2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR) | 978-1-6654-5325-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISMAR55827.2022.00016

Studying the Effects of Network Latency on Audio-Visual Perception During an AR Musical Task

Torin Hopkins*
ATLAS Institute, University of Colorado Boulder
Suibi Che-Chuan Weng†
ATLAS Institute, University of Colorado Boulder
Rishi Vanukuru‡
ATLAS Institute, University of Colorado Boulder
Emma Wenzel§
‘ ATLAS Institute, University of Colorado Boulder
Amy Banic¶
Interactive Realities Lab, University of Wyoming
Mark D Gross ‖
ATLAS Institute, University of Colorado Boulder
Ellen Yi-Luen Do **
ATLAS Institute, University of Colorado Boulder

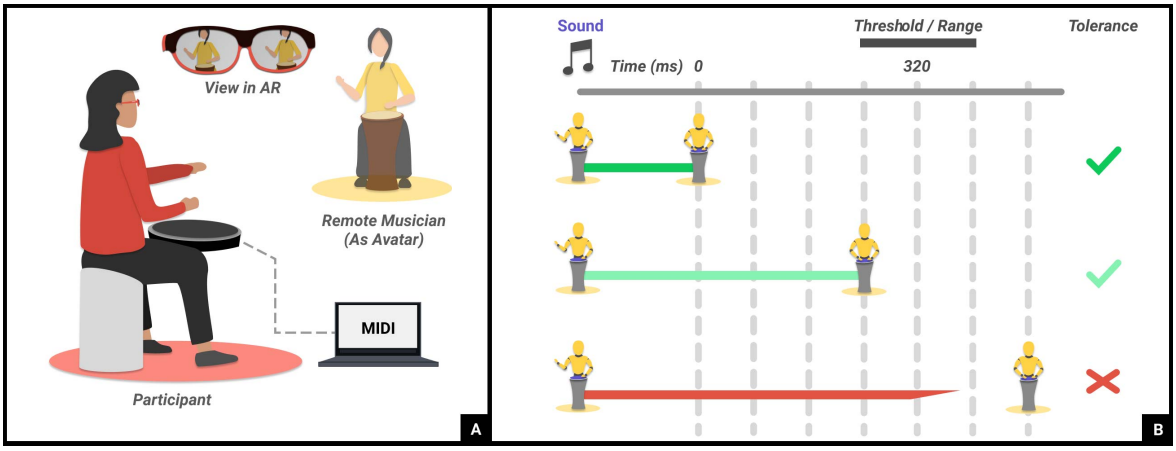

Figure 1: A) A representation of the experimental setup involving a participant collaborating with a remote musician using AR glasses and a MIDI drum, and B) An illustration of the sound-animation latency threshold being studied.

Abstract

Augmented Reality (AR), with its ability to make people feel like they are in the same space as friends from across the world, is an ideal medium for the purpose of Networked Musical Collaboration. Most conventional systems that enable networked musical collaboration minimize network latency by focusing on the transfer of auditory information at the expense of visual feedback. Studies into human perception have shown that sensory integration of audio and visual stimuli can take place even when there is a slight delay between the two signals. We studied the way changes in network latency effect participants' auditory and visual perception in latency detection, latency tolerance and attention focus; in this paper, we explore the trade-off between the presence of AR visuals and the minimization of latency. Twenty-four participants were asked to play a hand drum and collaborate with a prerecorded remote musician rendered as an avatar in AR. Multiple trials involving different levels of audio-visual latency were conducted. We then analyzed the subjective responses of the participants together with the recorded musical information from each session. Results indicate a minimum noticeable delay value–defined as the highest amount of delay that can be experienced before two stimulated senses are perceived as separate events–between 160 milliseconds (ms) and 320 ms. Players also reported that a delay between sound and an accompanying avatar animation became less tolerable at 320 ms of delay, but was never completely intolerable, even up to 1200 ms of delay. We conclude that players begin to notice delay at about 320 ms and most players can tolerate large delays between sound and animation.

Index Terms: Human-centered computing—Interaction paradigms—Mixed / augmented reality; Human-centered computing—Interaction paradigms—Collaborative Interaction; Applied computing—Sound and music computing

*e-mail: torin.hopkins@colorado.edu
†e-mail: chwe1250@colorado.edu
‡e-mail: rishi.vanukuru@colorado.edu
§email: emma.wenzel@colorado.edu
¶email: abanic@cs.uwyo.edu
‖email: mdgross@colorado.edu
**email: ellen.do@colorado.edu

1 Introduction

The past decade has seen significant progress in Augmented Reality (AR) technology. AR experiences are more spatially stable than in previous years, and the general availability of hardware that supports high-fidelity AR applications—both on mobile devices [21] and head-worn displays [17, 28]—has greatly increased. These advancements, together with constantly improving internet connectivity, have made it possible to create applications that support various forms of remote collaboration, such as online meetings

978-1-6654-5325-7/22/$31.00 ©2022 IEEE
DOI 10.1109/ISMAR55827.2022.00016

(AltspaceVR[1], Spatial[2]) and multiplayer games (RecRoom[3]). Musical collaboration is an inherently multi-modal experience. Visual feedback (such as physical gestures and facial expressions [4]) is particularly important when coordinating and conveying musical intentions, and AR environments can potentially provide such kinds of feedback in a more spatial manner. However, not many studies or commercial applications have explored the use of XR to support real-time remote musical collaboration [26].

Network Latency—defined as the time it takes for data to be transferred between its original source and its destination—is perhaps the main issue faced when attempting to communicate over the internet. Remote musical collaboration in particular requires a very low network latency for musicians to be able to coordinate effectively. While there are many hardware and software tools that support audio transmission at the required speeds [22], visual information such as the video streams of musicians are often slower to transmit. This results in a delay between audio and visual information, or unsynchronized audio-visual events, which can make musical collaboration difficult due to the cue mismatch. As a result, users often disregard video information and focus solely on audio to maintain coordination [19]. Another solution of asynchronous play using recorded video or sound clips can alleviate this challenge, but lacks the benefits of a synchronous experience where musicians can feed off each other with improvisation.

Musical collaboration in AR faces a similar challenge; transmitting information in real time over the network will likely cause a delay between spatial information, such as a musician's body movements and audio information. However, unlike video streams, AR creates a 3-dimensional environment overlaid on a user's current surroundings, thereby providing a more immersive experience. The ability to view remote musicians as fully-articulated avatars (as opposed to faces on a screen) and hear multiple musical streams coming from different locations in the space around them are further potential benefits of using AR as a medium to support remote collaboration.

Research in the field of sensory perception has shown that when it comes to perceiving temporally disparate audio and visual stimuli that are the result of the same event, the human brain might have a threshold for sensory integration [1, 7]. A recent study by Liu et al. [18] utilised the medium of Virtual Reality (VR) to study this effect further when dealing with spatial auditory stimuli and human auditory gain response functions. Leveraging this behavior in the context of network latency and musical collaboration seems to be a promising approach that has not yet been studied.

In their recent survey on music in Extended Realities (XR), Turchet et al. note that the area of remote musical collaboration in AR is relatively unexplored [26]. Further, a limited number of studies have investigated issues related to latency and the perception of multi-modal musical stimuli in XR. In this paper, we begin to address some of these gaps. We are interested in understanding how musicians respond to delayed audio (music) and visual (avatar motion) stimuli due to network latency when remote musicians are collaborating with each other using AR. In particular, we conducted an experimental study using prerecorded clips in AR to identify the minimum noticeable delay and the latency threshold at which musical collaboration with delayed AR visual and avatar animation are tolerable, as well as understand the impact of the speed or tempo of collaboration on the latency threshold. The rest of this paper is structured as follows: we first discuss related work from the fields of Networked Musical Performance, XR Collaboration, and Sensory Perception. After this, we define our research questions, and outline the design of an experimental study to help answer those questions. The results from our controlled experiment are then presented. Finally, we conclude with a discussion of how our findings relate to our initial hypotheses, and how this work can inform the design of XR musical collaboration interfaces moving forward.

[1]https://altvr.com/

[2]https://spatial.com/

[3]https://recroom.com/

2 Related Work

2.1 Networked Musical Collaboration and Performance

Networked Musical Performance can be classified into two broad categories based on the strategies used to enable remote musicians to effectively collaborate when faced with network latency [5]: instantaneous performance and user-controlled delay.

2.1.1 Instantaneous Performance

Instantaneous performance involves reducing network latency to as low a level as possible. Studies by Schuett [23] indicate that the latency threshold for "impulsive, rhythmic music" lies around 20 - 30 ms. When the delay is greater than 30 ms but less than 70 ms, collaboration is still possible, if slightly asymmetric, and variations in tempo are noticeable. These values are significantly below the the latency threshold for verbal communication. Schuett estimates this threshold to be around 200 ms. Zoom, a popular video conferencing tool, mentions a value of 150 ms [4] to be the limit for latency, beyond which communication might be hampered. Instantaneous networked musical collaboration systems are limited by the physical distance between musicians. JackTrip [8], Jamakazam [12] and Jamulus [13] achieve near-instantaneous collaboration by requiring players to use an external audio card, provide direct access to their modem, and set up an Ethernet connection to their personal device. This configuration can support remote collaboration with musicians who are located within a certain geographical distance, but cannot provide video or any additional real-time graphics. These systems are highly sensitive to configuration, server location, network speeds of individual users, and the inevitable latency affected by distance. However, it has been shown that latencies of less than 50 ms can be achieved when musicians have access to the required hardware and are within the same city or state.

2.1.2 User-controlled Delay

In this strategy, users can specify the delay with which their music is transmitted to remote partners, and also the delay in receiving streams of remote music. When musicians solely listen to streams with uncontrolled delay, the threshold for latency is similarly around 50 - 75 ms according to work by Chew et al. [6]. However, when the delay is increased to coincide with musically relevant time periods (such as beats or measures as done in the nJam project [5]), collaboration is improved due to the feeling of the delay being "in phase" with the ongoing musical performance.

In addition to latency, network jitter is another factor that influences the quality of collaboration. Jitter is the effect caused when individual data packets have differing latencies, which can occur with network fluctuations. Kleimola et al. [16] study the effect of jitter on musical collaboration in more detail. However, jitter can be reduced to an extent by buffering streams–waiting for a certain number of data packets to be received before making them available to the remote user. This results in a net increase in latency. Hence, we focus solely on the parameter of latency in this paper and ignore the problem of jitter.

2.2 Remote Collaboration in XR

Some of the earliest research projects in the field of Augmented and Virtual Reality focused on the design and study of applications that support remote collaboration [3]. Recent research has focused on use-cases such as online meetings [11], remote assistance [2], visualization tasks [15], and gaming [27]. The nature of the tasks involved

[4]https://support.zoom.us/hc/en-us/articles/202920719-Meeting-and-phone-statistics

in these applications (mostly speech and gesture-based communication) is such that it is possible to collaborate meaningfully even with the levels of latency experienced when using commercial networks [23]. As a result, many studies do not separately discuss the effect of latency on task performance, or conduct experiments in controlled environments with access to high-speed local area network communication.

However, collaborative applications that require fine-motor control or real-time physics-based operations are likely to need much lower levels of network latency. One example of this is a project by Elvezio et al. [9], where they developed a collaborative VR game to assist with remote motor rehabilitation therapy. The game requires two remote players to balance a ball on a plank while controlling the position of ropes attached to the corners of a plank. Initial user tests of the game indicate that effective collaboration requires round-trip network latency to be below 15 ms, with performance further improving at the 3-7 ms range. While 5G networks of the future might be able to achieve the required speeds to support such applications, most commercial networks cannot do so today.

2.3 Remote XR Musical Collaboration

As discussed earlier, musical collaboration is another domain where meaningful interaction is only possible when there is very low network latency. Augmented and Virtual Reality environments have already been used to support music education [24] and create new musical instruments [10]. While there have been a few projects that explore musical collaboration in XR, this area seems to be relatively unexplored. One of the earliest works in this space is the PODIUM project [14], where a Desktop VR system is used to allow remote musicians to perform together. Players view each other as stylized avatars, and basic gestures from a conductor (captured using a 6 degree-of-freedom tracker) are transmitted to all participants. The LeMo project [20] studied musical collaboration in immersive VR using a shared music sequencer. Sequencers require players to plan musical choices and input them into a static grid-like framework indicating music notes and pauses. This creates a disconnect that eliminates the real-time feeling that musical instruments provide. More recently, Tamplin et al. [25] developed a VR application for remote group singing in the context of musical therapy. Because they were using a JackTrip network that ensured fast communication between the singers, their work did not face issues with latency. Across these projects, the benefits of the XR environment (such as being able to view other participants as avatars and communicate via gestures) were considered to be significant benefits despite issues with network connectivity in some cases.

2.4 Audio-visual Perception

Neural processing is associated with inherent latency. It has been demonstrated that several senses relay information to the brain for cognitive processing at different speeds, requiring the brain to associate multi-modal stimuli to produce a unified sensory experience [29]. When perceiving the world around us, we need to integrate input from various senses and decide whether the stimuli are associated with a single event, or disparate events, such as unsynchronized audio-video events. Recent studies have investigated the optimal time difference between auditory and visual stimuli for sensory integration [1,7]. It has been shown that audio-visual integration can take place when the constituent stimuli are as far as 200 ms apart. The medium of Virtual Reality has also been used to conduct studies related to this effect. Liu et al. used VR to present auditory stimuli of different durations and spatial locations, along with visual cues, to evaluate the effect of internal auditory response reduction when faced with consecutive stimuli [18]. Results demonstrated that localization errors were positively correlated with longer-duration auditory stimuli and shed light on how sound affects the ability to locate a sound in a virtual environment.

3 Research Questions

The key insights from prior studies are summarized below:

- Instantaneous Networked Musical Performance requires network latencies to be minimized to the extent possible, ideally below 30ms [23]. This requires either access to high-speed internet connections, or specialised hardware and networking protocols [22], both of which may not be available to all.
- Augmented Reality can potentially provide significant benefits to remote musical collaboration, particularly by enabling musicians to see remote participants as avatars in a shared space [26]. There are delays associated with modern XR applications which are still above the threshold for optimal musical collaboration (200 ms and above).
- The human mind is capable of integrating audio-visual input even when the two constituent stimuli are presented at different times within a threshold [1, 7].

Building upon these findings, our research questions are:

1. When collaborating with a remote partner via AR, what is the minimum noticeable delay and threshold of network latency (in the transmission of visuals relative to audio) until which musical interaction is possible and tolerable?
2. How does a musician's focus shift between auditory and visual information as latency increases?

4 Method

To further investigate our research questions, we conducted a 2x8 within-subjects experiment to simulate remote musical collaboration in AR. We experimentally controlled for two types of remote drumming collaboration (Section 4.1): Mimic and Improvise. We experimentally controlled eight levels of network latency: 0ms, 20ms, 40ms, 80ms, 160ms, 320ms, 640ms, and 1200ms (Section 4.4). We used Balanced Latin Squares across all conditions to reduce potential ordering effects. We first discuss the tasks and apparatus, followed by details about the participants involved and the experimental procedure.

4.1 Musical Collaboration Tasks

In order to focus on the effects of latency on musical rhythm (or the idea of being 'in time'), we chose to base our tasks around simple rhythmic patterns played on a hand drum. The hand drum is an instrument that does not require much prior knowledge to begin playing, hence it would be accessible to the widest range of participants.

Across the eight latency conditions, there were two task types:

1. **Mimic**: Here, participants were asked to mimic the remote partner and try to play the same rhythm on their drum.
2. **Improvise**: In this task, participants were given the freedom to play whatever they felt best complemented the rhythm being played by their remote partner.

In both cases, the speed of the rhythm played by the remote partner was 90 beats per minute. During the "mimicry task", the remote partner would play a strict 4/4 (4 beats per bar, 4 bars per measure) rhythm, which translates to a gap of 667 ms between subsequent hits of the drum. In the "improvise" task, the remote player would continue playing at 90 beats per minute, but using a more natural, free-form rhythm.

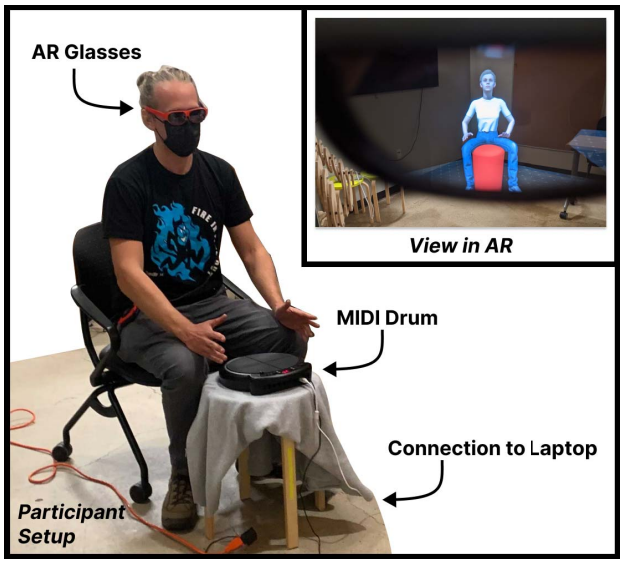

Figure 2: Experimental Setup: A Participant wearing an AR Headset and sitting in front of a MIDI drum. The view of the remote musician's avatar is included in the top-right corner.

4.2 Apparatus

The Nreal Light headset was used to enable participants to view their remote partner in AR. Participants were given an electronic drum pad to play along with the avatar. The output from the drum pad (in the form of MIDI notes) was recorded on a laptop computer.

4.3 Simulated Musical Partner

We simulated a remote musical partner by recording the motion and audio of a musician playing the above mentioned rhythm patterns on a hand drum. This was done for two reasons: (1) To ensure that all participants experienced the same behavior from their remote partners, and (2) to control the net latency between audio and visual more precisely.

An 18-camera Qualysis motion capture system[5] was used to record player movement as animation clips. The drum audio was recorded into the Ableton digital audio workstation software via an external sound card. Once recorded, these motion and audio files were applied on a Humanoid avatar (sourced from the Mixamo character and animation platform[6]) and rendered via an application for the Nreal Light AR headset [7] using the Unity live development engine (version 2020.2.2f1). This application was presented to participants during the study, with differing offsets between the music and resulting animation (Figure 3), discussed in Section 4.4.

4.4 Latency Conditions

Eight different values of delay between the remote partner's music stream and the animation of the avatar were considered. Starting with the baseline condition of no delay (0 ms gap), we consider constantly doubling latency values of 20 ms, 40 ms, and so on until 1200 ms. This was done to collect more points of data around the range of tolerable delay (50 - 70 ms) as reported by prior research [16, 23], while also extending the monitoring range until the point where speech communication suffers (150 - 300 ms), and beyond, as prior work has not explored the idea of "sound-animation delay" in the context of AR. Because conditions were created manually using Unity's Timeline tool, controlled latency near 0ms delay between the visual striking of the drum and the sound of the drum strike were made possible. At each latency condition, both musical tasks—mimicry and improvise—would be observed, resulting in a total of 16 tasks per participant.

[5]https://www.qualisys.com/
[6]https://www.mixamo.com/
[7]https://www.nreal.ai/light/

4.5 Procedure

24 participants were recruited via university mailing lists and word of mouth. All participants were university students, and the experiment was conducted inside a research laboratory on the university campus. Participants were briefed about the nature and number of tasks involved.

The AR glasses were adjusted to the participants' forehead size to ensure a comfortable fit. Each participant sat on a chair behind an electronic drum wearing the AR glasses. The AR avatar of the remote musician was positioned facing the participant at a distance of 3 meters in front of them. The environment was that of a plain room with no visual distractions.

Participants were then asked to perform the "mimic" and "improvise" task for each of the 8 latency conditions (16 tasks in total). Both tasks were 12 seconds in length (4 bars of 4 beats per bar at 90 beats per minute) for all 8 latency conditions.

Delay Amount

Task Delay (ms)	N	Mean	Std. Deviation
Mimic 0	24	2.0000	1.38313
Improvise 0	24	2.2083	1.25036
Mimic 20	24	2.3125	1.48772
Improvise 20	24	2.2500	1.35935
Mimic 40	24	2.3750	1.40844
Improvise 40	24	2.2708	1.51068
Mimic 80	24	2.2083	1.21509
Improvise 80	24	2.3125	1.47304
Mimic 160	24	1.8750	1.07592
Improvise 160	24	2.7083	1.78104
Mimic 320	24	2.6042	1.84732
Improvise 320	24	3.1458	1.80265
Mimic 640	24	3.7917	1.55980
Improvise 640	24	2.9583	1.70623
Mimic 1200	24	4.0000	2.35907
Improvise 1200	24	3.3958	1.89381

Table 1: Delay Scores Results: The means and standard deviations for the jam and follow sections for perception of delays 0ms - 1200ms.

We recorded the participant's electronic drum beat as MIDI (musical instrument digital interface) information in Ableton Live 11 for further analysis. After participants finished the two tasks in each condition, we verbally asked and recorded responses to three questions:

1. How much delay did you experience between animation and sound on a scale from 1 to 7? (No delay = 1, Max delay= 7)
2. How would you rate the tolerability of the delay experienced on a scale from 1 to 7? (Not at all tolerable=1, Not noticeable / very tolerable = 7)
3. Did you focus more on animation or sound on a scale from 1 to 7? (Sound = 1, Neutral= 4, and Animation = 7)

Across all participants, the presentation of the latency conditions was counterbalanced using a Latin Square design. The task order was reversed for half of the participants.

After participants finished all the tasks, they were instructed to answer a questionnaire related to the system as a whole. Overall, each session was approximately 45 minutes.

Figure 3: Avatar Representation of the Virtual Remote Musician and the recorded animation where the hands alternate each hit on the drum corresponding to the sound of a drum hit.

5 Results

Our results were derived in three ways: 1) a quantitative score between 1 and 7 for three questions asked after each drumming task described in the procedure section, 2) quantitative results derived from recorded musical data in the form of MIDI, and 3) qualitative data gathered from the questionnaire pertaining to the experience of delay between the presented sound and visual information. To analyze the data, a repeated measures Analysis of Variance (RM-ANOVA) was conducted on the data, and a Mauchly's test did not reveal a violation of sphericity for any of the RM-ANOVA tests. Post-hoc Least Significant Difference (LSD) tests were conducted on the data to explain any significance among the groups.

5.1 Delay Perception and Tolerance

After each task, participants were asked to rate their perceived sense of delay, how tolerable the delay was to the experience of playing

Figure 4: The Motion + Audio Capture setup. A Qualisys motion capture system was used to record player motion, and the drum audio was recorded into Ableton Live.

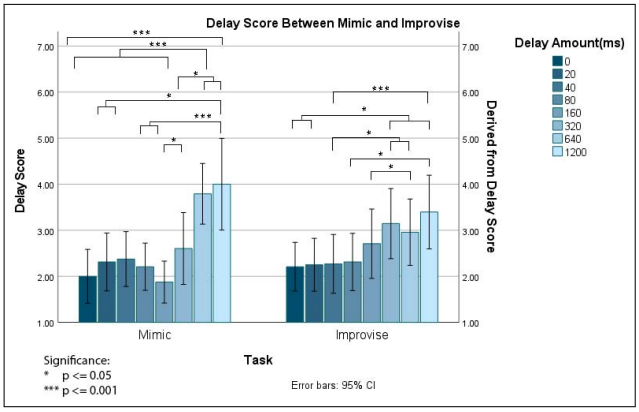

Figure 5: Delay Scores Comparison: The delay scores vs. the delay amount graphed for both the mimic and improvise sections.

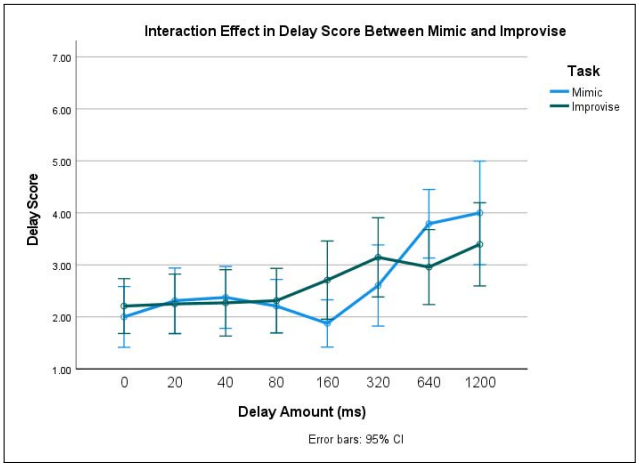

Figure 6: Interaction Effect in Delay Scores Comparison: The interaction effect in delay scores vs. the delay amount graphed for both the mimic and improvise sections.

Delay Tolerance

Task Delay (ms)	N	Mean	Std. Deviation
Mimic 0	24	5.73	1.713
Improvise 0	24	5.81	1.325
Mimic 20	24	6.02	1.355
Improvise 20	24	5.48	1.850
Mimic 40	24	5.52	1.598
Improvise 40	24	5.81	1.451
Mimic 80	24	5.58	1.316
Improvise 80	24	5.75	1.482
Mimic 160	24	6.00	1.285
Improvise 160	24	5.33	1.685
Mimic 320	24	5.06	2.023
Improvise 320	24	4.56	1.952
Mimic 640	24	4.00	1.769
Improvise 640	24	5.33	1.949
Mimic 1200	24	4.17	2.353
Improvise 1200	24	4.79	1.911

Table 2: Delay Tolerance Results: The means and standard deviations for the jam and follow sections for tolerance of delays 0ms - 1200ms.

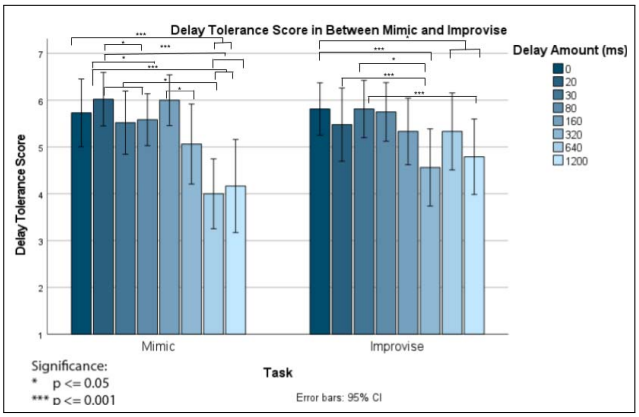

Figure 7: Delay Tolerance Comparison: The tolerance scores vs. the delay amount for both the mimic and improvise latency conditions.

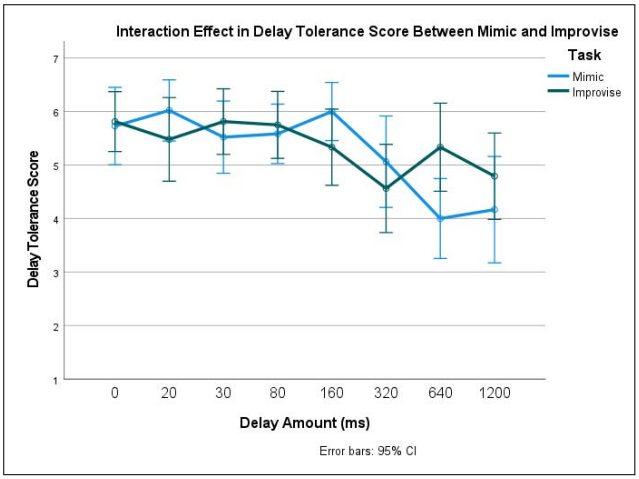

Figure 8: Interaction Effect in Delay Tolerance Scores Comparison: The interaction effect in delay tolerance scores vs. the delay amount graphed for both the mimic and improvise sections.

the hand drum, and whether they were focusing on the sound or animation.

5.1.1 Minimum Noticeable Delay

Participants were asked to rate their perceived delay on a scale from 1 to 7 for minimum noticeable delay. A RM-ANOVA revealed a significant difference among minimum noticeable delay means, $F(7, 161)= 9.74$, $p<0.001$, $\eta_p^2=0.30$, where the delay amount conditions of 0ms, 20ms, 40ms, 80ms, 160ms, were significantly smaller than delay amounts of 320ms, 640ms, 1200ms (see Table 1). Delay scores began increasing significantly at 320ms and continued to increase as the delay between audio and visual information increased (see Table 1 for reported means and standard deviations).

There was an observable interaction effect in the reported delay score b theetween mimic task and the improvisation task $F(7, 161)= 2.61$, $p=.014$, $\eta_p^2=.10$, whereby the scores for delay between the two tasks begin to deviate from each other at 160ms (see Figure 6). At 160ms delay, the reported delay in the improvisation task became more noticeable than it was in the mimic task. This trend completely reverses at 640ms of delay and continues through 1200ms of delay, where delay is reported to be noticeable in both conditions. However, participants reported less noticeable delay when improvising (Figure 5).

5.1.2 Delay Tolerance

To calculate delay tolerance we asked participants to rate how tolerable the delay between audio and visual feedback was with regard to playing drums on a scale from 1 to 7. A score of 1 indicated that the participant felt the experience was completely intolerable and would not use an application that exhibited this type of delay. A score of 7 indicated that the participant felt the delay (whether or not it was perceived) caused no issue with their playing experience. Delay tolerance was calculated using a RM-ANOVA and a LSD for post-hoc analysis.

We observed that as the amount of visual delay increased participants began to score the tolerance lower at 320ms of delay $F(7, 161)= 9.94$, $p<.001$, $\eta_p^2=.30$, where the delay amount conditions of 0ms, 20ms, 40ms, 80ms, 160ms, were significantly smaller than delay amounts of 320ms, 640ms, 1200ms (see Table 2 for reported means and standard deviations)

5.1.3 Visual vs. Audio Attention Focus

We assessed whether participants were focusing on sound or visual animation by having them choose a number between 1 and 7 on a spectrum (1=sound, 4=neutral, and 7=animation) after completing the musical task. A score of 1 indicated that the participant was focusing solely on the sound rather than the animated avatar, while a score of 7 indicated that they were focused solely on the animation as a means of assessing tempo. Results were assessed using RM-ANOVA and a LSD for post-hoc analysis. The test revealed that there was no significant difference among any of the delay conditions with respect to sound or visual animation focus, $F(7, 161)= 9.94$, $p>.001$. Results also indicated that participants frequently switched focus between trials without any clear preference for either sound or animation.

5.2 Rhythm Variability with Increasing Delay

To measure the level of disruption the participants experienced when drumming with different amounts of visual-auditory delay, we recorded the the drumming output from the instrument during the mimic task.

The participants' MIDI information was recorded using Ableton Live as they drummed on the MIDI instrument. The MIDI information was then exported from Ableton at a resolution of 96 pulses per quarter note (ppq). This is calculated by dividing the inter-note

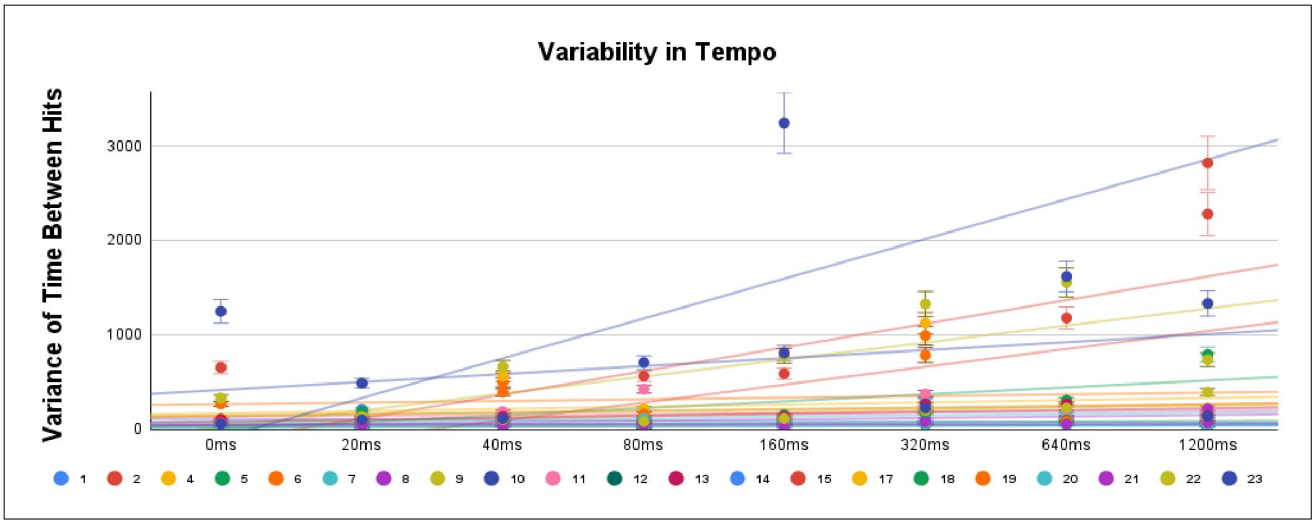

Figure 9: Variability in Tempo: The variance of the participants drumming tempo by the eight delay amount (ms) conditions. Individual participant numbers are color-coded along the bottom of the graph.

interval (667ms) by sixteen and then again by twenty-four steps. This number was then converted to milliseconds for analysis.

We calculated the sample variance associated with the drumming data of each mimicry task and delay condition (Figure 9). Variance in drumming frequency shows how much the participants altered their tempo while playing. We experienced mechanical errors from the MIDI drum during this portion of the experiment and variability was much higher than expected for some data points. For this reason we chose to eliminate the MIDI data of three participants (P3, P16, and P24) that exhibited this type of mechanical issue.

The variance data from all participants associated with each delay condition were positioned in a scatter plot graph. We conducted a linear regression analysis that resulted in a positive trend line for almost all participants, however, there were no significant line fit match due to the high variability in the data (see Figure 9).

5.3 Qualitative Feedback

The questionnaire asked participants three basic sets of questions: 1) about the motion of the avatar, 2) more detail as to what they focused on during the experiment (sound or visual content), and 3) any general feedback they had about their experience during the experiment.

Participants had varying opinions on the motion of the remote musician's avatar. Most agreed that the arm movements seemed particularly realistic ("The motion of her arms had variations in the elbow which felt realistic and I liked that."), but some participants expected a greater degree of animation in the rest of the body as well ("The arms and hands looked somehow accurate to real life. The rest of the body a bit static."). This was a limitation of the motion capture process, where we intentionally focused on capturing accurate drum-strike movements.

Across participants, we noticed an interesting trend relating to their focus depending on the task type (mimic or improvise) and level of latency (with or without perceivable delay). In both tasks, when there was no perceivable delay, participants tended to give equal attention to both the audio and animation ("I feel like I was focusing on both in equal proportion. When there is no delay, brain just automatically focuses and synchronises both the sound and the animation."). In the "mimic" task, some participants attempted to time their strikes to the visual feedback from the remote musician's avatar. This visual information was less important in the "improvise" task, as participants were more concerned about the resulting music created by the superposition of their playing and the remote musician's audio stream. However, once the latency started to increase beyond a certain limit, their focus mostly shifted towards the audio stream, and the visuals were seen as a distraction ("When improvising, I'm listening first and looking second. As long as I can hear the other players without any delay, I can improvise without any need for visual cues"). While this result was expected supported by work comparing audio-video streams [19], it is interesting that the quantitative results reveal a higher tolerance threshold for the visual information from the avatar being delayed, when compared to the musical latency threshold.

When asked for general feedback about the experiment, participant responses were primarily related to the behaviour of the remote musician's avatar and their experience with the musical tasks. In particular, participants indicated a need for more upper-body gestures from the avatar for effective communication. They also asked for variations in the musical rhythm and context to account for more situations and make the experiment less monotonous.

6 Discussion

In this study we aim to define a threshold for both perceiving a minimum noticeable delay and when players can no longer tolerate the amount of delay between a perceived sound and visual animation. In addition, we are also interested in investigating whether players focus more on sound or animation while delay increases.

We addressed the first research question of defining the minimum noticeable delay by assessing the delay amount scores indicated by the participants after each task. Results indicated that players began significantly noticing delay between 160ms and 320ms. However, there was no significant difference between the 160ms condition and the conditions with less delay than 160ms. There was a significant increase at 320ms of visual delay. Therefore, we hypothesize that the minimum noticeable delay value is higher than 160ms and possibly less than 320ms. In future studies we will investigate perceived delay between 160ms and 320ms to test for a more precise value of minimum noticeable delay.

Additionally, we address the maximum amount of delay a participant can tolerate while playing drums. The upper threshold for tolerance of visual delay was surprisingly undefinable. We noticed that the tolerance began significantly declining at the same delay

amount of 320ms as the reported delay score. This suggests that as soon as players begin noticing delay, the experience begins to degrade. Interestingly, the reported delay tolerance levels are not observed below a neutral score of 4. We interpret this to mean that a maximum threshold for delay was not found during our test sequence. Because players were mimicking a virtual drummer playing at a regular tempo of 90bpm (667ms between beats), 1200ms of delay causes animations to fall out of sync by almost 2 full cycles of sequential beats. This poses a limit for understanding an upper threshold for delay tolerance with regularly spaced drum beats because a player will not be able to distinguish the discrepancy in delay between cycles. It is possible that irregular drum beats may cause this upper tolerance threshold to become more apparent.

During the improvisation task players did encounter irregular drum beats, but did not experience as much delay overall as they did in the mimic section. We think that this is due to an increase in cognitive load during the improvisation task. As the delay becomes more and more noticeable, the delay scores associated with the improvisation task begin to become greater than those associated with the mimicry task. Cognitive loading occurs when a more mentally and/or physically demanding task is presented and distracts the player from being able to process additional information. In this case we think the perception of delay becomes less apparent to the player when improvising with high amounts of visual delay.

The second research question addresses whether the player is focusing on sound or visual information as the amount of delay increases between the two sources of information. We did not find any significant effect of changing focus throughout the test, nor did focus change reliably to one stimulus or the other. We noted that participants chose sound or visualization, sometimes went back and forth between information sources, or were otherwise confused about what specifically to focus on. This produced a result that averaged to a neutral score of 4 and therefore did not support the hypothesis that players will focus more on sound as delay increases.

The findings of this study are particularly relevant for designers and researchers working on developing more full-fledged XR musical collaboration systems. Knowing that there is a higher tolerance for delayed animations opens up the possibility of developing a multi-network system that takes advantage of existing networked musical performance tools such as JackTrip [8], while also communicating information about player movement and gestures via more general-purpose networks. Crucially, this means that progress in this area need not merely be tied to the improvement of networking infrastructure. Even with the threshold being what it is, we anticipate that there is considerable scope for the design of unique features that make the collaborative experience more enjoyable, drawing from related work on player-controlled delay [5, 6]. We plan on exploring these directions as part of future research into the field of XR musical collaboration.

7 Limitations

Given that this is an initial, controlled experimental study into human perception, our choice of tasks and overall design of the study inevitably prioritized internal validity over external applicability. While we are talking about a collaborative musical context, the remote musician presented to the participants was simulated. Future studies involving fully-networked collaborative prototypes with additional communication channels (speech, gestural), as well as real participants on both sides of the network, would allow us to better understand the effect of audio-visual delay in a real-life context. Such studies could also have longer and more natural tasks, involving real-time collaboration on music that is more familiar to participants.

Regarding the choice of participants, experience with musical instruments was not a criteria for inclusion or exclusion. Novice musicians are likely to have different expectations from such a system as compared to more experienced players, and this is bound to affect their perception tolerance thresholds as well. Conducting a similar study with stratified groups of participants would allow us to quantify these effects further.

We did not collect or record information from the AR headset being worn by the participants. Future studies could potentially use participant gaze information to also get a sense of whether they were focusing on the visuals (e.g. looking at the remote musician's avatar) or audio (e.g. looking at their own drum), and use this information to corroborate the responses by the participants after the sessions.

8 Conclusion

Most systems that enable instantaneous networked musical collaboration have focused on low-latency transmission of auditory information between remote musicians, as slower video streams can often be more confusing than helpful. Augmented Reality can be a promising medium to support remote musical collaboration due to its ability to render remote musicians as avatars in the same space as a local player, thereby resulting in a more immersive experience. Through this study, we demonstrate that musicians are able to collaborate in AR even at latencies higher than those required for purely musical or speech-based communication. While their focus might shift between the auditory and visual streams of information depending on the degree of latency, we found that delays in AR visuals seemed to be more tolerable than those for video streams as reported by prior work. These findings are encouraging as they indicate that network latency is not the main roadblock to meaningful collaboration, and that immersive visual feedback can be useful to musical interaction even if it is slightly delayed from audio. Moving forward, we hope to use these results in the design of immersive environments that better support real-time remote musical collaboration.

Acknowledgments

The authors wish to thank Ericsson Research who have been instrumental in development of this research. Specifically, Alvin Jude Hari Haran, Gunilla Berndtsson, Ali El Essaili, Hector Caltenco, Meral Shirazipour, Per-Erik Brodin, Saeed Bastani, Greg Phillips, Per P Karlsson, and Amir Gomroki. Other notable contributions from Hooman Hedayati, Colin Soguero, Peter Gyory, Darren Sholes, Dan Szafir, and Alex Eldridge throughout this project have been very much appreciated. The first author would also like to thank the pilot testers, players who graciously gave their time to help provide quality data, and those who helped with editing of the figures and paper text. Thank you so much.

References

[1] W. J. Adams. The development of audio-visual integration for temporal judgements. *PLOS Computational Biology*, 12(4):e1004865, 2016.

[2] H. Bai, P. Sasikumar, J. Yang, and M. Billinghurst. A user study on mixed reality remote collaboration with eye gaze and hand gesture sharing. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pp. 1–13, 2020.

[3] M. Billinghurst and H. Kato. Collaborative augmented reality. *Communications of the ACM*, 45(7):64–70, 2002.

[4] L. Bishop and W. Goebl. When they listen and when they watch: Pianists' use of nonverbal audio and visual cues during duet performance. *Musicae Scientiae*, 19(1):84–110, 2015.

[5] N. Bouillot. Njam user experiments: Enabling remote musical interaction from milliseconds to seconds. In *Proceedings of the 7th International Conference on New Interfaces for Musical Expression*, NIME '07, p. 142–147. Association for Computing Machinery, New York, NY, USA, 2007. doi: 10.1145/1279740.1279766

[6] E. Chew, A. Sawchuk, C. Tanoue, and R. Zimmermann. Segmental tempo analysis of performances in user-centered experiments in the distributed immersive performance project. In *Proceedings of the Sound and Music Computing Conference, Salerno, Italy*, 2005.

[7] M. J. Crosse, G. M. Di Liberto, and E. C. Lalor. Eye can hear clearly now: inverse effectiveness in natural audiovisual speech processing relies on long-term crossmodal temporal integration. *Journal of Neuroscience*, 36(38):9888–9895, 2016.

[8] J.-P. Cáceres and C. Chafe. Jacktrip: Under the hood of an engine for network audio. *Journal of New Music Research*, 39:1–4, 09 2010. doi: 10.1080/09298215.2010.481361

[9] C. Elvezio, F. Ling, J.-S. Liu, and S. Feiner. Collaborative virtual reality for low-latency interaction. In *The 31st Annual ACM Symposium on User Interface Software and Technology Adjunct Proceedings*, pp. 179–181, 2018.

[10] R. Hamilton. Coretet: A dynamic virtual musical instrument for the twenty-first century. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 1395–1395. IEEE, 2019.

[11] Z. He, R. Du, and K. Perlin. Collabovr: A reconfigurable framework for creative collaboration in virtual reality. In *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 542–54.

[12] JamKazam. JamKazam | Live, In-Sync Music Jamming Over the Internet.

[13] Jamulus. Jamulus – Play music online. With friends. For free.

[14] B. Jung, J. Hwang, S. Lee, G. J. Kim, and H. Kim. Incorporating co-presence in distributed virtual music environment. In *Proceedings of the ACM symposium on Virtual reality software and technology*, pp. 206–211, 2000.

[15] R. Khadka, J. H. Money, and A. Banic. Evaluation of scientific workflow effectiveness for a distributed multi-user multi-platform support system for collaborative visualization. In *Proceedings of the Practice and Experience on Advanced Research Computing*, PEARC '18. Association for Computing Machinery, New York, NY, USA, 2018. doi: 10.1145/3219104.3229283

[16] J. Kleimola. Latency issues in distributed musical performance. *Telecommunications Software and Multimedia Laboratory*, 2006.

[17] G. Lin, T. Panigrahi, J. Womack, D. J. Ponda, P. Kotipalli, and T. Starner. Comparing order picking guidance with microsoft hololens, magic leap, google glass xe and paper. In *Proceedings of the 22nd International Workshop on Mobile Computing Systems and Applications*, HotMobile '21, p. 133–139. Association for Computing Machinery, New York, NY, USA, 2021. doi: 10.1145/3446382.3448729

[18] J. Liu, V. Drga, and I. Yasin. Optimal time window for the integration of spatial audio-visual information in virtual environments. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*, pp. 723–728. IEEE, 2021.

[19] B. Loveridge. Networked music performance in virtual reality: current perspectives. 2020.

[20] L. Men and N. Bryan-Kinns. Lemo: supporting collaborative music making in virtual reality. In *2018 IEEE 4th VR workshop on sonic interactions for virtual environments (SIVE)*, pp. 1–6. IEEE, 2018.

[21] P. Nowacki and M. Woda. Capabilities of arcore and arkit platforms for ar/vr applications. In W. Zamojski, J. Mazurkiewicz, J. Sugier, T. Walkowiak, and J. Kacprzyk, eds., *Engineering in Dependability of Computer Systems and Networks*, pp. 358–370. Springer International Publishing, Cham, 2020.

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

[23] N. Schuett. The effects of latency on ensemble performance. *Bachelor Thesis, CCRMA Department of Music, Stanford University*, 2002.

[24] S. Serafin, A. Adjorlu, N. Nilsson, L. Thomsen, and R. Nordahl. Considerations on the use of virtual and augmented reality technologies in music education. In *2017 IEEE Virtual Reality Workshop on K-12 Embodied Learning through Virtual & Augmented Reality (KELVAR)*, pp. 1–4. IEEE, 2017.

[25] J. Tamplin, B. Loveridge, K. Clarke, Y. Li, and D. J Berlowitz. Development and feasibility testing of an online virtual reality platform for delivering therapeutic group singing interventions for people living with spinal cord injury. *Journal of telemedicine and telecare*, 26(6):365–375, 2020.

[26] L. Turchet, R. Hamilton, and A. Camci. Music in extended realities. *IEEE Access*, 9:15810–15832, 2021.

[27] M. Viitanen, J. Vanne, T. D. Hämäläinen, and A. Kulmala. Low latency edge rendering scheme for interactive 360 degree virtual reality gaming. In *2018 IEEE 38th International Conference on Distributed Computing Systems (ICDCS)*, pp. 1557–1560. IEEE, 2018.

[28] C. Xu. Nreal: Ready-to-wear mixed reality glasses. In *SPIE AVR21 Industry Talks II*, vol. 11764, p. 1176409. International Society for Optics and Photonics, 2021.

[29] B. Yin, D. B. Terhune, J. Smythies, and W. H. Meck. Claustrum, consciousness, and time perception. *Current Opinion in Behavioral Sciences*, 8:258–267, 2016. Time in perception and action. doi: 10.1016/j.cobeha.2016.02.032