This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/MMUL.2024.3415116 IEEE Multimedia

Article Type: Full Paper

On Perceived AV Synchronization in 360° Multimedia

Aleph Campos da Silveira, *Federal University of Espirito Santo, Vitoria, Espirito Santo, 29932-540, Brazil*
Fotios Spyridonis, *Brunel University London, Uxbridge, UB8 3PN, United Kingdom*
Roope Rosaimo, *Tampere University, Tampere, FI-33014 Finland*
Alexandra Covaci, *University of Kent, Kent, CT2 7NZ United Kingdom*
Gheorghita Ghinea, *Brunel University London, Uxbridge, UB8 3PN, United Kingdom*
Celso Alberto Saibel Santos, *Federal University of Espirito Santo, Vitória, Espirito Santo, 29932-540, Brazil*

Abstract—Media synchronization and, in particular, audiovisual (AV) synchronization, plays a pivotal role in multimedia systems, significantly impacting the Quality of Experience (QoE) perceived by users. What is intriguing is that despite the growing prevalence of multimedia content consumption in 360° environments, the issue of perceived AV synchronization remains relatively unexplored. To tackle this challenge, we present the results of a user study that assessed the influence of AV skews on QoE and the feeling of presence within 360° multimedia content. By examining the various effects of AV skews on the user experience we gain valuable insights into the factors that shape QoE and presence in such immersive settings. These findings have the potential to improve the design and development of more engaging and effective multimedia content, ensuring a seamless and enjoyable experience for users in 360° contexts.

Index Terms: *360° video, Audiovisual Synchronization, Quality of Experience, Sense of Presence.*

Effective media synchronization is a pivotal aspect in the realm of multimedia management. It encompasses the meticulous preservation of temporal relationships within and between all media elements. As elucidated in a prior comprehensive study [1], any divergence in the presentation times of related media data objects across various media is commonly termed interstream skew. Achieving flawless synchronization across multiple media streams culminates in the absence of skew, denoting a perfect temporal alignment.

Despite the extensive exploration of AV skews in traditional multimedia [3], [4], [5], research concerning 360° videos remains limited. To address this intriguing and uncharted territory, we undertook a user study involving 360° videos subjected to varying degrees of skews. Unlike conventional multimedia, where AV skew typically has a low tolerance threshold and is readily noticeable by viewers, we hypothesized that the immersive nature of 360° videos might offer a unique advantage, in that AV coupling will be much looser than in traditional 2D multimedia. The all-encompassing visual experience in a 360° video could potentially mask certain audio deficiencies, allowing for a more generous margin for synchronization without significantly compromising the overall viewer experience. Accordingly, we undertook a study with two underpinning research questions:

- *RQ1. Do AV skews influence the QoE in 360° multimedia?*
- *RQ2. Do AV skews affect the sense of presence in 360° multimedia?*

As further elaborated upon in the following sections, our experimental setup included 360° videos

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/MMUL.2024.3415116 IEEE Multimedia

capturing three different scenarios, ranging from the tranquil scenario of a *Coffee Shop* to fast-paced action sequences of a *Kung Fu* scene. We applied the same set of AV skews to each video, introducing time delays between its visual and auditory components. Accordingly, this study has a further two research questions:

- *RQ3. Does motion dynamism of 360° multimedia videos influence the QoE?*
- *RQ4. Does motion dynamism of 360° multimedia videos affect the sense of presence?*

This paper is organized as follows: Section Related Works presents previous research concerning media synchronization; the Material and Methods (Section) details the questionnaires and devices used for this investigation; afterward, the Results and Discussion in Section presents our findings and engages in a comprehensive discussion; lastly, the Conclusion in Section synthesizes the main takeaways of our study and outlines directions for future work.

Related Work

Synchronization is crucial in multimedia systems. Unlike traditional single-medium systems that require synchronization within the same medium, multimedia systems with two or more media components face even more challenges because they need to synchronize between different media types. Concerning the latter, the focus of the study presented in this paper was pioneered by Steinmetz [6], who established the groundwork for exploring human media perception of skew, especially concerning AV content. Steinmetz [6] demonstrated that perceived synchronization could still be achieved despite skews between related data streams and provided conditions under which jitter might be acceptable. Leveraging these findings, he devised a method for handling nontrivial synchronization skew among more than two media streams.

Montagud et al [9] explore the domain of media synchronization providing definitions, classifications, and examples. It underscores the importance of various types of media synchronization in maintaining a satisfactory QoE and how it is associated with Quality of Service (QoS) magnifying the importance of synchronization types, offering a concise overview of the key aspects and components of media synchronization solutions.

Younkin et al. [7] conducted a study to determine the minimum amount of audiovisual synchronization (AV sync) errors that end-users can detect. Lip synchronization, the most noticeable AV sync error, was used as the testing stimuli to determine the perceptual threshold of audio-leading errors. The experiment results determined that the average audio-leading threshold for AV sync detection was 185.19 ms, with a standard deviation of 42.32 ms.

Khosravan et al. [8] discuss the importance of good synchronization between audio and video modalities in defining the quality of a multimedia presentation. The authors note that humans often pay close attention to discriminative spatiotemporal blocks of the video to judge whether audio and video signals of a multimedia presentation are synchronized. Inspired by this observation, the authors propose leveraging attention modules for automatically detecting AV synchronization. These neural network-based attention modules are capable of weighing different portions (spatiotemporal blocks) of the video based on their respective discriminative power. Experiments indicate that incorporating attention modules yields state-of-the-art results for the AV sync classification problem.

However, research has yet failed to explore the perceived AV sync quality when 360° content is presented to the user - the focus of our study, the details we now proceed to give.

Materials and Methods

Participants

We employed purposive sampling[1] to recruit a pool of 24 individuals, separated into two groups: an Experimental Group (EG) consisting of 15 users, which experienced AV skews, and Control Group (CG) of 9 users, which viewed the 360° multimedia content with no artificially generated AV skew.

Participants' ages ranged between 23 and 63 years. All had self-reported normal vision and were screened for contraindications (e.g., epilepsy, psychoactive drug treatment) for Virtual Reality (VR). The sample size and sampling method align with good experimental design practices [12]. Most participants were already familiar with the use of VR: participants' self-declared familiarity with VR varied: 9 participants (37.50%) were *Not Familiar* with VR, 12 participants (50.00%) were *Somewhat Familiar*, whilst 3 participants (12.50%) were *Very Familiar*. The study was approved by the Ethical Committee of Brunel University London Review Number 40020-LR-Oct/2022-41826-3.

[1] Purposive sampling is a form of non-probability sampling in which researchers rely on their judgment when choosing members of the population to participate in their research.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/MMUL.2024.3415116 IEEE Multimedia

Apparatus

Participants viewed the 360° content on a META QUEST 2 device, in which a Unity Application integrated with the Oculus App and connected with the META 2 via Oculus Link was used to control the sequence of videos and their skew. This setup ran on a desktop Dell Precision Tower 3620 PC with a quad-core Intel Core i7-7700 HQ, 16 GB RAM, 500 GB SSD with an AMD RX560, and a Dell Latitude 3490 laptop.

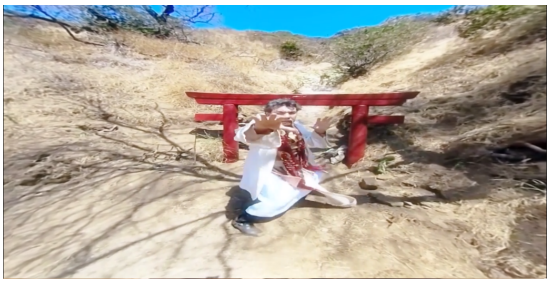
Materials

To address the two research questions of our study, three 360° videos of varying dynamism were selected, ranging from low to high, as proposed by Comşa et al. [10]. The *low dynamism Coffee Shop* (LD) video (Figure 1(a)) primarily shows the process of making coffee, accompanied by dubbed background sounds. Considering the straightforward audio context, we expected viewers would be less likely to notice AV skews in this video. At the opposite end of the spectrum, the *high dynamism Kung Fu* (HD) video (Figure 1(c)) features a sequence of fight scenes in First Person View (FPV) that requires meticulous AV synchronization to align sound effects with the visual movements on-screen. Here, we anticipated that any AV misalignment in this clip would be readily noticeable to the participants. Situated in the middle is the *medium dynamism Fireworks* (MD) video (Figure 1(b)), which contains the spectacle of a fireworks show. Given that the scene is set against a night sky, punctuated by the display of fireworks and their accompanying sounds, this video's dynamism is situated between the two mentioned scenarios. All videos were encoded at AAC (LC) (mp4a/0x6134706D), 48 KHz, stereo, fltp, 128 Kbps (default) and H.264 (High), and vp9.

To investigate the research questions of our study, we manipulated the audio tracks of the videos to introduce delays and hastening effects ranging from 1s to 5s. Each video was edited to a duration of 60s.

(a) *Coffee Shop* (Low Dynamism - LD)

(b) *Fireworks* (Medium Dynamism - MD)



(c) *Coffee Shop* (High Dynamism - HD)

FIGURE 1: Screenshots of each of the videos used

Experimental Protocol

Experiments were conducted in a dedicated room. Upon entering and before the start of the experiment, participants were questioned about any neurological or psychological disorders that might influence their responses to the videos. Participants were invited to sit on a swivel chair allowing them to spin and experience the 360° videos. Detailed information regarding the study's objectives and procedures was provided to ensure participants' understanding and informed consent. Participants adjusted the VR helmet till they were comfortable with its positioning. Participants were then presented with a sequence of videos. The presentation order of these videos was randomized as in Table 1 to mitigate order effects and to ensure a balanced experimental design for the EG. CG did not experience AV skews.

Research Instruments

After each viewing of a 360° video, participants were asked to complete two sets of questionnaires, each directly related to the research questions of our study. Accordingly, the first questionnaire targeted *RQ1* and *RQ3* aimed to assess the participants' QoE, providing valuable insights into their subjective perception and overall satisfaction with the video. The second, namely the SUS questionnaire, developed by Slater, Usoh, and Steed [11], aimed at the remaining two research questions, *RQ2* and *RQ4*.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI: 10.1109/MMUL.2024.3415116 IEEE Multimedia

	AV Skew					
	−5s	*−3s*	*−1s*	*+1s*	*+3s*	*+5s*
Participant 1	Coffee Shop (*LD*)	Fireworks (*MD*)	Kung Fu (*HD*)	Coffee Shop (*LD*)	Fireworks (*MD*)	Kung Fu (*HD*)
Participant 2	Kung Fu (*HD*)	Coffee Shop (*LD*)	Fireworks (*MD*)	Kung Fu (*HD*)	Coffee Shop (*LD*)	Fireworks (*MD*)
Participant 3	Fireworks (*MD*)	Kung Fu (*HD*)	Coffee Shop (*LD*)	Fireworks (*MD*)	Kung Fu (*HD*)	Coffee Shop (*LD*)

TABLE 1: Video presentation order for P1, P2, and P3 participants. The presentation order was then cyclically repeated for the remaining participants of the EG, for which a within subjects design was adopted.

QoE Questionnaire contains four questions that serve as a tool for collecting qualitative feedback, allowing us to gain insights into users' QoE and to understand their preferences and perceptions related to the 360° video content. Specifically, the questions are as follows.

- QoE-Q1: I enjoyed watching the 360° video.
- QoE-Q2: I noticed artifacts in the 360° video.
- QoE-Q3: I don't mind artifacts in the 360° video.
- QoE-Q4: Rate the overall quality of the 360° video.

The statements cover different aspects of the users' viewing experience. For the first three, participants were asked to give their responses on a five-point Likert scale, ranging from *(1) Strongly Disagree* to *(5) Strongly Agree*. The last statement was similarly coded, where each participant gave an overall rating on a 5-point scale ranging from *Very Bad* to *Very Good*.

SUS Questionnaire is a self-reported instrument for assessing the sense of presence induced by a virtual environment. Presence is the subjective experience of "being there" in a mediated environment, such as VR. The SUS is used in VR research and applications, as it is easy to administer and has good reliability and validity [11]. For our evaluation, the SUS was tailored to include the questions shown in Table 2.

SUS Questions and Rating Scales	
SUS-Q1	How much do you feel being in place, or being there, during the 360° videos? *(1) Not at all... (7) Very much.*
SUS-Q2	To what level were there times during the experience when the virtual environment was the reality for you? *(1) At no time... (7) Almost all the time.*
SUS-Q3	The 360° environment that you experimented reminded you of images and not a place that you were. *The 360° environment seems to me to be more like... (1) Images that I saw... (7) Somewhere that I visited.*
SUS-Q4	The 360° environment that you experimented reminded you of a place and not images. *I had a stronger sense of... (1) Being elsewhere... (7) Being in the 360° environment.*
SUS-Q5	How would you rate your sense of being in the virtual environment? *(1)Very Low... (7) Very High.*
SUS-Q6	How would you rate your sense of being elsewhere? *(1)Very Low... (7) Very High.*
SUS-Q7	How similar, in terms of the structure of the memory, is this to the structure of the memory of other places you have been today? *(1) Not at all... (7) Very much so.*
SUS-Q8	How often do you think to yourself that you were actually in the virtual environment? *(1) Not very often... (7) Very much so.*

TABLE 2: Adapted SUS questionnaire

Experimental Design

Our study had two independent variables (*Skew, Dynamism*) and two dependent variables (*QoE, Sense of Presence*), as measured by the individual items of the QoE and SUS questionnaires detailed previously. *Skew* had six possible values {−5s, −3s, −1s, +1s, +3s, +5s} and *Dynamism* three {low (LD), medium (MD), high (HD)}.

The study's sample size and participants per experimental condition (for both the EG and CG) are in line with similar studies reported elsewhere [1]. Moreover, the computed statistical power was 0.883, above the accepted 0.8 threshold for such calculations [12].

Results

Control Group

The quality of the videos was evaluated using a baseline established by the CG, which consisted of 9 participants. These participants did not experience AV skews during the experiment, providing the average measure of the QoE and SUS questionnaires absenting such skews.

CG ratings concerning the study's two questionnaires were then averaged to establish the baseline scores. The following table summarizes the baseline questions for each video:

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/MMUL.2024.3415116 IEEE Multimedia

Metric	LD	MD	HD	CG Average
Quality of Experience Questions (range: 1 to 5)				
QoE-Q1	3.88	4.11	4.33	4.11
QoE-Q2	3.22	3.77	3.11	3.37
QoE-Q3	3.11	2.88	2.66	2.88
QoE-Q4	3.55	3.66	3.88	3.70
System Usability Scale Questions (range: 1 to 7)				
SUS-Q1	4.22	4.33	5.00	4.51
SUS-Q2	4.00	3.88	4.22	4.03
SUS-Q3	3.33	2.55	3.44	3.11
SUS-Q4	5.00	4.55	4.44	4.66
SUS-Q5	4.33	4.33	4.44	4.37
SUS-Q6	4.44	3.44	4.55	4.14
SUS-Q7	4.11	4.88	4.66	4.55
SUS-Q8	4.77	4.11	5.44	4.77

TABLE 3: Control Group Results: Quality of Experience and System Usability Scale

We can observe that, for the CG, the *Kung Fu* (HD) video received the highest average ratings for both *QoE-Q1* and *QoE-Q4*, indicating that the participants found its levels of enjoyment and overall quality superior compared to the *Coffee Shop* (LD) and *Fireworks* (MD) videos. Also, with SUS scores, we see the same trend, with the *Kung Fu* (HD) video outscoring *Coffee Shop* (LD) and *Fireworks* (MD) videos concerning most SUS items.

Regarding overall average scores, *QoE-Q1* achieved a high score of 4.11, indicating a significant level of user satisfaction. Given that the CG did not experience any AV skews in the viewed videos, it is unsurprising that the average ratings for *QoE-Q2* and *QoE-Q3* were situated around the mid-point of the scoring scale, at 3.37 and 2.88 respectively. This suggests that for the baseline, participants noticed some artifacts, but generally did not perceive them as significant or negatively impactful. The overall average for *QoE-Q4*, while slightly lower than that for *QoE-Q1*, still received an average score of 3.70, higher than the midpoint of the scale. This score served as a reference point against which the video quality experienced by the EG (who did experience AV skews) could be compared and is now presented.

Experimental Group

For EG, data analysis was carried out by performing a two-way ANOVA (analysis of variance), with *Dynamism* and *Skew* as independent variables and items of the QoE and SUS questionnaires as dependent variables. The significance level adopted for the analysis was p=0.05.

The analysis revealed that only a subset of the questions demonstrated statistical significance. Specifically, in the QoE questionnaire, *QoE-Q1* (p<.05, F=2.282), *QoE-Q2* (p<.05, F=3.677) and *QoE-Q4* (p<.05, F=1.912) showed significant results. Interestingly, *QoE-Q2* and *QoE-Q3* were found to be significant exclusively for the *Kung Fu* video, characterized by high dynamism. In the case of the SUS questionnaire, *SUS-Q2* was the sole question that exhibited statistical significance (p<.05, F=2.254), with *SUS-Q1* also showing statistically significant results only for the *Kung Fu* video (p<.05, F=3.911).

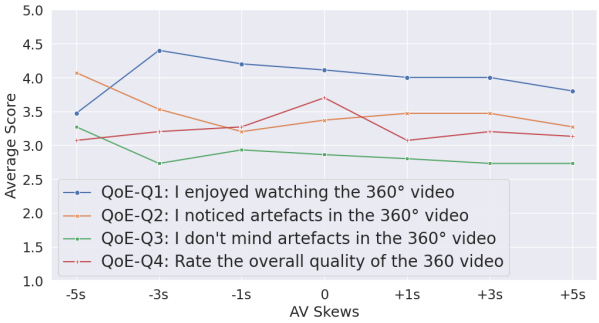


FIGURE 2: Impact of AV skews on mean QoE scores

We now seek to answer the four *Research Questions* introduced at the outset of this paper.

RQ1 - Do AV skews influence the Quality of Experience in 360° multimedia? In Figure 2, data revealed a pattern regarding the viewer's perception of the overall quality of 360° videos, as they perceived increasing average overall quality when the skew was approaching 0s and lower scores at extreme AV skews. The Discussion Section elaborates upon the presence of these outliers in LD and MD, but not in HD videos. Except for the *Fireworks* (MD) video, EG scores with AV skews were usually lower than the scores awarded by the CG. Furthermore, enjoyment scores declined as skew levels moved from the ideal 0s AV skew, except for *Coffee Shop* (LD) at −1s and +1s. Also, Figures 2 and Figure 3 highlight that *QoE-Q4* – targeting overall QoE – extreme AV skews led to lower scores for EG participants compared to their CG counterparts in the *Kung Fu* (HD).

Turning our attention to the perception of artifacts in the video, participants were likely to detect artifacts at a skew of −5s, as evidenced by the higher average scores for *QoE-Q2*. As the skews approached 0s, the average scores for noticing artifacts decreased slightly, suggesting that viewers were less likely to identify imperfections in the video when audio and video synchronization was closer to alignment. However, it is noteworthy that across all skew levels, the average rating for this statement remained higher than the mid-point scale score, for the CG and EG, indicating

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/MMUL.2024.3415116 IEEE Multimedia

that participants tended to notice some artifacts in the video, regardless of the degree of AV skew. This could be attributed to the question itself, which implies the presence of *artifacts* in the video, which could have led the participants to look for errors that were not present.

Further exploring attitudes towards artifacts in the 360° video, we found that participants generally displayed a tolerant disposition. As seen in Figure 3, the EG scores for the *QoE-Q3* statement ranged from 2.2 to 4.0 at different skew levels. However, as seen in Figure 2, there were no pronounced differences in the average results for any particular AV skew. This may suggest a willingness among viewers to overlook artifacts in favor of the overall immersive experience. The average score for this statement remained relatively consistent across skew levels, indicating a neutral stance on the issue of artifacts in the video. This high level of tolerance or neutrality towards these questions will be further addressed in the Discussion Section.

Finally, regarding *QoE-Q4* when EG participants were asked to assess the overall quality of the video with AV skews, viewers perceived the quality of videos as average with a mean score of 3.20, lower than the average overall quality of 3.7 rated by the CG. The average ratings remained steady across all levels of AV skew. Considering the impact of video content, Figure 3 shows mixed results across all skews, except for the *Kung Fu* (HD) that consistently decreasing scores for AV skews +1s, +3s and +5s.

RQ2 - Do AV skews affect the sense of presence in 360° multimedia? Compared to the CG, the EG exhibited a consistent pattern across most questions, indicating that the AV skews had minimal impact on the questionnaire items. This consistency suggests that, alongside our ANOVA analysis, the participant's experience within the 360° remained relatively stable, regardless of the level of AV skew applied.

As mentioned, within the context of the SUS questionnaire, *SUS-Q2* stood out as the only question that demonstrated statistical significance for the EG. Furthermore, *SUS-Q1* also indicated more pronounced results when considering the highly dynamic *Kung Fu* video, highlighting the need for a more extensive and in-depth data analysis.

Figure 4 presents the average scores for each statement in the SUS questionnaire, corresponding to the different levels of AV skews experienced by the participants. *SUS-Q1*, which measures the feeling of being in place during the 360° videos, shows a consistent average score of between 3.87 to 5.20 across different skew levels, except for a skew of −5s where the average score is slightly lower at 3.87, with

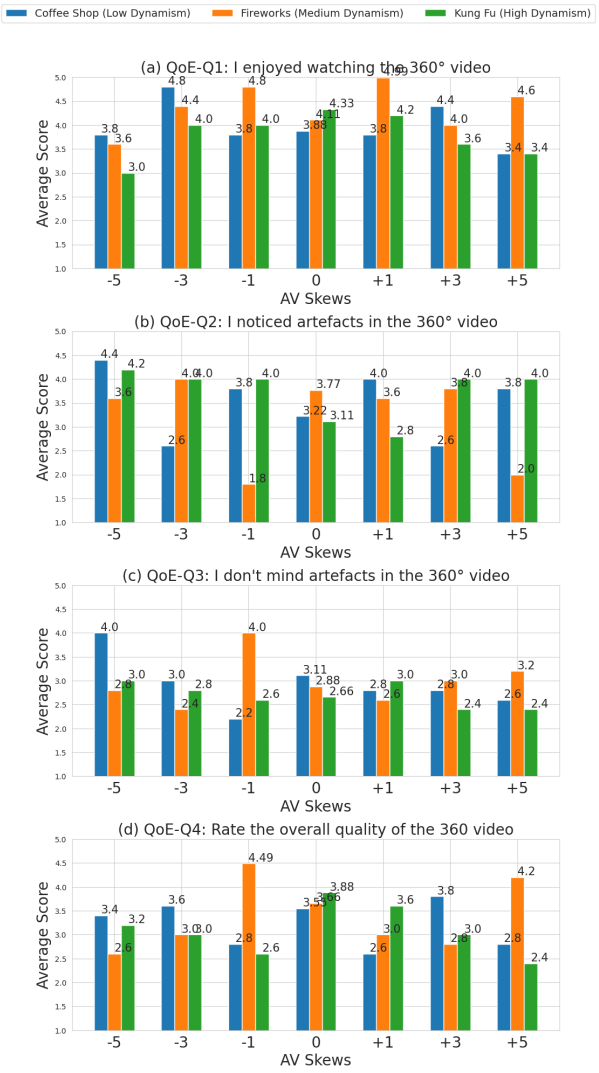


FIGURE 3: Mean QoE ratings according to skew levels and 360° video content

the overall lower values being present in the extreme audios skews both at −5s and +5s.

The remaining questions (*SUS-Q5* to *SUS-Q8*), which assess various aspects of the experience of participants in the virtual environment, show relatively consistent scores at different levels of AV skews, as depicted in Figure 4.

RQ3 - Does motion dynamism of 360° multimedia videos influence QoE? Figure 3 (a) reveals that, whilst for perfect AV synchronization of 0*s* participants rated *Kung Fu* (HD) better overall in terms of enjoyment, for other skew levels, *Coffee Shop* (LD) and *Fireworks* (MD) were most enjoyed, each for three skew levels respectively (−5*s*, −3*s* and +3*s* for *Coffee*

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/MMUL.2024.3415116 IEEE Multimedia

Shop (LD), and $-1s$, $+1s$ and $+5s$ for *Fireworks* (MD)). However, the *Fireworks* (MD) video emerges as the most enjoyable among the three videos if one computes an overall average score (including EG and CG) across all seven skew levels of our study. Specifically, *Fireworks* (MD) has an average overall enjoyment rating of *AVR* = 4.51 (compared to *AVR* = 4.1086 for both *Coffee Shop* (LD) and *Kung Fu* (HD)). It is also worth noting that the *Kung Fu* (HD) received the lowest enjoyment scores for the EG, except in $-1s$, $-1s$ where the *Coffee Shop* (LD) was the least enjoyable and for $5s$, in which *Coffee Shop* (LD) and *Kung Fu* (HD) scored jointly equal lowest scores, signaling that in the presence of AV skews, it was the least enjoyable for viewers in most cases. It was expected that the *Coffee Shop* (LD) was expected to return less pronounced results considering that the *Kung Fu* (HD) exhibited the highest sensitivity to enjoyability variations across different AV skew levels, with scores showing steady improvement as the skew values decreased, aligned with our expectations.

The graphs also illustrate mixed results across all videos concerning artifact detection, although the *Kung Fu* (HD) video garnered the highest artifact detection scores (QoE-Q2) for four of the seven skew levels employed in our study, seemingly indicating that viewers were relatively discerning when it came to identifying skews for this particular video of high dynamism. This aligns with our aforementioned observation that the *Kung Fu* (HD) video exhibited the highest sensitivity to variations in quality across different AV skew levels as this video has the highest AV coupling.

Furthermore, on average, viewers of the *Kung Fu* (HD) video seemed less forgiving of artifacts, with lower scores indicating reduced tolerance. In contrast, viewers of the medium and low dynamism *Fireworks* (MD) and *Coffee Shop* (LD) videos appeared to be more tolerant. This suggests that viewers' tolerance for artifacts may vary depending on the video content and the extent of the AV skew.

Lastly, viewers were asked to provide an overall quality rating for each video group. Figure 3 (d) reveals disparities in these ratings based on the different skew levels. The medium dynamism *Fireworks* video received higher overall quality ratings, indicating that it was perceived as the highest quality content in general, although CG rated *Kung Fu* (HD) as the overall better video. In contrast, the same high dynamism *Kung Fu* (HD) video received lower overall quality ratings with the EG, implying that it was perceived as the lowest quality in the presence of the AV skews. However, it scored highest over the other two videos for the CG. The *Coffee Shop* (LD) occupied an intermediate position, with moderate overall quality ratings.

RQ4 - Does motion dynamism of 360° multimedia videos affect the sense of presence? As Figure 4 highlights, in the context of the sense of presence in 360° multimedia, we observed that *SUS-Q1*, measuring the feeling of being in place during the videos, showed consistent scores across different skew levels. This suggests that the motion dynamism of 360° multimedia videos, as represented by AV skews, did not significantly affect the participants' sense of presence, as their scores remained relatively stable, even when compared to the scores from CG. This indicates that AV skews, at least within the range of variations explored, did not substantially impact the sense of presence in 360° multimedia.

The connection between AV skews and the virtual environment becoming a perceived reality was also evident. The statement *"To what level were there times during the experience when the virtual environment was the reality for you?"* demonstrated that participants reported varying degrees of immersion. At a skew of -5s, the average ratings were lowest at 2.93 and 3.13 at -5s, indicating a low and limited sense of the virtual environment being a reality in extreme AV skews. However, as AV skews decreased, with values closer to 0s, the ratings increased, peaking at -1s and +1s with an average rating of 4.20.

Moreover, the influence of AV skews on participants' sense of place versus images was apparent. "*The 360 environment that you experimented with reminded you of images and not a place that you were*" received higher ratings (mean = 4.20) when audio was delayed by -5s and 3.60 at +5s, higher than the other skews, indicating that more significant AV skews led to a perception of images rather than a sense of place. Conversely, with an AV skew level of +3s, participants gave an average rating of 2.73, suggesting that the smaller AV skews contributed to a stronger sense of place.

Discussion

This paper explored the relationship between AV skews and user experiences in 360° videos. Our analysis shed light on the multifaceted nature of this connection, offering valuable insights for content creators and researchers alike. As VR technology advances, a pressing need arises for a deeper comprehension of the interplay between different sensory stimuli and their synchronization to create a truly immersive and impactful virtual environment. By embracing the challenges and opportunities presented by integrating additional

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/MMUL.2024.3415116 IEEE Multimedia

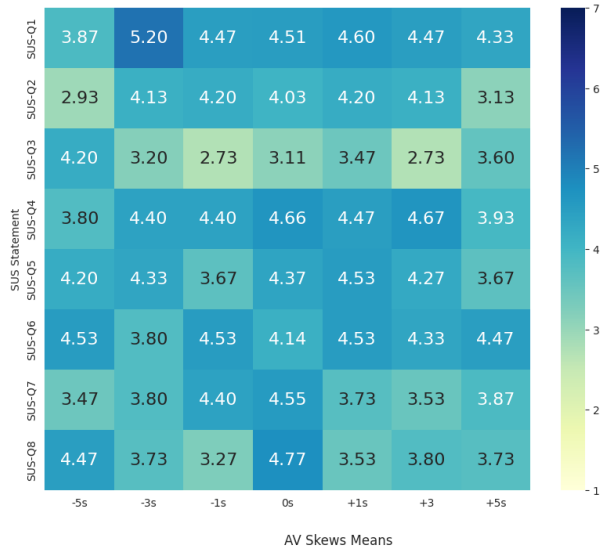


FIGURE 4: Heatmap of the Mean SUS Scores

sensory cues, researchers can develop innovative techniques and tools to create more compelling virtual experiences that engage users on multiple sensory levels.

Our study highlights a correlation between AV skew levels and participants' experiences in a virtual environment. We found that when the AV skews were low, the participants reported higher levels of presence, a stronger perception of the virtual environment as reality, and a greater sense of place. Interestingly, despite being exposed to varying levels of AV skews, viewers generally expressed relative enjoyment while watching the 360° videos. This suggests that the immersive aspect of the video can effectively mask these skews. Even when artifacts were more noticeable at extreme skew levels, viewers demonstrated tolerance, emphasizing the importance of an overall compelling experience. Therefore, despite the skews, the videos were still perceived as of satisfactory quality.

The immersive nature of 360° videos inherently encourages a more active viewer engagement, which could explain why AV skews were more tolerable in our study compared to previous research ([6], [7]). In traditional video formats, viewers are passive recipients of the content, making them more likely to notice discrepancies like AV skews. However, in 360° videos, viewers are participants in the virtual environment, actively exploring their surroundings. In our study, the influence of AV skews in 360° videos was less significant than we initially anticipated as the immersive nature of 360° videos encourages active exploration by viewers, which could potentially divert their attention away from any AV skews, shifting their focus away from any discrepancies between the audio and video.

The AV quality and video content also may have played a significant role. High-quality content could engross viewers to the point that they overlook any AV skews. The contents of 360° videos could potentially "mask" the effects of AV skews. For example, in the case of the *Fireworks* video, the inherent delay between the visual and auditory experience in real life could be a natural masking effect for AV skews. This is because fireworks are typically viewed from a distance, resulting in a noticeable time gap between the sight of the explosion and the sound reaching the spectator. This real-world phenomenon could make viewers more tolerant of AV skews in 360° videos. Due to those reasons, we believe and our data underlines that the immersive nature of 360° content and its masking effect suggest that the AV skew threshold identified in previous studies, such as those by Steinmetz [6] and Younkin and Corriveau [7] may be significantly larger in 360° videos compared to traditional 2D screenings.

Another aspect to consider in our study is the lack of 3D spatial audio in the videos. It plays a crucial role in enhancing the immersive experience of 360° videos by providing cues about the direction and distance of sound sources within the virtual environment. In our experiments, the videos did not incorporate 3D spatial audio. This could potentially have influenced the perception of AV skews by the participants. Without spatial audio, viewers might not have been able to accurately perceive the skews between the audio and video, reducing their sensitivity to AV skews. This lack of spatial audio could have further contributed to viewers' tolerance of skews and their overall satisfaction with the video quality, and further research is required in this respect.

In the context of AV skews and 3D spatial effects, further research is imperative to grasp the impact of 3D spatial audio on skew perception and overall video quality. Future studies could incorporate 3D spatial audio and compare its efficacy against traditional audio formats to determine whether it enhances users' perception of AV skews. These experiments should also explore whether the effectiveness of 3D spatial audio varies based on content type and degree of skew.

It should be noted that this lack of a significant impact of presence does not necessarily imply a lack of perceptual differences between the participants. However, while the AV skews did not lead to substantial changes in the SUS scores, they still subtly influenced the participants' experience, as revealed in the QoE scores.

However, it is also important to recognize that while our findings indicate an increased tolerance for AV

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/MMUL.2024.3415116 IEEE Multimedia

skews, there is a critical threshold beyond which the immersive benefits of 360° videos begin to wane. Excessive AV skews can cause noticeable discrepancies and viewer discomfort.

Conclusion

Our findings revealed that the impact of AV skews was not as conspicuous as we initially anticipated. Several factors may have contributed to this outcome. Firstly, the immersive nature of 360° videos allowed users to engage with their surroundings, diverting their attention from audio-video skews concerns. Participants were allowed to explore their virtual environments, turning their heads and focusing on various visual elements, which diminished the significance of AV skews.

While AV skews remain noteworthy in 360° video content creation, our study suggests that its impact may be less pronounced than one would assume *a priori*. Designers and researchers should consider the dynamic nature of user interaction within 360° environments and user engagement duration to optimize the overall user experience.

Regarding the fidelity of the virtual world to the real one, the content of 360° videos can influence user tolerance to AV skews. For example, real-world phenomena, like the delay in sound from distant sounds, can also mask AV skews, making them less noticeable. These factors suggest that viewer engagement in 360° videos may be less affected by AV skews than initially assumed.

Also, as discussed, the absence of 3D spatial audio in our study's videos may have influenced the participants' perception of AV skews. A more immersive experience provided by 3D spatial audio, offering cues about the direction and distance of sound sources, could potentially enhance the viewers' sensitivity to AV skews. This study suggests that the lack of spatial audio might have increased viewers' tolerance of skews and overall satisfaction with the video quality. Future studies should incorporate 3D spatial audio and compare its effectiveness with traditional audio formats. Additionally, these studies should investigate whether the efficacy of 3D spatial audio varies based on the content type, audio dynamism, and the degree of skew, as the comparison between spatial audio and non-spatial audio setups is a promising area for future research.

This finding has significant implications, especially in 360° videos, virtual reality, and augmented reality applications. By understanding the optimal levels of AV skews tolerated in 360° videos without compromising user QoE, creators can enhance their content production processes and explore new creative possibilities.

ACKNOWLEDGMENTS

This study was financed partly by the CAPES, Brazil – Finance Code 88887.570688/2020-00 and 88881.689984/2022-01; CNPQ, Brazil – Finance Code 307718/2020-4; and FAPES, Brazil – Finance Code 2021-GL60J.

REFERENCES

1. N. Murray, Y. Qiao, B. Lee, A. K. Karunakar, and G.-M. Muntean, "Subjective evaluation of olfactory and visual media synchronization". In MMSys '13 - 4th ACM Multimedia Systems Conference, ACM, 2013, pp. 162–171. doi: 10.1145/2483977.2483999.
2. S. Van Damme, M. T. Vega, and F. De Turck, "Human-centric Quality Management of Immersive Multimedia Applications". In NetSoft 2020 - 6th IEEE Conference on Network Softwarization, IEEE, 2020, pp. 57-64. doi: 10.1109/NetSoft48620.2020.9165335.
3. C. Abry, M.-T. Lallouache, and M.-A. Cathiard, "How can coarticulation models account for speech sensitivity to audiovisual desynchronization?". Speechreading by Humans and Machines: Models, Systems, and Applications, 1996, pp. 247-255. doi: 10.1007/978-3-662-13015-5_19.
4. N. Murray, B. Lee, Y. Qiao, A. K. Karunakar, and G.-M. Muntean, "Multiple-scent enhanced multimedia synchronization". ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), vol. 11, no. 1s, pp. 1-28, 2014. doi: [10.1145/2637293].
5. E. Hagman, "audiovisual Desynchronization Impact on Listening Effort: How does Audio Delayed to Visuals Affect Listeners' Effort to Understand a News Reporter in a Noisy Background?" 2015.
6. R. Steinmetz, "Human perception of jitter and media synchronization". IEEE Journal on Selected Areas in Communications, vol. 14, no. 1, pp. 61-72, 1996. doi: 10.1109/49.481694.
7. A. C. Younkin and P. J. Corriveau, "Determining the amount of audio-video synchronization errors perceptible to the average end-user". IEEE Transactions on Broadcasting, vol. 54, no. 3, pp. 623-627, 2008. doi: 10.1109/TBC.2008.2002102.
8. N. Khosravan, S. Ardeshir, and R. Puri, "On Attention Modules for audiovisual Synchronization". In CVPR Workshops, 2019. doi: 10.48550/arXiv.1812.06071.
9. Montagud, M., Cesar, P., Boronat, F., Jansen, J. (2018). Introduction to Media Synchronization (MediaSync). In: Montagud, M., Cesar, P., Boronat,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/MMUL.2024.3415116 IEEE Multimedia

F., Jansen, J. (eds) MediaSync. Springer, Cham. https://doi.org/10.1007/978-3-319-65840-7-1

10. I. -S. Comşa, E. B. Saleme, A. Covaci, G. M. Assres, R. Trestian, C. A. S. Santos, and G. Ghinea, “Do I Smell Coffee? The Tale of a 360° Mulsemedia Experience”. IEEE MultiMedia, vol. 27, no. 1, pp. 27-36, Jan.-March 2020. doi: 10.1109/MMUL.2019.2954405.
11. M. Usoh, E. Catena, S. Arman and M. Slater, "Using Presence Questionnaires in Reality". Presence, vol. 9, no. 5, pp. 497-503, Oct. 2000, doi: 10.1162/105474600566989.
12. J. Greene and M. D'Oliveira, "Learning To Use Statistical Tests In Psychology", McGraw-Hill Education (UK), 2005.

Aleph Campos da Silveira is a Ph.D. student in Computer Science at the Federal University of Espirito Santo, Vitória-ES, Brazil, 29932-540, Brazil. His research interests include Usability Evaluation, Multisensory Human-Computer interaction, Biofeedback, and Games. Silveira holds a BSc degree in Computer Science and Information Systems and an MSc in Education, both from the Federal University of Lavras. Contact him at alephcampos@gmail.com.

Fotios Spyridonis is a Lecturer in Computer Science at Brunel University London, Uxbridge, UB8 3PN, United Kingdom. His research interests include Interactive Multimedia and Human-Computer Interaction. He received his Ph.D. in Computing from the Department of Computer Science at Brunel University London. Contact him at fotios.spyridonis@brunel.ac.uk.

Roope Raisamo is currently a Professor of Computer Science at Tampere University, Tampere, FI-33014 Finland. His research interests include haptic interaction, computer vision, and multimodal systems. He received his Ph.D. in computer science from the University of Tampere. Contact him at roope.raisamo@tuni.fi k

Alexandra Covaci is currently a Senior Lecturer in Digital Arts and Technology at the University of Kent, Kent, CT2 7NZ United Kingdom. Her research interests are centered around the transformative power of VR from training to social scenarios and lie at the confluence of virtual reality, multisensory media, human-computer interaction, and psychology. She received her Ph.D. in virtual reality from Transilvania University of Brasov. Contact her at A.Covaci@kent.ac.uk

Gheorghita Ghinea is a Professor of Mulsemedia Computing at Brunel University London, Uxbridge, UB8 3PN, United Kingdom. His research interests lie at the confluence of computer science, media, and psychology, focusing on building adaptable end-to-end communication systems incorporating user multisensorial and perceptual requirements. He received his Ph.D. degree in Computer Science from the University of Reading, U.K. Contact him at George.Ghinea@brunel.ac.uk

Celso Alberto Saibel Santos is a Professor in the Department of Informatics at the Federal University of Espirito Santo, Vitória-ES, Brazil, 29932-540. His current research interests include Multimedia Systems, Computing Systems Engineering, and Multisensory Applications. He received his Ph.D. in Informatics from Université Paul Sabatier de Toulouse III. Contact him at saibel@inf.ufes.br.