawn M Behne. 2015. Perceived synchrony for realistic and dynamic audiovisual events. *Frontiers in psychology* 6 (2015), 736.

[13] Carmine Elvezio, Frank Ling, Jen-Shuo Liu, and Steven Feiner. 2018. Collaborative Virtual Reality for Low-Latency Interaction. In *Adjunct Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology* (Berlin, Germany) *(UIST '18 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 179–181. doi:10.1145/3266037.3271643

[14] Lauren Fink, Brian Hurley, Joy Geng, and Petr Janata. 2018. A linear oscillator model predicts dynamic temporal attention and pupillary entrainment to rhythmic patterns. *Journal of Eye Movement Research* 11 (11 2018), 12. doi:10.16910/jemr.11.2.12

[15] John Fox and Georges Monette. 1992. Generalized collinearity diagnostics. *J. Amer. Statist. Assoc.* 87, 417 (1992), 178–183.

[16] Waka Fujisaki, Shinsuke Shimojo, Makio Kashino, and Shin'ya Nishida. 2004. Recalibration of audiovisual simultaneity. *Nature neuroscience* 7, 7 (2004), 773–778. doi:10.1038/nn1268

[17] Ilanit Gordon, Avi Gilboa, Shai Cohen, Nir Milstein, Nir Haimovich, Shay Pinhasi, and Shahar Siegman. 2020. Physiological and Behavioral Synchrony Predict Group Cohesion and Performance. *Scientific reports* 10, 1 (2020), 8484. doi:10.1038/s41598-020-65670-1

[18] Eric Granholm and Stuart Steinhauer. 2004. Pupillometric measures of cognitive and emotional processes. *International journal of psychophysiology : official journal of the International Organization of Psychophysiology* 52 (04 2004), 1–6. doi:10.1016/j.ijpsycho.2003.12.001

[19] Nina Heins, Jennifer Pomp, Daniel S Kluger, Stefan Vinbrüx, Ima Trempler, Axel Kohler, Katja Kornysheva, Karen Zentgraf, Markus Raab, and Ricarda I Schubotz. 2021. Surmising synchrony of sound and sight: Factors explaining variance of audiovisual integration in hurdling, tap dancing and drumming. *Plos one* 16, 7 (2021), e0253130.

[20] Bert Hoeks and Willem JM Levelt. 1993. Pupillary dilation as a measure of attention: A quantitative system analysis. *Behavior Research methods, Instruments, & Computers* 25, 1 (1993), 16–26. doi:10.3758/BF03204445

[21] Bernhard Hommel, Jochen Müsseler, Gisa Aschersleben, and Wolfgang Prinz. 2001. The theory of event coding (TEC): A framework for perception and action planning. *Behavioral and brain sciences* 24, 5 (2001), 849–878.

[22] 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*, Henry Duh (Ed.). IEEE, Piscataway, NJ, 26–34. doi:10.1109/ISMAR55827.2022.00016

[23] Michael J. Hove and Jane L. Risen. 2009. It's all in the timing: Interpersonal synchrony increases affiliation. *Social cognition* 27, 6 (2009), 949–960.

[24] Sakari Ilomäki, Johanna Ruusuvuori, and Jaana Laitinen. 2021. Effects of Transmission Delay on Client Participation in Video-Mediated Group Health Counseling. *Qualitative Health Research* 31(13) (2021), 2328–2339. doi:10.1177/10497323211010726

[25] Rino Imai, Ryota Matsui, Yutaka Yanagisawa, Yoshinari Takegawa, and Keiji Hirata. 2023. Survey on the Effect of Video Delay in Online Dance with Multiple Participants. In *Human-Computer Interaction*, Masaaki Kurosu and Ayako Hashizume (Eds.). Springer Nature Switzerland, Cham, 375–384.

[26] Katherine Isbister, Elena Márquez Segura, Suzanne Kirkpatrick, Xiaofeng Chen, Syed Salahuddin, Gang Cao, and Raybit Tang. 2016. Yamove! A Movement Synchrony Game That Choreographs Social Interaction. *Human Technology* 12 (05 2016), 74–102. doi:10.17011/ht/urn.201605192621

[27] Mari Riess Jones. 2016. *Musical Time.* Oxford University Press, Oxford. 125 pages.

[28] Topi Kaaresoja, Stephen Brewster, and Vuokko Lantz. 2014. Towards the temporally perfect virtual button: touch-feedback simultaneity and perceived quality in mobile touchscreen press interactions. *ACM Transactions on Applied Perception (TAP)* 11, 2 (2014), 1–25.

[29] Stamos Katsigiannis, Dimitris Maroulis, and Georgios Papaioannou. 2013. A GPU based real-time video compression method for video conferencing. In *2013 18th International Conference on Digital Signal Processing (DSP)*. IEEE, IEEE Computer Society, USA, 1–6.

[30] Marcel Kinsbourne and Molly Helt. 2011. *Social entrainment of typically developing and autistic children.* Oxford University Press, Oxford, UK, 339–365.

[31] Tanja Kojic, Steven Schmidt, Sebastian Möller, and Jan-Niklas Voigt-Antons. 2019. Influence of Network Delay in Virtual Reality Multiplayer Exergames: Who is actually delayed?. In *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, Berlin, Germany, 1–3. doi:10.1109/QoMEX.2019.8743342

[32] Edward W Large and Mari Riess Jones. 1999. The dynamics of attending: How people track time-varying events. *Psychological Review* 106 (1999), 119–159. doi:10.1037/0033-295X.106.1.119

[33] Jacques Launay, Roger T. Dean, and Freya Bailes. 2014. Synchronising movements with the sounds of a virtual partner enhances partner likeability. *Cognitive Processing* 15 (2014), 491–501.

[34] Insoo Lee, Jinsung Lee, Kyunghan Lee, Dirk Grunwald, and Sangtae Ha. 2021. Demystifying Commercial Video Conferencing Applications. In *Proceedings of the 29th ACM International Conference on Multimedia* (Virtual Event, China) *(MM '21)*. Association for Computing Machinery, New York, NY, USA, 3583–3591. doi:10.1145/3474085.3475523

[35] Christophe Leys, Christophe Ley, Olivier Klein, Philippe Bernard, and Laurent Licata. 2013. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of experimental social psychology* 49, 4 (2013), 764–766.

[36] Hsin-I Liao, Makoto Yoneya, Makio Kashino, and Shigeto Furukawa. 2018. Pupillary dilation response reflects surprising moments in music. *Journal of Eye Movement Research* 11, 2 (2018), 1–13.

[37] SG Lisberger, C Evinger, GW Johanson, and AF Fuchs. 1981. Relationship between eye acceleration and retinal image velocity during foveal smooth pursuit in man and monkey. *Journal of Neurophysiology* 46, 2 (1981), 229–249.

[38] Shengmei Liu, Xiaokun Xu, and Mark Claypool. 2022. A survey and taxonomy of latency compensation techniques for network computer games. *ACM Computing Surveys (CSUR)* 54, 11s (2022), 1–34.

[39] Scott A. Love, Karin Petrini, Adam Cheng, and Frank E. Pollick. 2013. A Psychophysical Investigation of Differences between Synchrony and Temporal Order Judgments. *PloS one* 8, 1 (2013), e54798. doi:10.1371/journal.pone.0054798

[40] Hamish G. MacDougall and Steven T. Moore. 2005. Marching to the beat of the same drummer: the spontaneous tempo of human locomotion. *Journal of applied physiology (Bethesda, Md. : 1985)* 99, 3 (2005), 1164–1173. doi:10.1152/japplphysiol.00138.2005

[41] Guillaume S Masson and Leland S Stone. 2002. From following edges to pursuing objects. *Journal of Neurophysiology* 88(5) (2002), 2869–2873. doi:10.1152/jn.00338.2002

[42] Sebastiaan Mathôt and Ana Vilotijević. 2023. Methods in cognitive pupillometry: Design, preprocessing, and statistical analysis. *Behavior Research Methods* 55, 6 (2023), 3055–3077.

[43] Sebastiaan Mathôt. 2024. Python DataMatrix. https://github.com/open-cogsci/datamatrix Accessed 2024-07-17.

[44] Joseph E McGrath. 1986. Time and human interaction: Toward a social psychology of time.

[45] Geoff Norman. 2010. Likert scales, levels of measurement and the "laws" of statistics. *Advances in health sciences education* 15 (2010), 625–632.

[46] Margarethe Olbertz-Siitonen. 2015. Transmission delay in technology-mediated interaction at work. *PsychNology Journal* 13(2-3) (2015), 203–234.

[47] Yun Suen Pai, Ryo Hajika, Kunal Gupta, Prasanth Sasikumar, and Mark Billinghurst. 2020. NeuralDrum: Perceiving Brain Synchronicity in XR Drumming. In *SIGGRAPH Asia 2020 Technical Communications* (Virtual Event, Republic of Korea) *(SA '20)*. Association for Computing Machinery, New York, NY, USA, Article 8, 4 pages. doi:10.1145/3410700.3425434

[48] Taiwoo Park, Uichin Lee, Bupjae Lee, Haechan Lee, Sanghun Son, Seokyoung Song, and Junehwa Song. 2013. ExerSync: facilitating interpersonal synchrony in social exergames. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work* (San Antonio, Texas, USA) *(CSCW '13)*. Association for Computing Machinery, New York, NY, USA, 409–422. doi:10.1145/2441776.2441823

[49] Jonathan Peirce. 2007. PsychoPy—psychophysics software in Python. *Journal of neuroscience methods* 162, 1-2 (2007), 8–13.

[50] Jonathan Peirce. 2024. Can PsychoPy® deliver millisecond precision? - PsychoPy v2024.1.5. https://www.psychopy.org/general/timing/millisecondPrecision.html Accessed 2024-07-15.

[51] Jonathan Peirce. 2024. Defining the onset/duration of components - PsychoPy v2024.1.5. https://www.psychopy.org/builder/startStop.html Accessed 2024-07-15.

[52] Roosa Piitulainen, Perttu Hämäläinen, and Elisa D Mekler. 2022. Vibing Together: Dance Experiences in Social Virtual Reality. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 188, 18 pages. doi:10.1145/3491102.3501828

[53] Hedenir Monteiro Pinheiro and Ronaldo Martins da Costa. 2021. Pupillary light reflex as a diagnostic aid from computational viewpoint: A systematic literature review. *Journal of Biomedical Informatics* 117 (2021), 103757.

[54] Kerstin Preuschoff, Bernard Marius 't Hart, and Wolfgang Einhäuser. 2011. Pupil dilation signals surprise: Evidence for noradrenaline's role in decision making. *Frontiers in neuroscience* 5 (2011), 115.

[55] R Core Team. 2021. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

[56] Pallavi Raiturkar, Andrea Kleinsmith, Andreas Keil, Arunava Banerjee, and Eakta Jain. 2016. Decoupling light reflex from pupillary dilation to measure emotional arousal in videos. In *Proceedings of the ACM Symposium on Applied Perception* (Anaheim, California) *(SAP '16)*. Association for Computing Machinery, New York, NY, USA, 89–96. doi:10.1145/2931002.2931009

[57] Paul Reddish, Ronald Fischer, and Joseph Bulbulia. 2013. Let's dance together: synchrony, shared intentionality and cooperation. *PloS one* 8, 8 (2013), e71182. doi:10.1371/journal.pone.0071182

[58] Bruno H. Repp and Yi-Huang Su. 2013. Sensorimotor synchronization: a review of recent research (2006-2012). *Psychonomic bulletin & review* 20, 3 (2013), 403–452. doi:10.3758/s13423-012-0371-2

[59] Michal Rinott and Noam Tractinsky. 2022. Designing for interpersonal motor synchronization. *Human–Computer Interaction* 37, 1 (2022), 69–116. doi:10.1080/07370024.2021.1912608

[60] Raquel Breejon Robinson, Elizabeth Reid, James Collin Fey, Ansgar E. Depping, Katherine Isbister, and Regan L. Mandryk. 2020. Designing and Evaluating 'In the Same Boat', A Game of Embodied Synchronization for Enhancing Social Play. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3313831.3376433

[61] Michael Rofe and Federico Reuben. 2017. Telematic performance and the challenge of latency. *Journal of Music, Technology & Education* 10, 2-3 (2017), 167–183.

[62] Karen Ruhleder and Brigitte Jordan. 2001. Co-Constructing Non-Mutual Realities: Delay-Generated Trouble in Distributed Interaction. *Computer Supported Cooperative Work (CSCW)* 10(1) (2001), 113–138. doi:10.1023/A:1011243905593

[63] Alexander C Schütz, Doris I Braun, and Karl R Gegenfurtner. 2011. Eye movements and perception: A selective review. *Journal of Vision* 11(5) (2011), 9. doi:10.1167/11.5.9

[64] Natalie Sebanz and Guenther Knoblich. 2009. Prediction in joint action: What, when, and where. *Topics in cognitive science* 1, 2 (2009), 353–367.

[65] Ben Shneiderman, Catherine Plaisant, Maxine S Cohen, Steven Jacobs, Niklas Elmqvist, and Nicholas Diakopoulos. 2017. *Designing the User Interface: Strategies for Effective Human-Computer Interaction* (6 ed.). Pearson, US.

[66] Jo Fougner Skaansar, Bruno Laeng, and Anne Danielsen. 2019. Microtiming and mental effort: Onset asynchronies in musical rhythm modulate pupil size. *Music Perception: An Interdisciplinary Journal* 37, 2 (2019), 111–133.

[67] Ekaterina R. Stepanova, John Desnoyers-Stewart, Philippe Pasquier, and Bernhard E. Riecke. 2020. JeL: Breathing Together to Connect with Others and Nature. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference* (Eindhoven, Netherlands) *(DIS '20)*. Association for Computing Machinery, New York, NY, USA, 641–654. doi:10.1145/3357236.3395532

[68] J. V. Stone, N. M. Hunkin, J. Porrill, R. Wood, V. Keeler, M. Beanland, M. Port, and N. R. Porter. 2001. When is now? Perception of simultaneity. *Proceedings. Biological sciences* 268, 1462 (2001), 31–38. doi:10.1098/rspb.2000.1326

[69] Yi-Huang Su. 2014. Peak velocity as a cue in audiovisual synchrony perception of rhythmic stimuli. *Cognition* 131, 3 (2014), 330–344. doi:10.1016/j.cognition.2014.02.004

[70] Gail M Sullivan and Anthony R Artino Jr. 2013. Analyzing and interpreting data from Likert-type scales. *Journal of graduate medical education* 5, 4 (2013), 541–542.

[71] Tatsuto Takeuchi, Théodore Puntous, Anup Tuladhar, Sanae Yoshimoto, and Aya Shirama. 2011. Estimation of mental effort in learning visual search by measuring pupil response. *PloS one* 6, 7 (2011), e21973.

[72] Sam van Damme, Javad Sameri, Susanna Schwarzmann, Qing Wei, Riccardo Trivisonno, Filip de Turck, and Maria Torres Vega. 2024. Impact of Latency on QoE, Performance, and Collaboration in Interactive Multi-User Virtual Reality. *Applied Sciences* 14, 6 (2024), 2290. doi:10.3390/app14062290

[73] Rob LJ Van Eijk, Armin Kohlrausch, James F Juola, and Steven van de Par. 2008. Audiovisual synchrony and temporal order judgments: effects of experimental method and stimulus type. *Perception & psychophysics* 70 (2008), 955–968.

[74] Jean Vroomen and Mirjam Keetels. 2010. Perception of intersensory synchrony: a tutorial review. *Attention, perception & psychophysics* 72, 4 (2010), 871–884. doi:10.3758/APP.72.4.871

[75] Jean Vroomen, Mirjam Keetels, Beatrice de Gelder, and Paul Bertelson. 2004. Recalibration of temporal order perception by exposure to audio-visual asynchrony. *Cognitive Brain Research* 22, 1 (2004), 32–35. doi:10.1016/j.cogbrainres.2004.07.003

[76] Yuexin Wang, Yining Guo, Jiajia Wang, Ziyuan Liu, and Xuemin Li. 2021. Pupillary response to moving stimuli of different speeds. *Journal of Eye Movement Research* 14, 1 (2021), 1–12.

[77] Magdalena Wojtczak, Anahita H. Mehta, and Andrew J. Oxenham. 2017. Rhythm judgments reveal a frequency asymmetry in the perception and neural coding of sound synchrony. *Proceedings of the National Academy of Sciences of the United States of America* 114, 5 (2017), 1201–1206. doi:10.1073/pnas.1615669114

[78] Kielan Yarrow, Joshua A. Solomon, Derek H. Arnold, and Warrick Roseboom. 2023. The best fitting of three contemporary observer models reveals how participants' strategy influences the window of subjective synchrony. *Journal of experimental psychology. Human perception and performance* 49, 12 (2023), 1534–1563. doi:10.1037/xhp0001154

[79] Inc. Zoom Communications. 2024. Accessing meeting and phone statistics. https://support.zoom.com/hc/en/article?id=zm_kb&sysparm_article=KB0070504 Accessed 2024-03-12.

Received 12 September 2024; revised 10 December 2024; accepted 13 February 2025