Immersive Networked Music Performance: Impact of Extended Reality on the Quality of Experience

1st Robert Hupke
Leibniz University Hannover
Hannover, Germany
robert.hupke@sennheiser.com

2nd Stephan Preihs
Leibniz University Hannover
Hannover, Germany
preihs@ikt.uni-hannover.de

3rd Jürgen Peissig
Leibniz University Hannover
Hannover, Germany
peissig@ikt.uni-hannover.de

***Abstract*—The potential of integrating extended reality (XR) and spatial audio into immersive networked music performance (INMP) systems has not yet been the subject of extensive research. The advantages of a visual representation based on the transmission of position-dependent metadata seem to be evident, not just in terms of lower transmission latency. It enables the transformation of a virtual environment with spatial audio into an extended reality environment (XRE), simulating the experience of making music together in the same room. A rhythm-based experiment was conducted with 18 pairs of participants, using several XREs of the Immersive Room ExtensioN Environment (IRENE) system, an immersive real-time system combining virtual reality with telepresence and providing spatial audio for an enhanced immersive experience. The experiment evaluated the influence of the XR system on participants' perceived quality of experience (QoE) and musical outcome. Results show a clear trend that the XREs of the IRENE system enhance participants' perceived coherence and immersion, as well as the overall perceived QoE, while having a minimal impact on the musical outcome. This study provides valuable insights into the benefits and challenges of using XR in NMP, contributing to a deeper understanding of how immersive technologies can improve networked music collaboration.**

***Index Terms*—Immersive Networked Music Performance, Extended Reality, Virtual Reality, Spatial Audio, Quality of Experience**

I. Introduction

Although the latency of video transmission in most networked music performance (NMP) systems is significantly higher compared to the audio transmission, the use of videoconferencing tools is one of the most common methods in today's NMP systems [1]–[4]. While it has been shown that the visual information can be even more informative than the auditory reproduction for the perceiver's understanding of the performer's expressive intention [5], the delay difference between audio and video can cause the musicians to focus primarily on the audio transmission and to ignore the visual content [6]. Several findings from experiments support the assumption that musicians do not look [7] or rarely use [8] the video link when they perform, and that the use of small video displays is inadequate to convey visual cues [9]. However, there is evidence that body movements help regulate coordination in ensemble performance [10], and that musicians in ensembles rely on peripheral vision and rarely use direct gaze [11]. Nevertheless, traditional videoconferencing systems seem to support direct gaze. Considering these findings and the low data transmission requirements, the use of virtual reality (VR) and augmented reality (AR) solutions for distributed music performances appears promising. Consequently, a framework that includes a suitable visual platform using motion capture (MoCap) skeleton data might have the potential to serve as an efficient alternative to traditional video transmission in NMP scenarios.

The Immersive Room ExtensioN Environment (IRENE) [12] aims to extend the concept of a "meeting place" within virtual environments (VE) to a local space by introducing a "virtual window". This integration of VR with telepresence is designed to provide a realistic representation of a remote virtual space, potentially addressing challenges associated with VR, such as motion sickness. Furthermore, the system offers solutions for maintaining perceived local presence, which might be an issue in fully VR NMP systems involving both remote and onsite musicians [13]. Additionally, the IRENE system provides spatial audio to achieve an immersive perception of sound sources and room characteristics of the VE. Since transmission latency remains the dominant factor influencing interaction in NMP systems [14], additional audio signal processing can limit the effectiveness of such immersive networked music performance (INMP) systems, especially in comparison to well established low-latency audio transmission systems [15]. Therefore, an INMP system must offer substantial benefits beyond the audio transmission.

In the context of the Internet of Musical Things (IoMusT) [16], which extends the Internet of Things paradigm to the musical domain, INMPs emerge as an essential component [17], [18]. Results from several studies using spatial audio, both without visual display and in combination with telepresence, have shown varying outcomes. While a non-reverberant condition was preferred for clearer note onsets in drum performances [19], it was found in [20] that a reverberant VE improved the precision and ensemble playing. Spatial separation of sound sources can enhance cognitive attention and auditory scene partitioning [21], [22], providing more space for directing the musicians' cognitive attention where it is needed. It was shown in [23] that stereo source panning between remote musicians and a local metronome click can improve tempo stabilization, especially at higher delay times in rhythmic NMP contexts. It was shown in [24] that binaural auralization of virtual acoustic environments

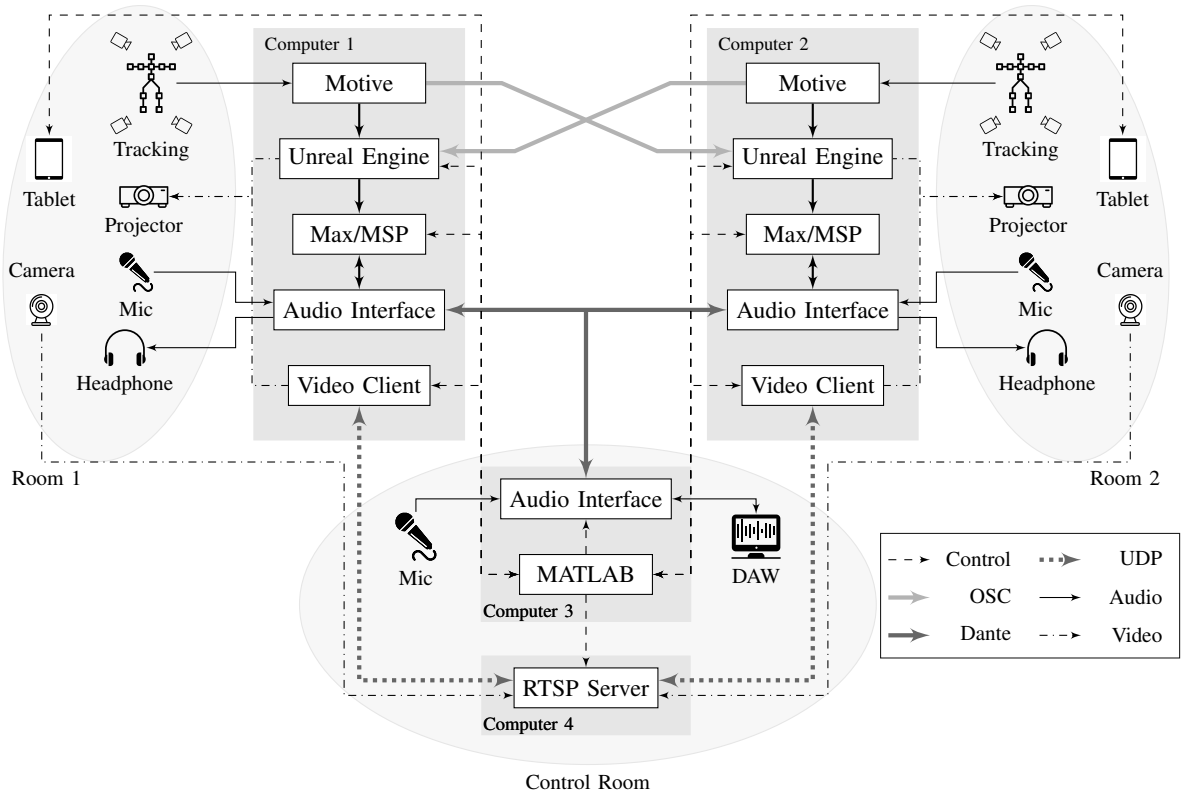


Fig. 1. Experimental setup and data flow of the experiment using the IRENE system.

enhances the subjective experience in NMP scenarios. The simulation of a virtual music studio showed that immersive audio may influence the experience with respect to factors such as perceived presence and localization, and may also affect the perception of latency [25], [26]. In [27], a study involving upper body tracking for avatar representation in an INMP experience within XR was conducted, revealing that musicians preferred personalized avatars over generic versions. A quality of experience (QoE) evaluation demonstrated that immersive audio rendering was considered beneficial for singer monitoring, particularly in achieving clarity, source separation, and providing a sense of virtual space [28]. In [29] it was demonstrated that musicians prefer playing their instruments with binaural spatialization and head tracking in remote collaborative music scenarios using NMP systems. Overall, the impact of spatial audio in NMP remains somewhat unclear and appears to be influenced by specific use cases and technical implementations. Particularly, the potential benefits of spatial audio in combination with a VE in INMP have been little explored.

This contribution presents one part of a larger experiment with the aim to evaluate the impact of the IRENE system on the participants' perceived QoE and their objective musical outcome. Section II describes the framework of the IRENE system, the experimental setup and procedure to evaluate the impact of the IRENE system and its different VEs. Specifically, participants were tasked with performing a complementary rhythm within three different XREs with spatial audio including a free-field, a concert hall, and a listening room scenario, compared to a traditional videoconferencing scenario as baseline. Results are shown in III for measurable objective metrics, such as tempo stability and rhythmical alignment, and subjective evaluations on the quality of interaction and perceived QoE within the VEs under investigation. The findings are finally discussed in Section IV.

II. Experiment

A. Setup

The IRENE setup, described in [12], was adapted to the existing infrastructure of the Institute of Communications Technology (IKT) at the Leibniz University Hannover, with its data flow shown in Figure 1. Small changes were made to the audio rendering for optimization. An RTSP server/client infrastructure was integrated to provide low-latency video transmission. Two tablets were included for questionnaire assessment, as well as an audio interface that enabled the recording of individual audio tracks. The experimenter, located in a separate control room, managed all interfaces and parameters via MATLAB, maintaining contact with participants through a talkback microphone.

For the part of the experiment presented in this publication, three XREs with binaural audio were used and evaluated: a free-field condition with only the direct sound path and one reflection of the floor (VE_{Free}), a concert hall with a mean

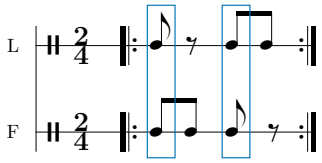


Fig. 2. Complementary rhythm consisting of one quarter and two eighth notes to define two synchronization points (blue) per beat (adapted from [30]).

reverberation time of $T30 = 2.05\,\mathrm{s}$ ($\mathrm{VE_{Hall}}$), and a virtual representation of a listening room with a mean reverberation time of $T30 = 0.46\,\mathrm{s}$ ($\mathrm{VE_{IML}}$). Each XRE had individualized room acoustics and binaural rendering depending on the relative positions of the participants. Additionally, a visual baseline condition was included: a traditional videoconferencing display (*Baseline*) with monaural audio playback. Pictures of the four environments can be seen in Figure 3. Omnidirectional lavalier microphones were used for recording monaural audio signals with an analog wireless system ensuring low latency and participants' freedom of movement. Open headphones (Sennheiser HD650) provided audio playback, allowing participants to hear their own sound while wearing headphones.

The setup ensured minimal latency, with monaural signals routed via Max/MSP or directly to the audio interface output. A minimum I/O audio delay of approximately $11\,\mathrm{ms}$ was maintained, along with an additional artificial delay for a more realistic NMP scenario. In Table I, the measured transmission latencies for the different audio and video conditions are shown. The procedure of measuring the Action to Screen (A2S) latency is described in [12]. Raw audio signals and the two headphone signals were recorded in a DAW for later analysis. MoCap tracking was used to capture participants' poses and movements accurately. Seven tracking cameras were positioned around the projection screen in each room to ensure optimal coverage, with tracking data streamed directly from Motive to MATLAB and synchronized with the corresponding audio data in Max/MSP. For video integration, the real-time streaming protocol (RTSP)[1] was used for low-latency streaming, with an OSC interface in UE for switching between the different experimental conditions.

B. Participants

A total of 36 individuals participated in the study, resulting in 18 separate experimental sessions. Participants were recruited from among the audio-enthusiast students and employees of the IKT. External participants were also recruited, as long as they had at least five years of musical background and an interest in the experience of musical XR. At the beginning of the experiment, participants filled out a consent form, which included collecting data on their age and musical background. The group of participants consisted of 6 females and 30 males, with a mean age of 29 years and an average of 11 years of musical experience.

[1]https://www.ffmpeg.org/

C. Procedure

The participants were tasked with clapping the complementary rhythm depicted in Figure 2. The experiment procedure comprised several stages. Initially, a comprehensive introduction familiarized participants with the facilities and technology used, and detailed explanations of the questionnaire ensured consistent understanding. Participants rehearsed the rhythm together, with fixed roles of *Leader* and *Follower* at the beginning of each trial. Then, participants put on their MoCap suits, including shoes, cap, gloves, microphone, and headphones, followed by a skeleton calibration in the tracking software (Motive) and a soundcheck. After this, the participants had a training phase of a minimum of two minutes for each condition to familiarize themselves with the various XREs.

At the beginning of each trial, participants had the chance to explore the randomly chosen XRE for 60 seconds before starting the performance. Then, the *Leader* initiated the trial while hearing a metronome click, which was manually turned off after the *Follower* joined in when ready. The performance was recorded for a minimum of 45 seconds. After each trial, the participants completed a questionnaire on their tablets using QUEST [31]. While answering the questions, the XRE was deactivated and the participants were no longer able to communicate with each other. This process was repeated for all experimental trials, with additional pauses for re-calibration if necessary to ensure an accurate avatar representation.

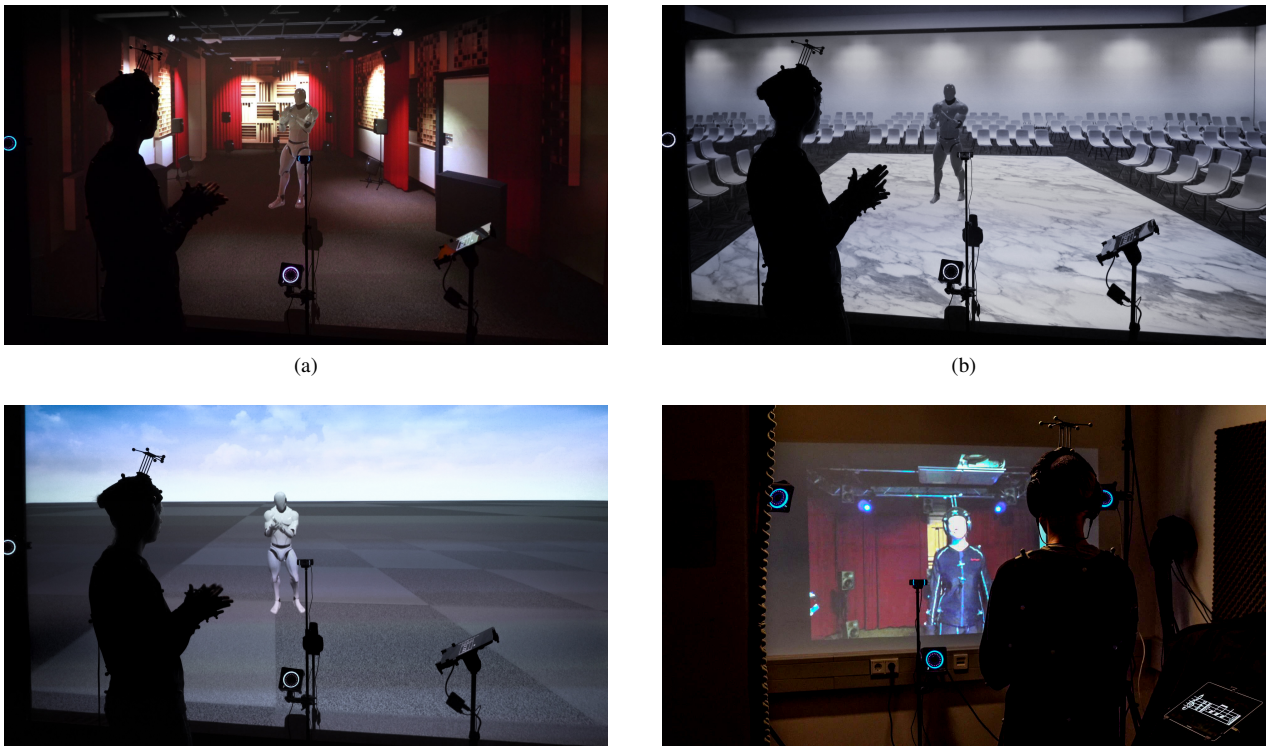
D. Questionnaire

While a number of self-report questionnaires have been developed to assess participants' perceived experience and social presence, there is no validated inventory for evaluating the perceived QoE in NMP. The majority of research studies in the field of NMP rely on the use and subsequent analysis of such pseudo-standard questionnaires [1], [4], [29], [32]. The questions listed in the questionnaire are also based on similar constructs helping to identify trends in the overall QoE as perceived by the participants. The questions can be categorized into three main parts. The first set (1-4) assesses participants' perceived coherence and immersion. The second set (5-8) evaluates the perceived quality of the XREs, including

TABLE I
SUMMARY OF VISUAL DISPLAYS AND AURALIZATION MODES WITH THEIR RESPECTIVE MEASURED LATENCY.

Acronym	Description	Latency [a]
VE_Free	VE of Free-field	74 ms
VE_Hall	VE of Concert hall	74 ms
VE_IML	VE of IML	74 ms
Baseline	Video transmission	144 ms
	System Latency[b,d]	Remote Latency[b,c]
Binaural	11 ms	21 ms
Monaural	700 µs	10 ms

[a] Measured Action to Screen (A2S) latency, [b] with a buffer size of 128 samples and a sample rate of 48 kHz , [c] OWD with additional added transmission latency, [d] participants self-delay.



(a) (b) (c) (d)

Fig. 3. Pictures of the four XR environments under investigation of the IRENE system: (a) VE_{IML}, (b) VE_{Hall}, (c) VE_{Free}, and (d) Baseline (Video).

auditory and visual aspects. The final set (9-10) addresses the overall quality of the performance conditions and identifies any inaccuracies or glitches in the visual representation.

Participants rated questions 1-8 on a continuous, underlying scale from -5 to 5 with a step size of 0.1. The score of -5 was marked as "strongly disagree", 0 as "neutral", and 5 as "strongly agree". Question 9 and 10 were rated on a 5-point scale (very poor, poor, neutral, good, excellent / strongly disagree, disagree, neutral, agree, strongly agree).

1. Consistency: *The virtual performance space was very similar to reality.*
2. Social Presence: *I felt like I was in the same room with my remote counterpart.*
3. Immersion: *I felt being involved in the virtual performance space.*
4. Naturalness: *My musical interaction in the virtual performance space felt natural.*
5. Responsiveness: *The virtual performance space was acoustically responsive to sounds I initiated.*
6. Localization: *I was able to localize my musical counterpart.*
7. Separation: *I was able to distinguish between the sounds I initiated and the sounds produced by my musical counterpart.*
8. Delay: *I experienced delay between my actions and the expected outcome.*
9. *How would you rate the performance conditions during your performance?*
10. *My performance was not affected by inaccuracies or glitches in the visual display.*

III. Results

The recordings were evaluated using the objective metrics *asymmetry*, *imprecision*, *tempo slope*, and *pacing*. All metrics required both onset detection and assignment of the onset to the notes in the recorded rhythm. An energy-based approach, similar to the one used in [33], was used for onset detection. For a detailed explanation of the onset detection method, refer to [34]. The metrics used are explained in detail in [14] – therefore, only a brief explanation is provided here:

- *Asymmetry* measures the deviation of the onsets between notes or beats and indicates how consistently a rhythm is maintained by the musicians.
- *Imprecision* quantifies the variability in the timing of note onsets, indicating the accuracy of the musicians' performance.
- *Mean Tempo Slope* indicates the change in tempo over the piece (sometimes only over the beginning of the piece), showing whether the performance has a tendency to accelerate or decelerate.
- *Pacing* measures the overall tempo of the performance, enabling a comparison of the initial starting tempo and the actual tempo maintained by the musicians.

The results of the metrics were analyzed with linear mixed-effect models. Each metric had the musical XRE as a fixed

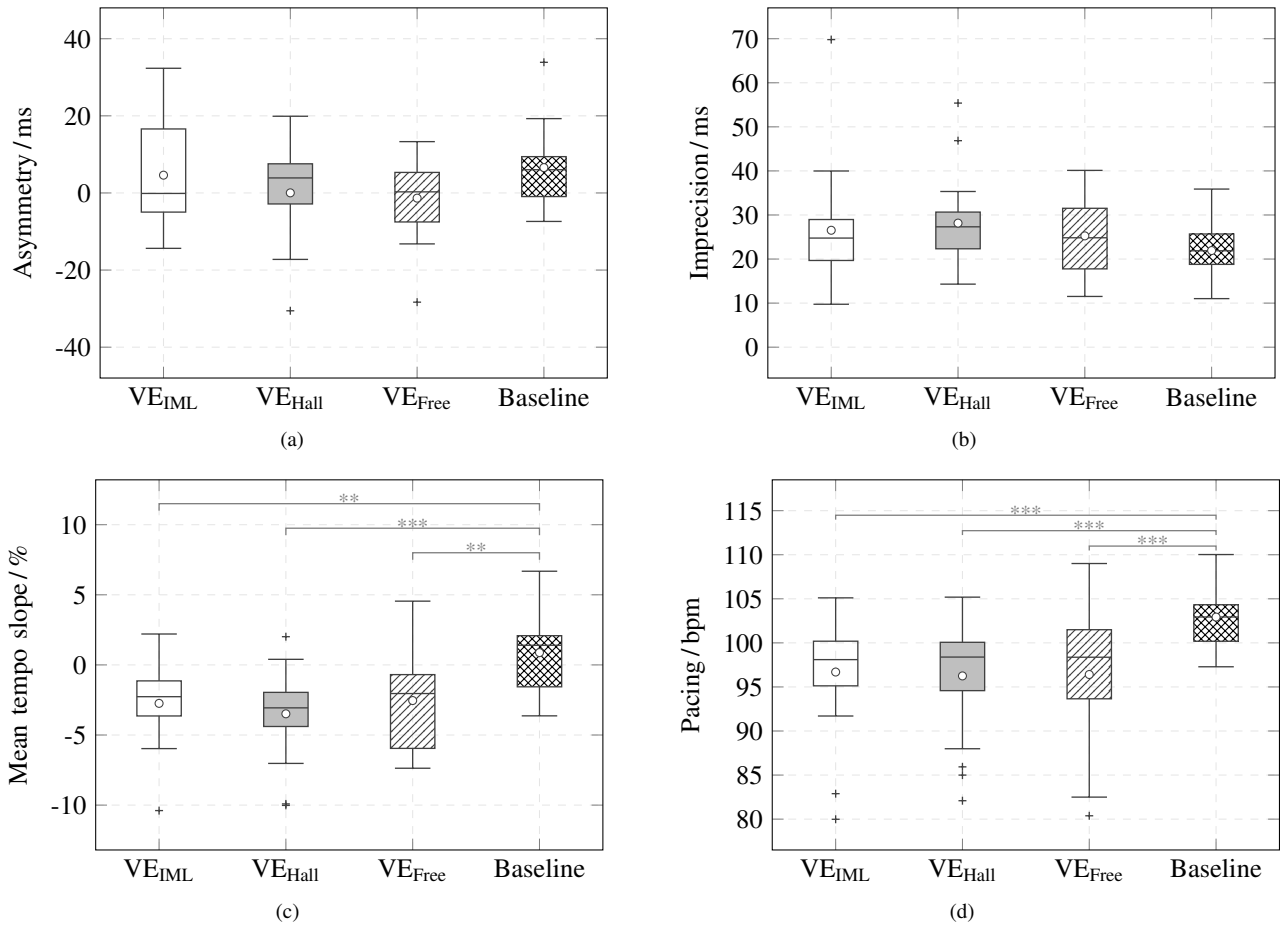


Fig. 4. Results of the objective metrics *asymmetry* (a), *imprecision* (b), *mean tempo slope* (c), and *pacing* (d) for the three virtual environments of the IRENE-system in comparison with the monaural video representation as baseline. The ○ shows the respective mean value. *, **, and *** indicate a significance level at $p < 0.05$, $p < 0.01$, and $p < 0.001$, respectively.

factor, while the participants were included as random factors. Pairwise comparisons with Tukey's correction were performed for the fitted model by post-hoc tests if the main factor showed a significant effect. The same process of analysis was applied to each question of the questionnaire.

A. Objective Results

The results of the metrics *asymmetry*, *imprecision*, *tempo slope*, and *pacing* are shown in Figure 4. The metrics *asymmetry* and *imprecision* indicate that the VEs under investigation show no significant differences when compared to each other or to the baseline, indicating that the VEs had no noticeable effect on the rhythmical alignment of the musicians' musical outcome.

The metric *mean tempo slope* shows a significant effect of the different XREs as a main effect ($F(3, 45.296) = 8.048, p < 0.001$). Pairwise comparisons indicate significant differences between the baseline and each of the three VEs VE_{IML} ($p < 0.01$), VE_{Hall} ($p < 0.001$), and VE_{Free} ($p < 0.01$). Specifically, the baseline shows a slight acceleration of the initial playing tempo. In contrast, the three virtual environments show a deceleration of the initial playing tempo. Nevertheless, no significant differences between the virtual environments were found. Considering the mean scores, a tendency is observed in which VE_{Hall} appears to cause the highest tempo deviation, followed by VE_{IML} and VE_{Free}.

A similar effect is observed when considering the metric *pacing*. The results show that the XREs have a significant effect on the metric *pacing* ($F(3, 44.979) = 10.21, p < 0.001$). Pairwise comparison highlights differences between the baseline and each of the three virtual environments VE_{IML} ($p < 0.001$), VE_{Hall} ($p < 0.001$), and VE_{Free} ($p < 0.001$). For the baseline, the mean tempo *pacing* is higher than the initial playing tempo, while it is lower for the three VEs under investigation, with minimal differences in their respective mean ratings.

B. Subjective Results

The subjective results for the questions 1-8 are presented in Figure 5 for all three VEs and the baseline.

1) Consistency: No significant main effect was found. However, there might be a trend towards higher mean scores for VE_{IML} and VE_{Hall} compared to VE_{Free} and the baseline.

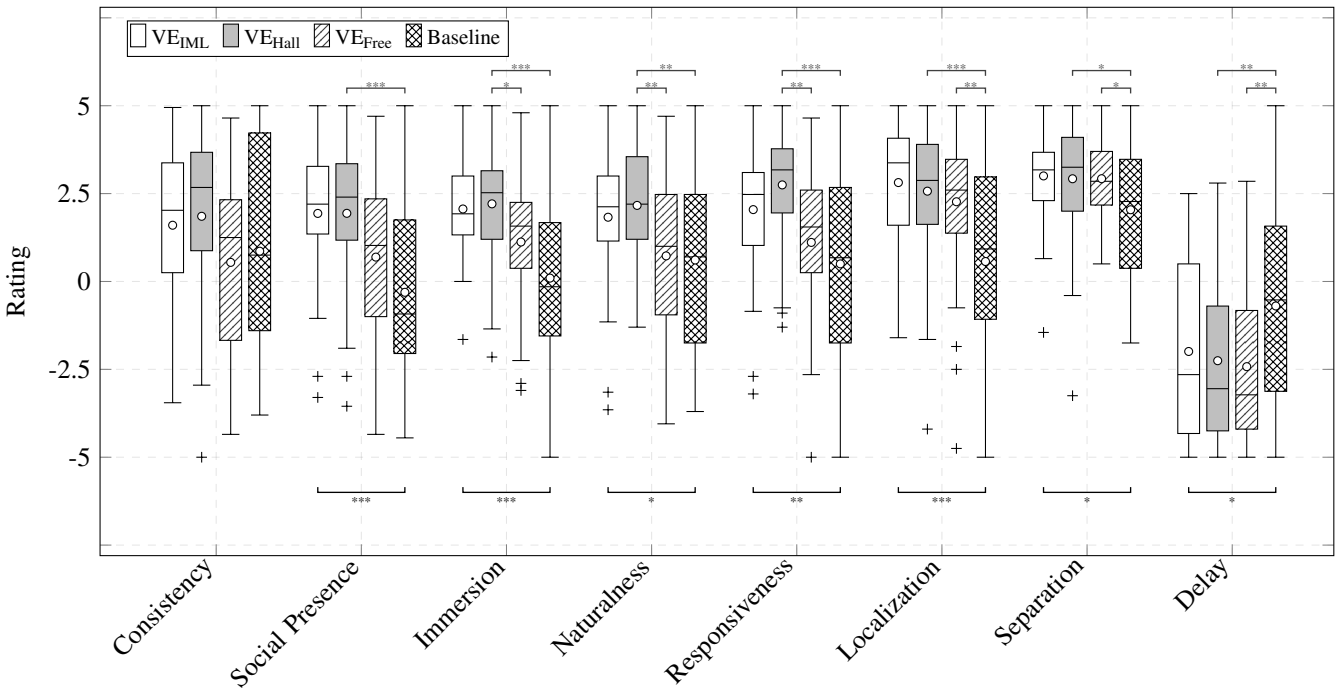


Fig. 5. Results of subjective ratings for *Responsiveness* (a), *Localization* (b), *Separation* (c), and *Delay* (d). The ∘ shows the respective mean value. *, **, and *** indicate significance at $p < 0.05$, $p < 0.01$, and $p < 0.001$, respectively.

2) Social Presence: A significant effect was found for the main effect of the XREs ($F(3, 105) = 9.819, p < 0.001$). Pairwise comparison post-hoc test of the four XREs resulted in significantly higher scores for VE_{IML} and VE_{Hall} in comparison to the baseline ($p < 0.001$). No significant differences can be observed between the respective XREs, as well as between VE_{Free} and baseline. However, there might be a tendency that VE_{IML} and VE_{Hall} were rated higher in terms of social presence than VE_{Free}, which was rated similarly like the baseline, showing a tendency toward a neutral rating of social presence.

3) Immersion: A significant main effect was found for the XREs ($F(3, 105) = 11.56, p < 0.001$). A pairwise comparison post-hoc test indicates that participants perceived VE_{IML} and VE_{Hall} significantly higher immersive than the baseline ($p < 0.001$). No significant differences were found between VE_{IML} and VE_{Hall}, while VE_{Hall} received a significantly higher rating than VE_{Free} ($p < 0.05$).

4) Naturalness: A significant main effect was found for the XREs ($F(3, 105) = 6.687, p < 0.001$). VE_{Hall} was rated significantly higher than VE_{Free} and the baseline ($p < 0.01$). There was no significant difference between VE_{Hall} and VE_{IML}. VE_{IML} was rated significantly more natural than the baseline.

5) Responsiveness: A significant main effect was found for the XREs under investigation ($F(3, 105) = 10.68, p < 0.001$). Pairwise comparison post-hoc test resulted in significantly higher scores for VE_{IML} and VE_{Hall} in comparison to the baseline ($p < 0.01$, $p < 0.001$, respectively). VE_{Hall} received significantly higher perceived responsiveness than VE_{Free} ($p < 0.01$).

6) Localization: The results show a significant main effect ($F(3, 105) = 10.09, p < 0.001$). The three XREs – VE_{IML}, VE_{Hall}, and VE_{Free} – were rated significantly higher than the baseline ($p < 0.001$, $p < 0.001$, $p < 0.01$, respectively). No significant differences were perceived between the three VEs.

7) Separation: A significant main effect was found ($F(3, 105) = 3.933, p < 0.05$). The three VEs – VE_{IML}, VE_{Hall}, and VE_{Free} – were rated significantly higher for perceived separation than the baseline ($p < 0.05$).

8) Delay: A significant main effect was found ($F(3, 105) = 6.021, p < 0.001$). The perceived delay between participants' actions and the expected outcome was rated significantly higher for the baseline than for the XREs VE_{IML}, VE_{Hall}, and VE_{Free} ($p < 0.05$, $p < 0.01$, $p < 0.01$, respectively).

9) General Questions: Since the last two questions were rated on a 5-point scale, the results are presented in Figure 6 as percentage stacked bar plots for a relative comparison of the ratings. The results for the question regarding overall perceived performance conditions indicate that a greater proportion of participants perceived their performance more positively in VE_{Free}, VE_{Hall}, and VE_{IML} compared to the baseline. For the baseline condition, the highest percentage of participants rated their overall performance as *neutral*. The XREs VE_{Free}, VE_{Hall}, and VE_{IML} achieved similarly high percentages of *good* and *excellent* ratings. Among the XREs, VE_{IML} achieved the highest percentage of ratings above *neutral*.

The final item of the questionnaire was designed to identify possible glitches or inaccuracies, particularly in the visual representation of the avatars. Despite instructions about the

assigned physical areas within their respective rooms, preliminary examination indicated that errors could still occur in the tracking system, especially during the hand clapping process. Overall, the results shown in Figure 6 indicate that there were almost no perceived errors affecting participants' performance in nearly all XREs. The baseline received the highest percentage of negative ratings below *neutral*, while VE_{Free} had the highest percentage of ratings above *neutral*.

IV. Discussion

Summarizing the objective analysis, the results indicate that the XREs of the IRENE system, with their additional system latency of 11 ms, had no negative impact on participants' musical outcome in terms of musical synchrony (*asymmetry* and *imprecision*). Although the system added latency to the participants' own signals, no impact was observed. This might have been enhanced by the use of the open headphones, which were intended to combine the participants' direct sound with the reverberation from the respective VEs. It is somewhat surprising that a tendency can be observed in the metric *asymmetry*, with the baseline having the highest mean value. This might be due to the higher delay time of the video transmission compared to the VEs. On the other hand, the mean values for the metric *imprecision* show a tendency of the lowest value for the baseline.

Significant tempo deceleration can be observed for all three XREs compared to the baseline, which showed slight tempo acceleration in the metric *mean tempo slope*. This result is unexpected, as the selected one-way delay (OWD) of 21 ms was anticipated to result in a more stable tempo performance. However, the OWD of the baseline with 10 ms appears to fall precisely within the range where either a positive tempo slope is observed or participants can maintain a stable tempo [14]. Comparing the three different XREs, no significant differences were found, but a tendency indicates that the musical XRE with the longest virtual reverberation time also shows the highest tempo deceleration.

In the case of the subjective evaluation, the XREs VE_{IML} and VE_{Hall} were rated significantly higher in terms of *Social Presence*, *Immersion*, and *Naturalness* compared to the baseline. Only the unusual representation of the free-field showed no significant difference. However, no significant effect was observed for the question of *consistency*. This might have been due to the challenging nature of comparing the *virtual performance space* to a *real performance space* without a reference. In terms of the more technical questions, a similar trend can be observed. The perceived *Responsiveness*, *Localization*, and *Separation* were rated significantly higher for VE_{IML} and VE_{Hall}, with no significant differences between them. The VE_{Free} was rated slightly lower than VE_{Hall} for *Responsiveness*. In terms of perceived *Delay* between participants' actions and their expected outcome, all three VEs of the IRENE system were rated higher than the baseline. This indicates the high influence of the high video transmission latency compared to the low latency audio transmission. However, the

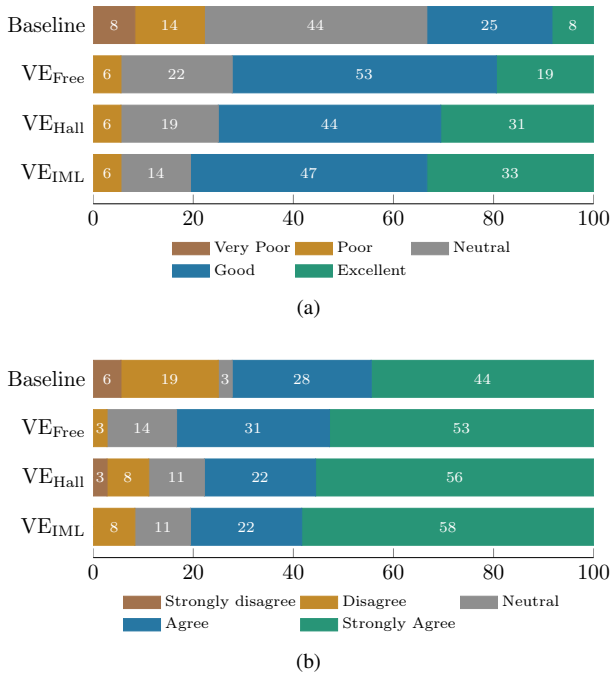


Fig. 6. Stacked bar plot showing responses to question 9 (a): *How would you rate the overall performance conditions during your performance?*, and question 10 (b): *My performance was not affected by inaccuracies or glitches in the visual display.*

overall performance condition was rated significantly higher for all three XREs than for the baseline.

Another finding is that there were no significant objective or subjective differences between two of the three XRE renderings (VE_{IML} and VE_{Hall}), indicating that the participants' responses to the rating items revealed no effect of reverberation time in this case. Only the more abstract visual representation VE_{Free} was sometimes rated significantly lower than the other two musical XREs. This could be due to the unusual visual representation of the free-field and the auditory rendering, which only reproduced the direct sound and one floor reflection. However, it is important to note that it is difficult to draw a conclusion about the influence of the individual components of the IRENE system in terms of visual representation and auditory rendering. The experimental design was not specifically aimed at examining the impact of different acoustic room renderings independent of their visual representation. Nevertheless, the findings suggest that long reverberation time could negatively impact musical performance while improving perceived QoE.

Future studies, with a clear separation of visual and auditory components, will help to explore the influence of the individual displays on the perception of musicians and draw conclusions about their benefits. Additional instrumental and musical configurations are also due to be explored in order to draw broader conclusions on the effects of XREs in relation to rhythm complexity, instrumentation, and interplay roles, since QoE variance may be expected for different performance contexts. Even if the technical effort remains high, new technological

tools, such as the acoustic or visual global metronome [33], [35] can be implemented within the INMP system, allowing flexible and free positioning within the VE, thus improving sound source separation and enabling the redesign of the visual component of such tools.

V. Conclusion

In this publication, the impact of the Immersive Room ExtensioN Environment (IRENE) was investigated to achieve an initial understanding of how the system's different extended reality environments (XRE) affect the participants' subjective experiences and musical outcomes. Additionally, the influence of various virtual environments (VE) with distinct acoustic and visual characteristics was examined. The IRENE system was modified and integrated into a framework representing a potential future NMP setup. This included three different VEs with corresponding spatial audio rendering and a traditional videoconferencing solution with monaural audio as a baseline. In an experiment involving 18 pairs of participants clapping a complementary rhythm in a representative NMP scenario using the IRENE system, the effects of several XREs were explored. Objective analysis indicated that the XREs did not significantly affect the synchrony of the musical outcome. However, tempo-related metrics showed a deceleration of tempo when using the XREs compared to the baseline. Subjective evaluations show clear trends that the XREs of the IRENE system enhance participants' perceived coherence and immersion as well as the overall perceived QoE. In particular, the observed improvements in *social presence* and *immersion* are promising and underscore the potential of XREs to enhance QoE in future immersive NMP. Future work will explore the separated visual and auditory components and their effects on objective musical outcomes and QoE.

Acknowledgments

The authors express their gratitude to all the performers who participated in the study and to the collaborators who assisted in drafting the experimental design and setup.

References

[1] S. Delle Monache, L. Comanducci, M. Buccoli, M. Zanoni, A. Sarti, E. Pietrocola, F. Berbenni, and G. Cospito, "A presence- and performance-driven framework to investigate interactive networked music learning scenarios," *Wireless Communications and Mobile Computing*, vol. 2019, pp. 1–20, 2019.

[2] R. Hupke, J. Dürre, N. Werner, and J. Peissig, "Latency and quality-of-experience analysis of a networked music performance framework for realistic interaction," in *Audio Engineering Society Convention 152*. Audio Engineering Society, 05 2022. [Online]. Available: https://aes2.org/publications/elibrary-page/?id=21659

[3] A. Carôt, C. Hoene, H. Busse, and C. Kuhr, "Results of the fast-music project – five contributions to the domain of distributed music," *IEEE Access*, vol. 8, pp. 47 925–47 951, 2020.

[4] M. Bosi, A. Servetti, C. Chafe, and C. Rottondi, "Experiencing remote classical music performance over long distance: A jacktrip concert between two continents during the pandemic," in *Journal of the Audio Engineering Society*, vol. 69, no. 12, 2021, pp. 934–945. [Online]. Available: https://www.aes.org/e-lib/browse.cfm?elib=21542

[5] J. W. Davidson, "Visual perception of performance manner in the movements of solo musicians," *Psychology of Music*, vol. 21, no. 2, pp. 103–113, 1993.

[6] B. Loveridge, "Networked music performance in virtual reality: current perspectives," *Journal of Network Music and Arts*, vol. 2, no. 1, 2020. [Online]. Available: https://commons.library.stonybrook.edu/jonma/vol2/iss1/2

[7] J.-P. Cáceres, R. Hamilton, D. Iyer, C. Chafe, and G. Wang, "To the edge with china: Explorations in network performance," in *Proceedings of the 4th International Conference on Digital Arts (ARTECH)*, 2008, pp. 61–66.

[8] M. Iorwerth and D. Knox, "Playing together, apart: Musicians' experiences of physical separation in a classical recording session," *Music Perception: An Interdisciplinary Journal*, vol. 36, no. 3, pp. 289–299, 2019.

[9] J. R. Cooperstock, "Interacting in shared reality," in *HCI International, Conference on Human-Computer Interaction*, 2005, pp. 1–7.

[10] P. Keller, "Ensemble performance: Interpersonal alignment of musical expression," in *Expressiveness in music performance: Empirical approaches across styles and cultures*. Oxford University Press, 07 2014. [Online]. Available: https://doi.org/10.1093/acprof:oso/9780199659647.003.0015

[11] F. Schroeder and P. Rebelo, "Sounding the network: the body as disturbant," *Leonardo Electronic Almanac*, vol. 16, no. 4-5, pp. 1–10, 2009.

[12] R. Hupke, S. Preihs, and J. Peissig, "Immersive room extension environment for networked music performance," in *Audio Engineering Society Convention 153*. Audio Engineering Society, 10 2022. [Online]. Available: https://aes2.org/publications/elibrary-page/?id=21909

[13] R. Hamilton, J.-P. Caceres, C. Nanou, and C. Platz, "Multi-modal musical environments for mixed-reality performance," *Journal on Multimodal User Interfaces*, vol. 4, no. 3-4, pp. 147–156, 2011.

[14] C. Rottondi, C. Chafe, C. Allocchio, and A. Sarti, "An overview on networked music performance technologies," *IEEE Access*, vol. 4, pp. 8823–8843, 2016.

[15] J.-P. Cáceres and C. Chafe, "Jacktrip: Under the hood of an engine for network audio," *Journal of New Music Research*, vol. 39, no. 3, pp. 183–187, 2010.

[16] L. Turchet, C. Fischione, G. Essl, D. Keller, and M. Barthet, "Internet of musical things: Vision and challenges," *IEEE Access*, vol. 6, pp. 61 994–62 017, 2018.

[17] L. Turchet, C. Rinaldi, C. Centofanti, L. Vignati, and C. Rottondi, "5g-enabled internet of musical things architectures for remote immersive musical practices," *IEEE Open Journal of the Communications Society*, vol. 5, pp. 4691–4709, 2024.

[18] F. Martusciello, C. Centofanti, C. Rinaldi, and A. Marotta, "Edge-enabled spatial audio service: Implementation and performance analysis on a mec 5g infrastructure," in *2023 4th International Symposium on the Internet of Sounds*, 2023, pp. 1–8.

[19] A. Carôt, C. Werner, and T. Fischinger, "Towards a comprehensive cognitive analysis of delay-influenced rhythmical interaction," in *International Computer Music Conference (ICMC)*, 8 2009. [Online]. Available: http://hdl.handle.net/2027/spo.bbp2372.2009.107

[20] S. Farner, A. Solvang, A. Sæbo, and U. P. Svensson, "Ensemble hand-clapping experiments under the influence of delay and various acoustic environments," *Journal of the Audio Engineering Society*, vol. 57, no. 12, pp. 1028–1041, 2009. [Online]. Available: https://aes2.org/publications/elibrary-page/?id=15235

[21] A. S. Bregman, *Auditory scene analysis: The perceptual organization of sound*. MIT press, 1994.

[22] B. G. Shinn-Cunningham and A. Ihlefeld, "Selective and divided attention: Extracting information from simultaneous sound sources." in *ICAD*, 2004.

[23] R. Hupke, A. Genovese, S. Sridhar, J. Peissig, and A. Roginska, "Impact of source panning on a global metronome in rhythmic networked music performance," in *2020 27th Conference of Open Innovations Association (FRUCT)*, 2020, pp. 73–83.

[24] A. F. Genovese, "Acoustics and copresence: Towards effective auditory virtual environments for distributed music performances," Ph.D, New York University, United States – New York, 2023. [Online]. Available: https://www.proquest.com/openview/fb49e48e7fdd20d010e2505cee3ae087/1?pq-origsite=gscholar&cbl=18750&diss=y

[25] P. Cairns, A. Hunt, J. Cooper, D. Johnston, B. Lee, H. Daffern, and G. Kearney, "Recording music in the metaverse: A case study of xr bbc maida vale recording studios," in *2022 AES International Conference on Audio for Virtual and Augmented*

Reality. Audio Engineering Society, 08 2022. [Online]. Available: https://aes2.org/publications/elibrary-page/?id=21842

[26] P. Cairns, A. Hunt, D. Johnston, J. Cooper, B. Lee, H. Daffern, and G. Kearney, "Evaluation of metaverse music performance with bbc maida vale recording studios," *Journal of the Audio Engineering Society*, vol. 71, no. 6, pp. 313–325, 2023. [Online]. Available: https://aes2.org/publications/elibrary-page/?id=22139

[27] A. Hunt, H. Daffern, and G. Kearney, "Avatar representation in extended reality for immersive networked music performance," in *AES 2023 International Conference on Spatial and Immersive Audio*, 08 2023. [Online]. Available: https://aes2.org/publications/elibrary-page/?id=22183

[28] P. Cairns, T. Rudzki, J. Cooper, A. Hunt, K. Steele, M. G. Acosta, A. Chadwick, H. Daffern, and G. Kearney, "Singer and audience evaluations of a networked immersive audio concert," *Journal of the Audio Engineering Society*, vol. 72, pp. 467–478, 10 2024. [Online]. Available: https://aes2.org/publications/elibrary-page/?id=22644

[29] M. Tomasetti and L. Turchet, "Playing with others using headphones: Musicians prefer binaural audio with head tracking over stereo," *IEEE Transactions on Human-Machine Systems*, vol. 53, no. 3, pp. 501–511, 2023.

[30] C. Chafe and M. Gurevich, "Network Time Delay and Ensemble Accuracy: Effects of Latency, Asymmetry," *Audio Engineering Society Convention 117*, 2004. [Online]. Available: https://aes2.org/publications/elibrary-page/?id=12865

[31] D. Schössow, "Quest - questionnaire editor system," Aug. 2023. [Online]. Available: https://doi.org/10.5281/zenodo.8233755

[32] P. Cairns, H. Daffern, and G. Kearney, "Parametric evaluation of ensemble vocal performance using an immersive network music performance audio system," *Journal of the Audio Engineering Society*, vol. 69, no. 12, pp. 924–933, 2021. [Online]. Available: https://aes2.org/publications/elibrary-page/?id=21541

[33] R. Hupke, M. Nophut, S. Preihs, and J. Peissig, "Toward professional distributed performances: Effects of a global metronome on networked musical ensemble interactions," *Journal of the Audio Engineering Society*, vol. 69, no. 10, pp. 720–736, 2021. [Online]. Available: https://aes2.org/publications/elibrary-page/?id=21467

[34] M. Müller, *Tempo and Beat Tracking*. Cham: Springer International Publishing, 2015, pp. 303–353. [Online]. Available: https://doi.org/10.1007/978-3-319-21945-5_6

[35] R. Hupke, L. Beyer, M. Nophut, S. Preihs, and J. Peissig, "Effect of a global metronome on ensemble accuracy in networked music performance," in *Audio Engineering Society Convention 147*. Audio Engineering Society, 10 2019. [Online]. Available: https://aes2.org/publications/elibrary-page/?id=20552