Int. J. Human–Computer Studies 185 (2024) 103231

Contents lists available at ScienceDirect

International Journal of Human - Computer Studies

journal homepage: www.elsevier.com/locate/ijhcs

Selection as Tapping: An evaluation of 3D input techniques for timing tasks in musical Virtual Reality

Alberto Boem *, Luca Turchet

Department of Information Engineering and Computer Science, University of Trento, Italy

ARTICLE INFO

Keywords:
Virtual reality
Music technology
3D user interfaces
Sensorimotor synchronization
Tapping
Time perception
Selection techniques

ABSTRACT

While numerous studies have examined 3D interaction techniques for Virtual Reality (VR) across various tasks and scenarios, limited research has focused on music-related applications. However, the most common input techniques in consumer VR systems have been developed outside of the musical domain. Therefore they have not been tested in tasks where synchronization with auditory stimuli and timing plays a crucial role. There is a lack of empirical knowledge about performance and user experience. This paper presents a comparison of five selection input techniques for VR employing the tapping paradigm commonly utilized in the research on sensorimotor synchronization. We assess asynchrony and timing variance as well as user experience, encompassing factors such as ease of use, workload, and cybersickness of such techniques. The study involved 30 participants, both with and without musical expertise, and encompassed the examination of all techniques using one and two hands. Our analysis yielded several key findings: (1) different input techniques yielded distinct outcomes regarding timing asynchrony and variance; (2) the choice of interaction metaphor significantly influenced the user experience; (3) tracking stability emerged as a critical factor. Building upon these insights, we identified essential considerations for selecting the most suitable technique for music creation in VR and proposed design guidelines and future research directions in this domain.

1. Introduction

The rapid development and availability of low-cost technologies have created a resurgent interest in Virtual Reality (VR) applications for musical purposes, attracting attention from composers, performers, and developers. Musical VR comes in a variety of forms: virtual musical instruments, generative audio-visual systems, gamified musical environments, and multi-user and shared virtual concerts (Turchet et al., 2021). Over the years, musical VR has been explored in a wide variety of immersive technologies such as stereoscopic projections, Head-Mounted Displays (HMDs), as well as custom input devices, and spatial audio systems (Serafin et al., 2016). Nevertheless, as noted by Steed et al. (2021), nowadays the majority of VR systems are HMD-centric (i.e., Meta Quest, Pico, HTC Vive). These systems are composed of standalone headsets with embedded batteries and equipped with wireless internet connections. They also employ quasi-standardized 6-DoF handheld tracked devices equipped with buttons, triggers, and joysticks. Most of these HMDs incorporate "inside-out tracking" systems based on video sensors that enable convenient tracking without the use of external hardware (Gourlay and Held, 2017). Moreover, this tracking methodology integrated with gesture and pose recognition algorithms allows free-hand interactions, promoting a more natural and accessible user experience. Furthermore, SDKs and software frameworks and development tools such as the SteamVR or Microsoft Mixed-Reality Toolkit (MRTK) as well as emerging standards like WebXR (WebXR) and OpenXR (OpenXR) indicate a growing tendency towards standardized interaction techniques based on shared metaphors, especially for selection, manipulation, and navigation in virtual environments.

These developments have prompted an appropriation of such interaction techniques by developers and musicians, which led to the development of a growing number of commercial applications oriented towards music performance and production, such as Patch XR, Lyra VR, Virtuoso VR and Electronauts VR Music. However, since these developments happened outside the musical domain, there is still little knowledge on how these techniques can be effectively applied to music making (Berthaut, 2020). Are these quasi-standardized interaction techniques capable of supporting the needs for music-making in VR? This question arises since music requires domain-specific tools. Therefore, it is important to understand the limits and possibilities of these interaction techniques that emerged in the non-musical part of VR. Hamilton and Camci described musical VR systems as "*Audio-first*" systems (Çamcı and Hamilton, 2020). With this expression, they

* Corresponding author.
E-mail addresses: alberto.boem@unitn.it (A. Boem), luca.turchet@unitn.it (L. Turchet).

https://doi.org/10.1016/j.ijhcs.2024.103231
Received 31 May 2023; Received in revised form 13 November 2023; Accepted 23 January 2024
Available online 3 February 2024

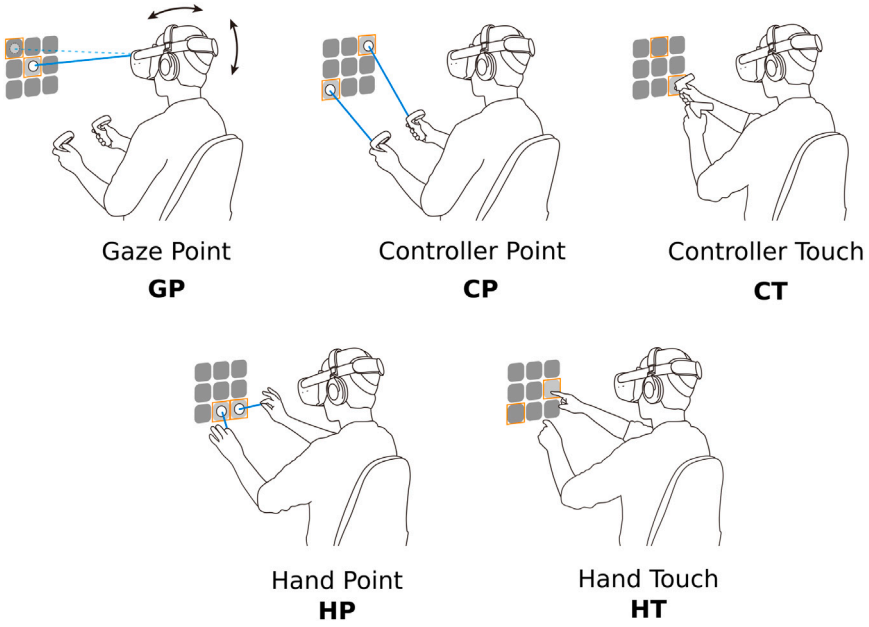

Fig. 1. A graphical representation of the five VR selection techniques used in the study. From left to right: Gaze Point (GP), Controller Point (CP), Controller Touch (CT), Hand Point (HP), and Hand Touch (HT).

highlight a distinction between systems where music and sound play a simple or auxiliary role and systems where music and sound represent the fundamental elements defining the user experience. In "*Audio-first*" VR systems, time is one of the fundamental elements and constraints against which the experience unfolds. This makes musical VR systems different from non-musical VR experiences.

Interaction techniques in the field of 3D User Interfaces are often evaluated by comparing how fast a user responds to a stimulus or completes a task in the shortest time possible. Time is treated as an aspect that must be minimized. However, reacting to a stimulus (being it visual or auditory) is different from moving in time with a paced sequence of signals. This is a typical experience in music from conducting an orchestra, to following a click-track while recording or playing together with other musicians. Keeping a tempo and producing timed actions to a series of audio pulses is a form of sensorimotor synchronization (Repp, 2006a). Previous research showed that moving along a beat is an activity that relies on error-correction mechanisms (Repp, 2005; Repp and Su, 2005). Synchronizing movements to a sequence of paced tones means executing timed responses that are delayed in order to approximately coincide with the next tone. Therefore, such form of synchronization involves a form of prediction of events, which is absent in spontaneous reaction (Repp and Keller, 2004; Pecenka et al., 2013). In addition, keeping a tempo is a continuous and prolonged activity that can be sustained for a long period of time, from the canonical three minutes of a "pop" song to the indefinite duration of a free-form session of improvisation. Therefore, methods and metrics used in canonical and non-musical 3D UI research to evaluate interaction techniques appear to have limited use in this context.

Wanderley and Orio (2002) suggested that evaluations of gestural interactions for music should focus on timing tasks (using different tempi), investigating the effectiveness of synchronizing gestures with auditory stimuli, and measurements of temporal precision. Previous works in the context of musical VR have seldom addressed such aspects since they have considered only few interaction techniques and tempi. They also used basic timing tasks, experienced not only through HMDs, but also with projected VR (Berthaut et al., 2011). To our best knowledge, no relevant work has studied and compared the interaction techniques found in contemporary VR systems. Therefore, it is important to understand how a certain interaction technique affects the user experience and supports a musical task (Holland et al., 2019). Moreover, there is a lack of clear guidelines and agreed-upon methodologies for evaluating and analyzing such techniques.

This paper presents the evaluation of five input selection techniques for musical VR (see Fig. 1). Selection is not only the most common and straightforward interaction in VR, but it also possesses a musical quality, as selecting a virtual object can produce a sound analogous to pressing keys on a piano or beating a drum (Berthaut, 2020). The evaluated techniques include three using hand-held controllers (Controller Point, Controller Touch), two employing hand-tracking (Hand Point, Hand Touch), and one combining head-tracking and controllers (Gaze Point). These are five of the most used techniques in musical VR applications today, both in research prototypes and commercial applications. We explore questions such as the impact of selection techniques used in commercial systems on timing performance, the effectiveness of different techniques in maintaining tempo and synchronization when using one or two hands, and the applicability of results observed in non-virtual settings to VR. To address these questions, we evaluated the five selected techniques through two timing tasks, which were inspired by studies on sensorimotor synchronization (Repp and Su, 2005). We employed the tapping paradigm (Repp, 2005) adapted to VR, to conduct two empirical studies involving participants (N = 30, 15 with musical background, 15 without). Experiments using this paradigm require participants to tap their fingers with a paced or unpaced auditory signal. In our study, we consider the action of *tapping* as analogous to the selection used in 3D User Interfaces. The experiment used a "*synchronization-continuation*" (Bella et al., 2017) task, which is composed of two phases: at first, participants have to synchronize with a pacing audio signal at different tempi, and then they have to continue in keeping the beat without the auditory stimuli. Participants have to perform this task using their dominant hand and then with alternated hands. Performance analysis included assessing synchronization with a beat, its variance, and the user experience.

Our work aims to answer the following research questions:

- Do selection techniques used in commercial systems influence timing performance?
- Which technique allows to keep tempo constant and synchronize better for one and two hands?
- Are results observed in non-virtual settings valid in VR?

Based on our findings, we provide guidelines to assist designers and researchers in selecting the most appropriate input technique for music creation in VR. Additionally, our work lays the groundwork for further discussions on the associated challenges and opportunities.

The main contributions of our study are summarized as follows:

1. Empirical analysis of the performance of selection techniques in timing tasks;
2. Empirical confirmation and measurement of user experience, workload, and cybersickness;
3. A methodology to effectively compare and understand the effect of 3D input techniques in a musical context.

2. Related work

Our work builds on two main research areas: one related to the studies of sensorimotor synchronization and evaluation of 3D User Interfaces in VR applications, focusing on the musical domain.

2.1. Perceptual timing and tapping research

The experience of music is intimately connected with movement, ranging from the audience nodding their heads at a concert to students practicing with a metronome and musicians synchronizing in an ensemble. An important aspect of this phenomenon is the synchronization of a bodly movement with an external auditory rhythm, which is often called sensorimotor synchronization. This research has a long history, see Repp (2005, 2006a) and Repp and Su (2005) for extensive reviews of the field.

Sensorimotor synchronization is often investigated using the so-called *tapping paradigm*. In these experiments, participants are asked to tap their fingers as regularly as possible in the presence of a pacing stimulus such as an isochronous sequence of tones like a metronome or the beat of a musical excerpt (paced tapping), as well as in the absence of a such stimulus (unpaced tapping). While paced tapping is used to assess the ability to synchronize movements to an external stimulus, unpaced tapping can be employed to assess tapping rate and motor variability (Lorås et al., 2019). Out of several experimental protocols developed for assessing sensorimotor and timing abilities (Fujii and Schlaug, 2013; Bella et al., 2017), one experimental paradigm uses the so-called "*synchronization-continuation*" task (Wing and Kristofferson, 1973; O'Boyle et al., 1996). In this task, participants are first asked to synchronize with an external stimulus and then continue tapping at the same rate after it has stopped.

For our study, we adopted the tapping paradigm using the "*synchronization-continuation*" task in VR and used it for comparing the five input techniques. We assume that selection mechanisms for 3DUIs produce discontinuous and discrete events analogous to the ones produced by finger tapping. Therefore we consider such experimental protocols the most appropriate for investigating timing musical tasks and timing abilities in VR.

In tapping experiments two main timing intervals are measured: the *Inter-Onset Interval* (IOI), which is the temporal distance between two beats, and the *Inter-Tap Interval* (ITI), which is the temporal distance between two consecutive taps. One of the main contributions of tapping research is the study of asynchrony, which is the difference between the occurrence of the tap and the time of the corresponding sound stimulus. It was observed that musicians (especially those with a high level of rhythmic expertise) exhibited less asynchrony and tapping variability compared to non-musicians (Krause et al., 2010).

Tapping research also contributed to uncovering the rate limits (Repp, 2003, 2006b). Research showed that synchronization with an isochronous auditory sequence (i.e., a metronome) or the spontaneous production of a rhythm is possible only within a certain range of IOIs.

Previous studies have identified the upper rate limits as corresponding to an IOI of ≈200 ms, and a lowest rate limit at an IOI of ≈2000 ms (Peters, 1989; Mates et al., 1992). However, it was observed that with an IOIs highest than 1000 ms performance can drastically decrease since it becomes difficult to anticipate the next stimulus as the interval becomes larger. While these rates are valid for people with musical expertise, for people without or with minimal musical training, the range is even narrow, with an upper limit of ≈500 ms and a lowest limit of ≈1000 ms. In our study, we choose five IOIs between 300 ms and 1000 ms, since they fall in the most comfortable rates for both musicians and non-musicians. Furthermore, upper-rate limits are influenced by the maximum frequency of the end effector. For a human finger, this is considered to be ≈500 ms (Repp, 2005). Different apparatuses have been used to capture finger tapping, such as computer keyboard (Ruspantini et al., 2011) and mouse (Zatorre et al., 2007), MIDI controllers (Fujii and Schlaug, 2013), touchscreens (Zanto et al., 2019), motion capture systems (Balasubramaniam et al., 2004), and microphones (Bavassi et al., 2013). Researchers have also developed their own apparatus made with force sensors (van Vugt, 2020), or magnetic sensors (Shima et al., 2009). We should notice that the tapping paradigm was also studied using feet (Numata et al., 2022) and even eye blinking (Bååth et al., 2011). In our study, we are interested in understanding the impact of each technique (e.g. with tracked controller or with tracked hands) in timing tasks. To our knowledge, tapping tasks have not been explored with VR hardware.

Research on tapping is mostly done using one hand, but it was found that bimanual movements are highly adaptive and context-dependent (Swinnen and Wenderoth, 2004). Moreover, coordination between hands plays a role in timing abilities and synchronization (Bugos, 2019). In our study, we investigate selection techniques as both single and double-handed. Lastly, results of tapping research unveiled the importance of tactile feedback for synchronization and timekeeping (Repp, 2005; Repp and Su, 2005). Thanks to multisensory accumulation, humans not only synchronize to the sound they hear but also to the contact they feel while touching a surface. In our study we compare three techniques that provide passive tactile feedback (the buttons on the controller for Gaze Point, Controller Point, and the touch of two fingers together with Hand Point), and two techniques where this feedback is absent (Controller and Hand Touch)

2.2. 3D interaction techniques for selection

The main tasks that users perform in virtual worlds have been classified as selection, manipulation, system control, and navigation of virtual worlds (LaViola et al., 2017). Each of these tasks can be performed with different interaction techniques and input devices.

In our work, we focused exclusively on techniques for selection tasks. The goal of selection techniques is to provide users with means to designate one or several objects in a virtual environment.

Bowman and Hodges (1997) suggested that any selection technique should provide users with a means to indicate an object, confirm its selection, and provide feedback (i.e., visual, auditory) while performing such a task. Poupyrev et al. (1998) divide selection techniques into different levels, according to the metaphor used. At the first level, they distinguish between egocentric (the interaction is in first-person) or exocentric (third-person) techniques. Exocentric techniques are further subdivided into two metaphors: virtual hand and pointer. Egocentric is divided into world-in-miniature and automatic scaling. In the context of our work, we will concentrate on the egocentric metaphor since it is the most prevalent in VR systems using inside-out tracking. Argelaguet and Andujar (2013) recommended that selection techniques should be rapid, accurate, easy to control, and should lower fatigue. Moreover, they extended previous classifications by including components and mechanisms of the selection techniques such as the tools used (i.e., rays, cones, cubes, spheres), the control tool (i.e., hand, head, and viewpoint), degrees of freedom, control display ratio, the relation between motor and visual spaces, and disambiguation mechanisms.

Such taxonomies have been successfully applied to the design and analysis of several 3D interaction techniques in VR. However, still, very little research has been conducted on how to use such techniques in a musical context. While the general goal of achieving accuracy in selection tasks remains essential also in timing tasks, it is not clear if the same requirements of 3D input techniques hold in a musical context.

2.3. 3D selection techniques for musical VR

According to Berthaut, selection techniques are the most basic techniques that can be employed for music in VR (Berthaut, 2020). Following the classification of musical gestures proposed by Cadoz and Wanderley (2000), Berthaut argues that selection techniques can serve as equivalent to musical selection and musical excitation gestures. In the case of musical selection gesture, the selection does not directly affect the sound produced by the virtual instrument but allows a user to choose elements or functions of the instrument (i.e., select a filter or select a sound file). On the other hand, selection gestures can also be analogous to musical excitation gestures (i.e., plucking a string or hitting a percussion). Here, the selection directly produces a sound, such as when triggering the envelope of a synthesizer or when colliding with the key of a virtual piano. In our work, we employed selection gestures as musical excitation gestures.

While in non-music VR research, quantitative comparisons between interaction techniques are fairly common, in the musical domain, such studies are very few and scattered. Mäki-Patola (2005) compared two input techniques for interacting with a virtual drum, using a virtual stick, and a tracked physical stick. By using their dominant hand, participants were asked to follow two rhythmic tracks, one steady (IOI = 500 ms) and one irregular (IOIs = 375–750 ms). The results showed that the tracked stick resulted in better timing accuracy, and in better user experience compared to the virtual one. This seems to be caused by the noticeable latency present in the virtual stick condition, which impacted the general usability of the technique. Berthaut et al. (2011) studied two selection techniques with respect to timing accuracy and error rates. With the first technique, participants had to use a virtual ray to hit a virtual object, like a drum. The second technique used a custom-made 6DOF-tracked input device, named Piivert. Participants could select 3d objects using a virtual ray, and perform an excitation gesture (drum hit) by pressing a force-resistive sensor with their finger. These techniques were evaluated using a task similar to the one used in the previous example but with two different IOIs of 500 ms, and 353 ms respectively. In terms of accuracy, the second technique performed better at the fastest tempo (significant), while the error rate was highest for the virtual drum technique (but not significant). Differently from our study, in these experiments participants were standing in front of stereoscopic projections, wearing shutter glasses, with sound played through an array of loudspeakers.

Reynaert et al. (2021) explored the effect of rhythm for mid-air interactions on gesture regularity, speed, and fatigue. In such a study, the participants wore an HMD and used a tracked hand-held controller with their dominant hand to move a virtual pointer in the presence and absence of auditory pacing stimuli. The stimuli involved two tempi of IOI 1000 and 700 ms. The results showed that the fastest tempo caused more arm fatigue, whereas the slowest one increased the perceived feeling of success.

Our work differs from previous studies not only for the apparatus used but mostly for the diverse types of techniques investigated (single and two-handed), the type of stimuli (five IOIs), the use of established methods from the research on auditory sensorimotor and timing activities, as well as standard usability metrics such as workload, cybersickness, and ease of use.

3. Design rationale

Of the five selection techniques, three of them use 6-DoF tracked hand-held controllers, and the other two are based on hand-tracking. Nowadays, tracked controllers are widely used and represent one of the main means of input in VR. While controllers are intuitive and precise, they require users to have them available and need to be powered and charged. Conversely, bare hands allow more natural, direct selection techniques based on real-world gestures. Previous research in non-musical contexts showed that controller-based interactions appear more responsive than free-hand interactions (Caggianese et al., 2019; Dudley et al., 2019).

Both controllers and virtual hands can be used for direct or distant selection (Figueiredo et al., 2018). Direct selection is performed by the 3D representation of either hands or controllers used to collide with the virtual object a user wants to select. This method is very realistic since the end-effector (being the avatars of the user's hands or of the controller) collides with the virtual object a user wants to select. However, since the control space of the user matches their motor space, this method limits the selection to objects that can be reached only manually.

To overcome the limitations, virtual controllers and hands can be used for distant selection, through the use of virtual rays and mechanisms of "point and commit". Pointing is usually achieved with a virtual ray that is projected forward starting from the end-effector. To confirm the selection, different mechanisms have been proposed. When using the controllers, users can press a button or a trigger, for free-hand different hand gestures such as pinch can be used (Wingrave et al., 2005). While convenient and less physically demanding, distant selection techniques are prone to the so-called "Heisenberg Effect" (Bowman et al., 2001), which can be observed when the position of the virtual ray changes after the selection is confirmed, leading the user to an increasing sense of uncertainty.

A third method we included involves the tracked head as a pointer. This is used as an approximation of where the user is looking. Such type of interaction is found to be faster and effortless compared to manual input only (Pfeuffer et al., 2020). However, there are several issues. First, it can cause neck strain, especially if targets are sparse (Choe et al., 2019). Second, this technique is prone to the "Midas Touch" problem (Jacob, 1995), i.e., the unintended selection of virtual objects. To overcome this problem, the gaze is usually used as a pointer, and a controller's button is used for confirmation. Therefore, the choice between these selection methods depends on the target application and the user experience. For instance, previous research on text entry has shown that when typing on a virtual keyboard in mid-air, selection techniques employing tracked controllers (especially those utilizing a virtual ray) outperform controller-free techniques (i.e., tracked hands) in both user experience and performance (Speicher et al., 2018; Xu et al., 2019).

Based on this rationale we selected five input techniques. Three use hand-held controllers as selection tools and two use bare hands. Of these, three use virtual rays for distant selection, and two for direct selection.

4. Materials and methods

To compare the five interaction techniques we conducted two controlled laboratory experiments. In the first experiment, participants tested the different techniques using their dominant hand, performing an in-phase tapping. In the second, they tested the techniques using two hands with anti-phase tapping (e.g., alternating the hands). The experiments followed the same methodology and were presented to subjects in random order.

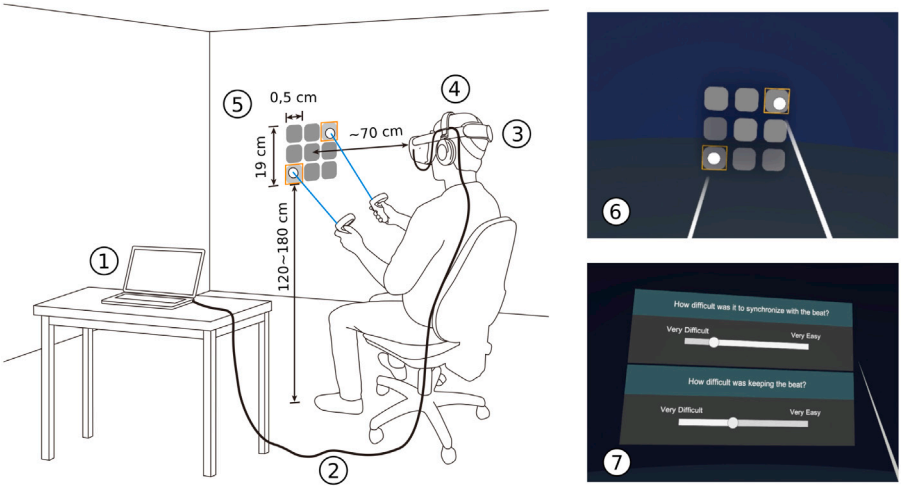

Fig. 2. This figure illustrates the experimental setup and the apparatus used: (1) the laptop used to run the control software, (2) the Meta Link cable, (3) a Meta Quest 2 HMD, (4) stereo headphones, (5) and (6) the virtual multi-touch pads, (7) one of the questionnaires the participant had to fill while in VR.

4.1. Design

The experiments followed a mixed within- and between-subjects design. We involved four independent variables (i.e., interaction method, IOIs, number of hands, and musical expertise) and three dependent variables related to the user performance (i.e., mean asynchrony, variance for synchronization, variance for continuation) as well as related to user's preferences such as the perceived difficulty for each IOI, workload, and cybersickness of each technique. The order of presentation of the input method condition and the IOIs was randomized using a Latin square. Musical expertise was treated as the between-subject factor.

4.2. Participants

A total of 30 participants (6 females, 24 males; aged between 20 and 40, mean = 27.5, SD = 5.8) volunteered for the experiments. They were recruited at the campus of the University of Trento through a mailing list. Out of these participants, 3 were left-handed, and 27 were right-handed. Participants reported to have no hearing or sensorimotor impairment. Ten wore glasses. Participants belonged to either groups, musicians or no musicians (each group n = 15). We considered musicians participants with > 3 years of musical expertise, non-musicians never played a musical instrument.

4.3. Apparatus and setup

Fig. 2 shows the setup of the experiment and its components. The VR system used a Meta Quest 2 headset connected using a Meta Link cable (length: 5 m, bandwidth: 5 Gbps) to a standard laptop computer running Windows 10, with an AMD Ryzen 9 CPU, 32 GB RAM, and an Nvidia GeForce 3060 graphics card. For the auditory feedback, the participants wore a pair of headphones (Beyerdynamic DT 770 PRO 80 Ohms) connected to the HMD. The control software was developed with Unity 2020.3.13 (Unity). To implement the five techniques we used the Interaction SDK from the Oculus Integration package (v. 44.0) (Oculus Integration). The experimenter controlled the software using a custom .NET application through the OSC protocol (OSC).

4.4. Virtual environment and virtual multi-touch pad

The virtual environment consisted of an empty space designed as less distracting as possible. Placed in front of the user, there was a virtual multi-touch pad controller, which was the surface used by subjects to perform the experiment. We chose a virtual representation of this kind of controller because, in its physical form, it is widely used in both studios and live performances. It is composed of a grid of square-shaped pressure-sensitive touch pads. By pressing one of the pads, a musician can control the parameters of a synthesizer or trigger different types of events, such as sound samples. Several VRMIs made use of 3D UIs inspired by such types of controllers (Zappi et al., 2010; Men and Bryan-Kinns, 2018; Valbom and Marcos, 2005; Cabral et al., 2015; Fillwalk, 2015; Virtuoso VR). We designed a virtual multi-touch pad controller composed of a 3 × 3 grid of cube-shaped objects representing the pads. Since there are no standard arrangements of pads in both virtual and physical touch-pad controllers, we adopted the configuration used by Choe et al. in their study (Choe et al., 2019). The arrangement and size of each pad was 50 mm with a spacing of 25 mm. These were obtained from recommendations found in the literature (Choe et al., 2019; Park et al., 2020; Figueiredo et al., 2018). However, after a pilot study, we decided to increase the size and spacing to avoid potential conflicts when two controllers are used together that might collide with each other. The Virtual Multi-touch Pad Controller was placed in the interaction zone of the participants at 100 cm–150 cm, within an arms reach of around 70 cm (Bachynskyi et al., 2015). When the participant selected a pad, a sound was triggered. The sound was produced with a frequency-modulated oscillator (main frequency of 110 Hz) with a very short attack and release, designed to sound similar to the produced by a percussion. It was developed using the Faust programming language (Faust) and then compiled as a plug-in for Unity. The sound was identical for all pads. In addition to the auditory feedback, we included a minimal form of visual feedback: when selected the pad turned from gray to white. To signal which pads the participants have to select, we drew an outline colored in yellow.

4.5. Selection techniques

We illustrate the implementation of the aforementioned five techniques for selection as follows.

- **Gaze Point (GP)**: With this technique, participants have to select the target virtual pads by moving the head, which acts as an approximation of their gaze. A virtual ray is cast from the center of the HMD. When the ray intersects a virtual pad, a circular cursor (diameter of 70 mm) is drawn on the surface. Confirmation is performed using the Meta Quest Touch Controllers' buttons: "A" for the right hand and "B" for the left hand. Head tracking is provided by the Oculus Integration SDK. GP is considered a distant selection technique and is best used for interacting with virtual

objects that are out of reach. Moreover, GP is widely used in several VR systems as one of the most basic selection methods (Gaze Cursor Component; Gaze and Commit). While GP has not been fully explored in immersive musical applications (Lucas Bravo and Fasciani, 2023), head tracking is a consolidated interaction modality (Davanzo and Avanzini, 2020) for both selecting notes on virtual keyboards (Wiederhold et al., 2016; Davanzo et al., 2021), and triggering sound samples (Kapur et al., 2004; Bardos et al., 2005).

- **Controller Point (CP)**: With this technique, the hand-held tracked controllers are used for both actions of pointing and confirmation. When controllers are available, this represents one of the most used selection techniques in musical and non-musical VR applications. Participants utilized 3D replicas of the Quest Touch controllers to manipulate a virtual ray, aiming it at the target pads. The ray is drawn using a linear gradient (see 2, 6). Similarly to GP, when the ray points to a virtual pad, a cursor is drawn, with a diameter of 70 mm. The selection mechanism mirrors the one of GP and uses the buttons on the Quest controllers. The 3D model of the controllers used is the one provided by the Oculus Integration SDK. CP has been explored in several VR musical applications for selecting functions of virtual synthesizers and triggering sound samples through 3D widgets (Patch XR; Costa et al., 2019; Kelly and Klipfel, 2017; Wakefield et al., 2020; Valbom and Marcos, 2005). A virtual ray extends from the top part of the virtual replicas of the tracked controllers held by the users. The controller can be moved and rotated in the 3D space to select a virtual surface. When the ray intersects it, a cursor is drawn. Then, to confirm the selection users press one of the buttons on the physical hand-held controllers.
- **Controller Touch (CT)**: This technique provides one of the most isomorphic and potentially realistic interactions. Participants are required to poke the virtual pads using the three-dimensional representation of the tracked handheld controllers. The collision with the virtual pads occurred on the backside of the top part of the virtual controller. This design choice was made to restrict the interaction area and make it akin to the sensation of touching a real surface, with the part of the controller facing backward being the initial point of contact with the surface. The 3D model of the controllers used is the one provided by the Oculus Integration SDK. This technique is widely used in VR musical applications to select functions of virtual instruments, trigger sound samples, loop audio tracks, and play with virtual drums and pads (Virtuoso VR; Drum Beats VR; EXA Infinite Instrument; Lyra VR). In a common variation of this technique, the tracked controllers are represented as a mallet or a drumstick (Mäki-Patola, 2005; The Music Room; Çamcı et al., 2020).
- **Hand Point (HP)**: This technique shares the same "*point and commit*" interaction mechanics as GP and CP. However, instead of controllers, it tracks the users' hands in real time. HP is often used in commercial headsets as a way to allow distant selection without the use of tracked controllers. Participants use the 3D representation of their hands to point a virtual ray extending from the center of the palms. When the light intersects the pads, a 2D circular cursor is drawn. The diameter of the cursor is 70 mm. Selection of the pads is done by performing a pinch gesture. Gesture recognition was implemented using the functions provided by the Oculus Interaction SDK. Virtual models of the hands were also obtained from the same SDK. This technique was already explored in early studies using sensorized gloves (Wingrave et al., 2005). Moreover, this technique can be found in several commercial VR systems (Distance Hand Grab Interaction; Interact with Objects Remotely), such as the "Point and Commit with Hands" model of the MRTK (Point and commit with hands). In our case, the pinch gesture is particularly interesting since it resembles the task of finger tapping (Shima et al., 2009; Morimoto et al., 2018; Sugioka et al., 2022). Recent works in the context of XR musical interfaces have started to explore such techniques for manipulation of virtual sound controllers (Bilbow, 2022), and for selection of 3D GUI elements (i.e., filters and pitch) (Wang and Martin, 2022).
- **Hand Touch (HT)**: Together with CT, this represents the second most isomorphic technique we used in the experiments. Participants use their virtual hands to directly touch the virtual pads. To avoid potential conflicts, the pads can be activated only when the tip of the index finger collides with them. Similarly to HP, we implemented HT using the hand-tracking capabilities provided by the Interaction SDK Quest 2. Thanks to the rapid improvements in computer vision and hardware design, inside-out hand tracking is becoming one of the preferred ways to interact with virtual environments (Reimer et al., 2023). In existing musical VR applications, HT was used mostly for functional purposes such as selecting 3D sound sources (Naef and Collicott, 2006; Bilbow, 2022; Wang and Martin, 2022), but also for triggering notes (Fillwalk, 2015; Moore et al., 2015) and sequencers (Men and Bryan-Kinns, 2018) on virtual instruments.

4.6. Pacing tones

The tones were produced with MIDI files written with Musescore (v. 3.6.2) (Musescore), composed of 6 bars for "synchronization" and 6 for "continuation" (at 4/4, quarter note = IOI). To emulate a metronome, we used a "woodblock" sound, with the first tone of each bar (downbeat) having a frequency of 659.26 Hz (MIDI 76) and the other three beats of 698.46 Hz (MIDI 77). The MIDI files were played in Unity using the Maestro MIDI Player Toolkit (v. 2.89.2) (Maestro MIDI).

4.7. Task

The experiment uses the "synchronization-continuation" paradigm (Wing and Kristofferson, 1973; Merchant et al., 2011) to test the asynchrony and the tapping variance with and without a pacing stimulus. First, the participants were instructed to tap to a series of 40 beats of a digital metronome presented isosynchronously at five tempi (quarter note IOI = 1000, 667, 500, 400, and 333 ms). After the last beat was heard, participants were asked to continue tapping at the same rate for a duration corresponding to 40 IOIs in the absence of the pacing stimuli. The choice of the five IOIs was guided by two factors. First, they fell into the range of the most widely used tempi in popular music, as they correspond to 60, 90, 120, 150, and 180 beats per minute (BPM). Second, they are contained in the range between the upper and lowest rate limits found in the literature for musicians and non-musicians (Repp, 2005). The "synchronization-continuation" task is illustrated in Fig. 3. Sample excerpts of the tasks can be viewed in the video in Fig. 4

4.8. Procedure

The experiment was preceded by a verbal introduction given by the experimenter, who explained to the participants the aims and goals of the study. After, participants signed a consent form. They were then instructed to sit comfortably on a chair while performing the experiments. Then, with the help of the experimenter, they wore both the Quest 2 headset and headphones and the experiment started. A short video tutorial was shown to make the participants familiar with the headset and its controllers. Each experiment was composed of the participant testing the five interaction techniques one after the other, in a block-randomized order (counterbalance was achieved using Latin square). Before each condition, the interaction was explained using a 40 s video tutorial and practiced in a warm-up phase for about 2 min. Participants received only minimal instructions about the functionalities of the different interaction techniques. We always counterbalanced the conditions of the task. The "synchronization-continuation" task was

Fig. 3. This figure illustrates the "*synchronization-continuation*" paradigm used in the timing task.

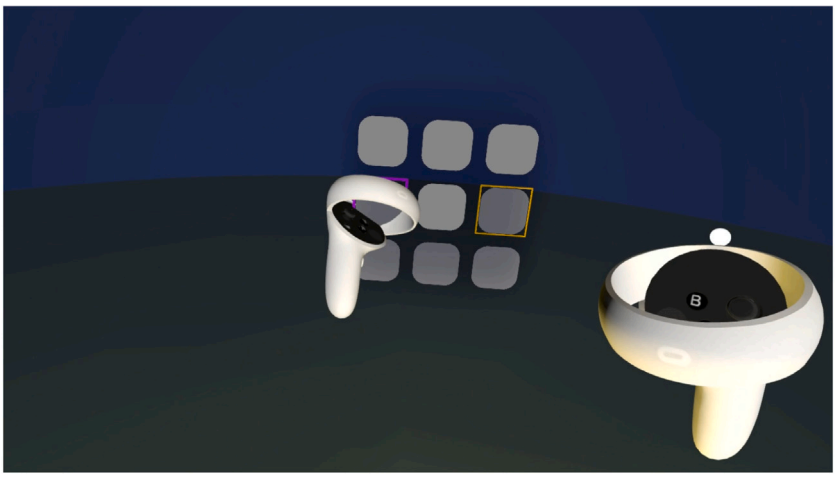

Fig. 4. An image showing the VR environment during the experiment. A video demonstration of the different techniques be accessed https://youtu.be/BKVxxnB0kBshere. Every technique is presented for the Unimanual and Bimanual conditions, with different IOIs.

repeated five times, one for each target IOI. The order of presentation of the IOIs was counterbalanced for each condition. In addition, for each trial, the highlighted target pads were also randomized. The total number of trials for the "synchronization-continuation" task amounted to 1500: 30 participants × 5 input methods *x* 5 IOIs *x* 2 hand conditions.

For the synchronization part of the task, we collected the mean asynchrony and the variance of the ITI with respect to the target IOI. The asynchrony was derived by subtracting the timing of the onset of the recorded tap from the onset of the metronome sound. The more the value of the asynchrony is closed to zero, the better a technique performs. Similarly, for the continuation part, we collected the variance of the ITI. It was also calculated the variance between the ITI produced by the participants and the target IOI. The more the variances are close to zero, the more a technique helps to produce a precise and steady beat. The results of each trial (regarding both asynchrony and variance) for conditions technique and hands were averaged across all IOIs, normalized, and the absolute value for each technique was obtained.

At the end of each session, participants were asked to answer two questions by moving a virtual slider on an 11-point Likert scale ranging from 1 (very easy) to 11 (very difficult). The first question referred to the synchronization phase, and the second corresponded to the continuation phase:

- *How difficult was it to synchronize with the beat?*
- *How difficult was keeping the beat?*

At the end of each condition, the participants had to fill out the NASA TLX (Hart, 2006) and the SSQ questionnaires (Bruck and Watters, 2009) in VR. All questionnaires were implemented using the VR Questionnaire Toolkit (Feick et al., 2020), which was modified to be compatible with Oculus Integration.

After all the tasks were completed, for both the Unimanual and Bimanual conditions, the participants were asked to take off the headset. After a 2-minute break, there was a brief session with a verbal interview with the experimenter composed of open-ended questions. Participants' answers were recorded using a digital audio recorder. Finally, participants had to fill out a questionnaire on a laptop to collect demographic data. On average participants took about 90 min to complete the experiment, including two breaks of 5 min each between the conditions, where participants had to take off the HMD and headphones.

4.9. User study hypothesis

Despite the highly investigative nature of our study, we defined four expected outcomes.

- **H1**: *Musicians will exhibit low asynchronies and variances*.
  Studies comparing musicians and non-musicians have shown that musical training can improve rhythmic perception and production (Repp, 2005). Especially, tapping studies showed that musicians showed more accuracy in motor timing (Franěk et al., 1991; Scheurich et al., 2018), but also smaller asynchronies, and lowest variability (Repp, 2010). We then hypothesize that musicians will exhibit low asynchronies and variances compared to non-musicians.
- **H2**: *Compared to the Unimanual condition, the Bimanual condition will result in better performance*.
  Musical VR applications do not enforce the use of Unimanual or Bimanual techniques, since they depend on the context and the use (Swinnen and Wenderoth, 2004). When comparing different interaction techniques, we might expect different mechanics to influence the outcome. According to previous studies, tapping with both hands in alternation can help to overcome limitations of the end effector that can appear when tapping with a single hand (Pressing and Jolley-Rogers, 1997; Repp, 2005). We then hypothesize that performance will improve in the Bimanual condition.

Table 1
Results of the statistical analysis for the synchronization-continuation task. We report the main effect and interactions among factors.

Synchronization-continuation task				
		Synchronization		Continuation
		Absolute mean asynchrony	Absolute variance	Absolute variance
Condition	Factor	Main effect	Main effect	Main effect
One hand	Technique	F(4,712) = 9.9 ***	F(4,712) = 11.1 ***	F(4,712) = 13.7 ***
	Musical expertise	F(1,28) = 6.1 *	–	–
	Technique – musical expertise	–	–	–
Two hands	Technique	F(4,712) = 9.9 ***	F(4,712) = 9.6 ***	F(4,712) = 13.7 ***
	Musical expertise	F(1 28) = 6.1 *	–	–
	Technique – musical expertise	–	–	–
Musicians	Technique	F(4,726) = 12 ***	F(4,726) = 7.2 ***	F(4,726) = 14.5 ***
	Hand	F(1,726) = 8.4 **	–	F(1,726) = 6.2 *
	Technique – hand	F(4,726) = 6.7 ***	F(4,726) = 2.8 *	–
Non-musicians	Technique	F(4,726) = 12 ***	F(4,726) = 7.8 ***	F(4,726) = 14.5 ***
	Hand	F(1,726) = 8.4 **	F(1,726) = 7.5 **	F(1,726) = 6.2 *
	Technique – hand	F(4,726) = 6.7 ***	F(4,726) = 3.1 *	–

Table 2
Mean of the miss ratio for each technique with respect to the experimental conditions.

Miss ratio (%)				
Technique	Unimanual		Bimanual	
	Musicians	Non-musicians	Musicians	Non-musicians
GP	13.41	14.67	18.07	20.67
CP	13.48	14.48	15.41	15.93
CT	13.52	13.52	14.48	16.93
HP	17.04	16.65	18.58	21.44
HT	16.85	17.56	18.07	19.67

- **H3**: *The techniques that provide tactile feedback will lead to better time accuracy than those that do not.*
  Non-VR research has highlighted the fundamental role of tactile feedback for timekeeping and synchronization tasks (Repp, 2005; Repp and Su, 2005). Moreover, research on mid-air interactions showed that pseudo-haptics and self-haptics could improve user experience in VR (Batmaz et al., 2019; Kim and Xiong, 2022). We hypothesize that techniques such as GP, CP, and HP that include some forms of passive haptics and self-haptics will result in better performance than those without tactile feedback.
- **H4**: *The techniques that require less workload will perform better and are perceived as easy to use.*
  If the interaction with a product leads to a high perceived workload, this can impact the user experience (Jerald, 2015; LaViola et al., 2017). Therefore, we hypothesize that techniques rated with less workload will be considered more easy to use.

5. Results

For each response variable regarding each task (i.e., Absolute Mean Asynchrony, Absolute Variance, the score of post-task questions 1 and 2) an ANOVA was performed on a linear mixed effect model. For the total score for NASA TLX and its six questions, and the total score for SSQ and its sixteen questions, an ANOVA was performed using a generalized linear mixed effect model. These models had the subject as a random factor, and the response variable, the hand condition (i.e., Unimanual or Bimanual), the technique (GP, CP, CT, HP, HT), and musical expertise (i.e., musician and non-musician) as fixed factors. Post hoc tests were performed on each fitted model using pairwise comparisons adjusted with the Tukey correction. The assumption of normally distributed residuals was visually verified.

For clarity, the results are presented in tabular and figure formats. Figures show the mean for each technique and the results of the pairwise comparison, in respect of each variable. Tables present the results of the statistical analysis (i.e., main effect and interactions between factors) found after the statistical analysis. The color coding corresponds to orange for subjects with musical expertise (musicians) and blue for subjects without musical expertise (non-musicians). We present pairs such as: on the right-most column the technique with the highest mean, and on the column on its left, the techniques with the lowest mean of the group of techniques analyzed. Respectively, they are highlighted with dark and light green respectively. In both figures and tables, we indicate * for $p < 0.05$, ** for $p < 0.01$, and *** for $p < 0.001$.

5.1. Task: Synchronization and continuation

Table 1 summarizes the results of the statistical analysis. Before the analysis, we discarded the first four ITI for synchronization and the first four for continuation. For the synchronization phase, we removed all ITIs that were twice as large as the target ITI. This represents the miss ratio. Table 2 presents the mean miss ratio for each technique, expressed in percentage.

5.1.1. Synchronization phase: Absolute mean asynchrony

Fig. 5 presents the results regarding the synchronization part of task. Altogether, the results for both Unimanual and Bimanual conditions show that musicians exhibited less asynchrony compared to non-musicians. This is a well-known result in the literature on sensorimotor synchronization (Krause et al., 2010), however, we found no statistical difference between the two groups. As shown in Table 1 we found a significant main effect for factor technique ($p < 0.001$) in both conditions hands and musical expertise. A significant interaction effect ($p < 0.001$) was found between conditions technique and hand for musicians and non-musicians. For the Unimanual condition, post hoc analysis revealed a difference ($p = 0.0058$) between CT ($M = 6.07\%$, $SD = 0.47$) and HT ($M = 8.9\%$, $SD = 0.9$) for musicians. Regarding non-musicians, we found a difference ($p = 0.0004$) between CT ($M = 8.2\%$, $SD = 0.6$) and HP ($M = 12.2\%$, $SD = 0.9$). For the Bimanual condition, we found a difference ($p \leq .0001$) between CT and GP for both musicians (CT: $M = 5.6\%$, $SD = 3.2$; GP: $M = 10.8\%$, $SD = 7.6$) and non-musicians (CT: $M = 8.6\%$, $SD = 5$; GP: $M = 13.8\%$, $SD = 8.2$). Results of the pairwise comparison are shown in Table 3. Therefore in both Unimanual and Bimanual conditions, CT emerged as the technique that exhibited less asynchrony. Overall, the difference between techniques is more pronounced in the Bimanual condition. In the Unimanual such a difference is noticeable more for musicians compared to non-musicians. Moreover, for Unimanual there is no major difference between the other controller-based techniques (GP, CP), but there is with hand-tracking techniques such as HP and HT. For the Bimanual condition, differently for Unimanual, GP stands out as the technique with the largest asynchronies. This is further confirmed by a statistical difference found for GP between Unimanual and Bimanual for both musicians and non-musicians. Results are presented in Table 4. We found no differences between musical expertise.

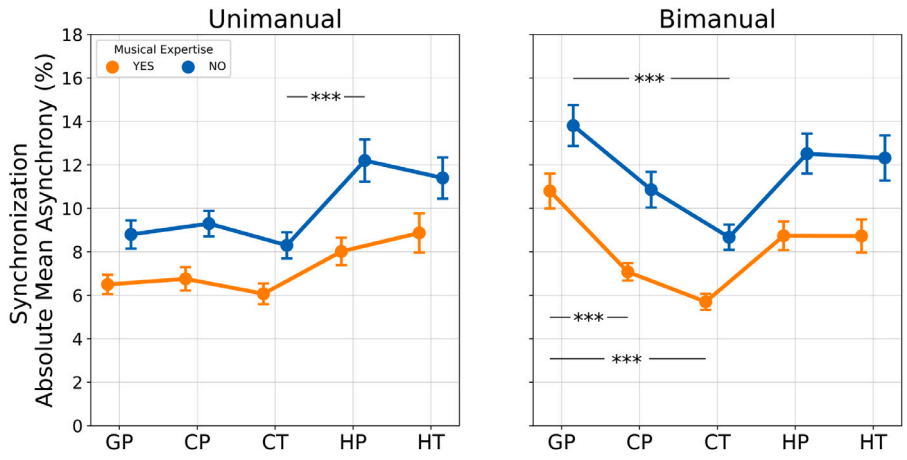

Fig. 5. The absolute mean asynchrony for Unimanual and Bimanual conditions, in respect to each interaction technique. Lowest values mean less asynchrony. Error bars represent the standard error. Only the most significant pairs ($p < 0.001$ ***) are reported. For details, refer to Tables 3 and 4.

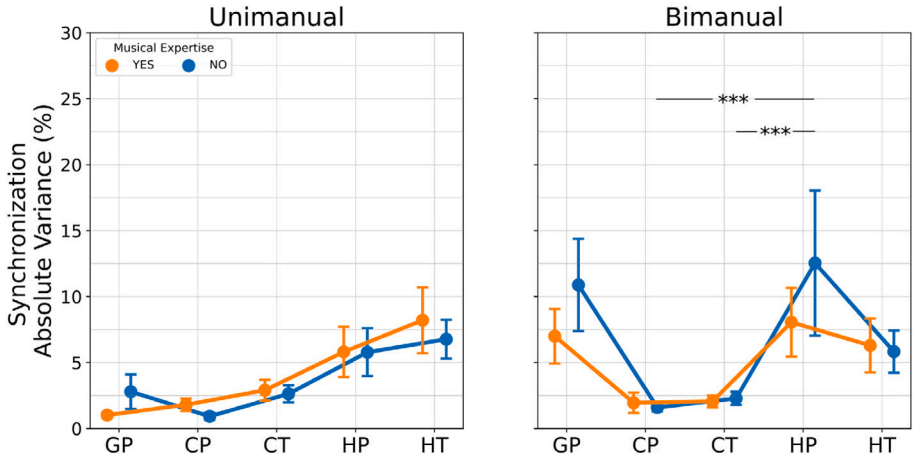

Fig. 6. The absolute variance for Unimanual and Bimanual conditions regarding the synchronization phase with respect to each interaction technique. Lowest values mean less variance. Error bars represent the standard error. Only the most significant pairs ($p < 0.001$ ***) are shown. Refer to Tables 5 and 6 for details.

Table 3
The results of the pairwise comparisons for the Mean Asynchrony (synchronization phase) regarding the hand condition. The color orange refers to participants with a musical background, and blue to participants without a musical background. Dark green indicates the values with the lowest means of the group, while light green indicates the techniques with the highest mean.

Synchronization - Absolute mean asynchrony				
Condition hand				
		Low mean	High mean	*p*-value
Unimanual		CT	HT	0.0058 **
		CT	HP	0.0004 ***
		GP	HP	0.0055 **
		CT	HT	0.0164 *
		CP	HP	0.0339 *
Bimanual		CT	GP	<0.0001 ***
		CP	GP	<0.0001 ***
		CT	HP	0.0014 **
		CT	HT	0.0014 **
		CT	GP	<0.0001 ***
		CT	HP	0.0055 **
		CT	HT	0.0115 **

Table 4
The table presents the results of the pairwise comparisons for the Mean Asynchrony (synchronization phase) regarding the musical background condition. The color orange refers to participants with a musical background, and blue to participants without a musical background. Dark green indicates the values with the lowest means of the group, while light green indicates the techniques with the highest mean.

Synchronization - Absolute mean asynchrony			
Condition musical expertise			
	Unimanual	Bimanual	*p*-value
	GP	GP	<0.0001 ***
	GP	GP	<0.0001 ***

Table 5
The table presents the results of the pairwise comparisons for the Absolute Variance (synchronization phase) regarding the hand condition. The color orange refers to participants with a musical background, and blue to participants without a musical background. Dark green indicates the values with the lowest means of the group, while light green indicates the techniques with the highest mean.

Synchronization - Absolute variance				
Condition hand				
		Low mean	High mean	*p*-value
Unimanual		GP	HT	0.0062 **
Bimanual		CP	HP	0.0002 ***
		CT	HP	0.0007 ***
		CP	GP	0.0036 **
		CT	GP	0.0108 *

5.1.2. Synchronization phase: Absolute variance

As shown in Table 1, a significant main effect was found for condition technique for both hands and musical expertise ($p < 0.001$). Fig. 6 shows the results for each technique, and Table 5 the results of the pairwise comparison for hand condition.

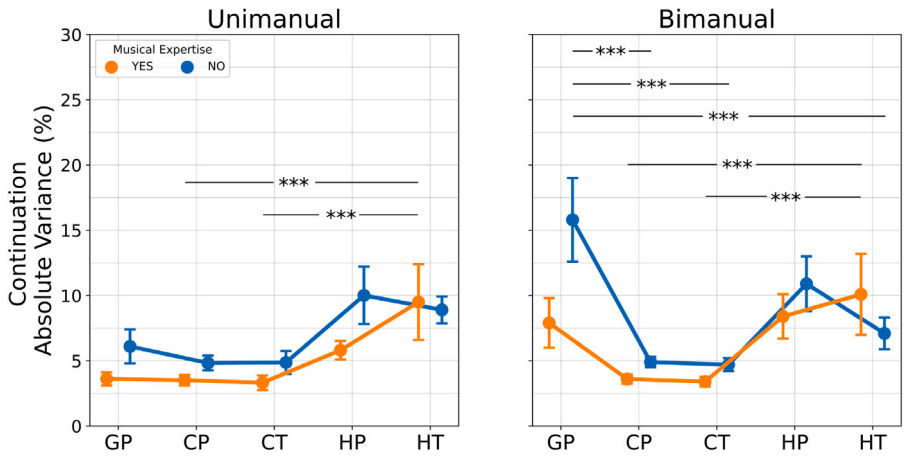

Fig. 7. The absolute variance for Unimanual and Bimanual conditions regarding the continuation phase, in respect to each interaction technique. Lowest values mean less variance. Error bars represent the standard error. Only the most significant pairs ($p < 0.001$ ***) are shown. See Tables 7 and 8 for details.

Table 6
The table presents the results of the pairwise comparisons for the Absolute Variance (synchronization phase) regarding the musical background condition. The color orange refers to participants with a musical background, and blue to participants without a musical background. Dark green indicates the values with the lowest means of the group, while light green indicates the techniques with the highest mean.

Synchronization - Absolute variance			
Condition musical expertise			
	Unimanual	Bimanual	*p*-value
	GP	GP	0.0232 ***
	GP	GP	0.0226 ***

For the Unimanual condition, post hoc analysis revealed a statistical difference ($p = 0.0062$) between GP ($M = 1.02\%$, $SD = 0.9$) and HT ($M = 8.2\%$, $SD = 9.9$). We found no difference between techniques for non-musicians. Differently, for the Bimanual condition, we found no differences for participants with musical expertise. For non-musicians, we found a high statistical difference ($p = 0.0002$) between HP ($M = 12.5\%$, $SD = 25.5$), and CP ($M = 1.6\%$, $SD = 1.3$). We also found a statistical difference for GP between Unimanual and Bimanual (musicians: $p = 0.0232$; non-musicians: $p = 0.0226$) as shown in Table 6. No significant difference was found between musical expertise.

5.1.3. Continuation phase: Absolute variance

Fig. 7 shows the results of the results for each technique. As presented in Table 1 we found a significant main effect for condition technique for both hands and musical expertise ($p < 0.001$). The results of the pairwise comparisons are presented in Table 7 for the hand condition.

For the Unimanual condition, we found a statistical difference regarding participants with musical expertise ($p = 0.0005$) between HT ($M = 9.5\%$, $SD = 11.2$) and CT ($M = 3.3\%$, $SD = 2.1$). For participants without musical expertise, we found a difference ($p = 0.0075$) between techniques HP ($M = 10.1\%$, $SD = 8.6$) and CP ($M = 4.8\%$, $SD = 2.1$). In the Bimanual condition, we found a difference ($p = 0.0001$) between CT ($M = 3.4$, $SD = 1.4$) and HT ($M = 10.1\%$, $SD = 8.7$) for musicians, while for non-musicians the difference ($p \leq 0.001$) was found between CT ($M = 4.7$, $SD = 1.9$) and GP ($M = 15.8\%$, $SD = 12.5$). Similarly to synchronization, we found a statistical difference for GP between the two hand conditions (see Table 8). The difference is higher for non-musicians ($p < 0.001$), compared to musicians ($p < 0.058$). However, we found no statistical difference between musical expertise.

5.2. User preference

After each trial, we asked participants to rate their experience regarding the synchronization and the continuation task. To provide a more comprehensive understanding, we averaged the results for each technique across all IOIs for each technique and applied the analysis described above. Results are presented in Table 9.

Table 7
The table presents the results of the pairwise comparisons for the Absolute Variance (continuation phase) regarding the hand condition. The color orange refers to participants with a musical background, and blue to participants without a musical background. Dark green indicates the values with the lowest means of the group, while light green indicates the techniques with the highest mean.

Continuation - Absolute variance				
Condition hand				
		Low mean	High mean	*p*-value
Unimanual		CT	HT	0.0005 ***
		CP	HT	0.0008 ***
		GP	HT	0.0012 **
		CP	HP	0.0075 **
		CT	HP	0.0245 *
Bimanual		CT	HT	0.0001 ***
		CP	HT	0.0002 ***
		CT	HP	0.0134 *
		CP	HP	0.0213 *
		CT	GP	<0.0001 ***
		CP	GP	<0.0001 ***
		HT	GP	<0.0001 ***
		HT	HP	0.0011 **
		CP	HP	0.0038 **
		CT	HP	0.0086 **

Table 8
The table presents the results of the pairwise comparisons for the Absolute Variance (continuation phase) regarding the musical background condition. The color orange refers to participants with a musical background, and blue to participants without a musical background. Dark green indicates the values with the lowest means of the group, while light green indicates the techniques with the highest mean.

Continuation - Absolute variance				
Condition musical expertise				
	Unimanual	Bimanual	P-value	
	GP	GP	0.058	(*)
	GP	GP	<0.0001	***

5.2.1. "How difficult was it to stay in sync with the beat?"

Fig. 8 shows the results of the first post-task questionnaire. Among all, CT is the technique that was rated lowest (easier to synchronize) by both musicians and non-musicians, in Unimanual ($M = 1.7$, $SD = 1.8$; $M = 2.1$, $SD = 1.8$) and Bimanual ($M = 2.5$, $SD = 2.3$; $M = 2.9$, $SD = 2.6$) conditions. However, the techniques with the highest rating (difficult to synchronize) are more diverse. For Unimanual condition,

Table 9
Results of the statistical analysis for the questionnaires regarding the synchronization-continuation task. We report the main effect and interactions among factors.

Synchronization-continuation task			
		Post-task	
		Questionnaire 1	Questionnaire 2
Condition	Factor	Main effect	Main effect
One hand	Technique	F(4,712) = 31.5 ***	F(4,712) = 32.0 ***
	Musical expertise	–	F(4,712) = 7.4 ***
	Technique – musical expertise	F(4,712) = 3.1 *	–
Two hands	Technique	F(4,711.01) = 56.7 ***	F(4,711) = 39.8 ***
	Musical expertise	–	–
	Technique – musical expertise	–	–
Musicians	Technique	F(4,725) = 43.5 ***	F(4,725) = 42.3 ***
	Hand	F(1,725) = 41.9 ***	F(1,725) = 17.3 ***
	Technique – hand	–	F(4,725) = 8.5 ***
Non-musicians	Technique	F(4,726) = 16.2 ***	F(4,726) = 15.5 ***
	Hand	F(1,726) = 37.7 ***	F(1,726) = 17.6 ***
	Technique – hand	–	F(4,726) = 10.3 ***

Table 10
The table presents the results of the pairwise comparisons for the Post-task questionnaire 1 regarding the hand condition. The color orange refers to participants with a musical background, and blue to participants without a musical background. Dark green indicates the values with the lowest means of the group, while light green indicates the techniques with the highest mean.

Post-task questionnaire 1				
Condition hand				
		Low mean	High mean	*p*-value
Unimanual		CT	HP	<0.0001 ***
		GP	HP	<0.0001 ***
		CP	HP	<0.0001 ***
		CT	HT	<0.0001 ***
		HT	HP	0.0022 **
		CT	CP	0.0248 *
		CT	HT	<0.0001 ***
		GP	HT	0.0043 *
		GP	HP	0.0055 *
Bimanual		CT	GP	<0.0001 ***
		CP	GP	<0.0001 ***
		HT	GP	<0.0001 ***
		CT	HP	<0.0001 ***
		CP	HP	<0.0001 ***
		HP	GP	0.0024 **
		HT	HP	0.0163 *
		CT	HT	0.0119 *
		CT	GP	<0.0001 ***
		CP	GP	<0.0001 ***
		HT	GP	<0.0001 ***
		CT	HP	0.0014 **
		HP	GP	0.0012 **
		CP	HP	0.0124 **

Table 11
The table presents the results of the pairwise comparisons for the Post-task questionnaire 1 regarding the musical background condition. The color orange refers to participants with a musical background, and blue to participants without a musical background. Dark green indicates the values with the lowest means of the group, while light green indicates the techniques with the highest mean.

Post-task questionnaire 1			
Condition musical expertise			
	Unimanual	Bimanual	*p*-value
	GP	GP	<0.0001 ***
	GP	GP	<0.0001 ***

Table 12
The table presents the results of the pairwise comparisons for the Post-task questionnaire 2 regarding the hand condition. The color orange refers to participants with a musical background, and blue to participants without a musical background. Dark green indicates the values with the lowest means of the group, while light green indicates the techniques with the highest mean.

Post-task questionnaire 2				
Condition hand				
		Low mean	High mean	*p*-value
Unimanual		CT	HP	<0.0001 ***
		CT	HT	<0.0001 ***
		CT	GP	<0.0001 ***
		GP	HP	<0.0001 ***
		CP	HP	<0.0001 ***
		HT	HP	<0.0001 ***
		CT	CP	0.0002 **
		CP	HT	0.0007 ***
		CT	HT	0.001 **
		CP	HP	0.0016 **
		CT	HP	0.0022 **
		GP	HT	0.0041 **
		GP	HP	0.0083 **
Bimanual		CT	GP	<0.0001 ***
		CP	GP	<0.0001 ***
		HT	GP	<0.0001 ***
		CT	HP	<0.0001 ***
		CP	HP	0.0001 ***
		HT	HP	0.0011 **
		CT	GP	<0.0001 ***
		CP	GP	<0.0001 ***
		HT	GP	0.0004 ***
		CT	HP	0.0005 ***
		HP	GP	0.0434 *
		CT	HT	0.0451 *
		CP	HP	0.0538 (*)

the techniques considered more difficult are the ones based on hand-tracking, respectively HP for musicians (M = 5.1, SD = 2.9) and HT for non-musicians (M = 5, SD = 3.1). For Bimanual condition, GP showed the highest level of difficulty for both participants with (M = 6.6, SD = 3.3) and without (M = 6.2, SD = 3.5) musical expertise. The pairwise comparison presented in Table 10 showed the results of the analysis. As for Unimanual, we found a difference between CT-HP for both musicians ($p \leq 0.0001$), and between CT-HT for non-musicians ($p \leq 0.0001$). For Bimanual, we found a difference between pairs CT-GP for both musicians and non-musicians ($p \leq 0.0001$). Different from synchronization, in the Bimanual condition, GP appears to be the technique that showed the highest variance among all different techniques and types of musical expertise.

As shown in Table 11 pairwise comparison showed a statistical difference between the Unimanual and Bimanual condition for GP in both participants with and without musical background ($p \leq 0.001$). When using one hand GP is considered to be less difficult compared to when using two hands. However, we did not find any significant difference between participants with and without musical expertise.

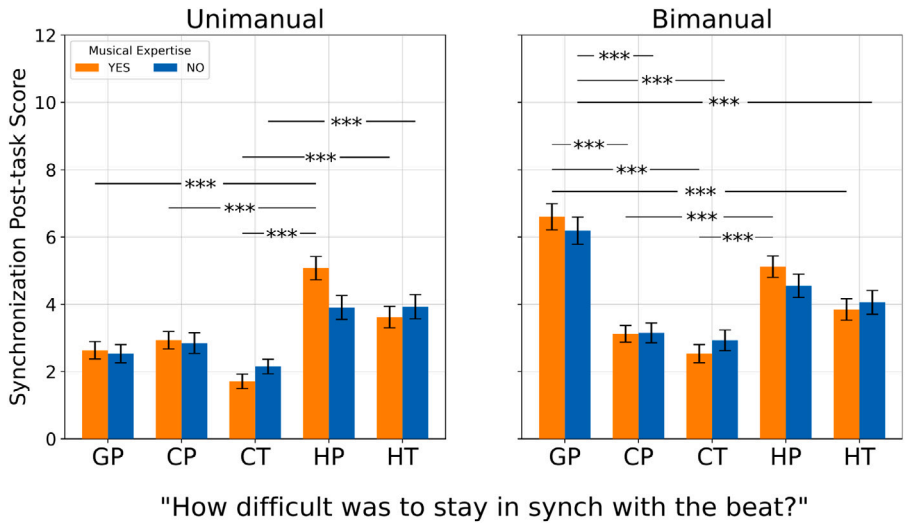

Fig. 8. Results of the post-task questionnaires: "How difficult was it to stay in synch with the beat?". Error bars represent the standard error. Only the most significant pairs ($p < 0.001$ ***) are shown. See Tables 10 and 11 for details.

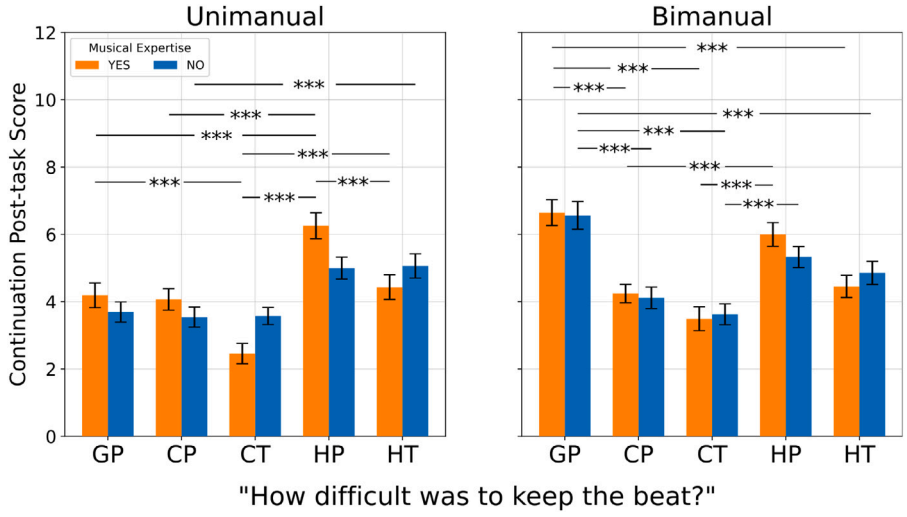

Fig. 9. Results of the post-task questionnaires: "How difficult was to keep the beat?". Error bars represent the standard error. Only the most significant pairs ($p < 0.001$ ***) are shown. See Tables 12 and 13 for details.

Table 13
The table presents the results of the pairwise comparisons for the Post-task questionnaire 2 regarding the musical background condition. The color orange refers to participants with a musical background, and blue to participants without a musical background. Dark green indicates the values with the lowest means of the group, while light green indicates the techniques with the highest mean.

Post-task questionnaire 2				
Condition musical expertise				
	Unimanual	Bimanual	p-value	
	GP	GP	<0.0001	***
	GP	GP	<0.0001	***

5.2.2. "How difficult was it to keep the beat?"

Fig. 9 shows the results of the second post-task questionnaire. For the Unimanual condition, musicians ranked CT as the technique less difficult for the task of keeping the beat (M = 2.4, SD = 2.6) and HP to be the most difficult (M = 6.2, SD = 3.3). Non-musicians ranked CP to be the easiest (M = 3.5, SD = 2.5) and HT the most difficult (M = 5, SD = 3.1). For the Bimanual condition, for both participants with and without musical expertise CT was ranked as the easiest (M = 3.5, SD = 3.1; M = 3.6, SD = 2.6), and GP the most difficult (M = 6.6, SD = 3.3; M = 6.5, SD = 3.5). Post hoc test confirmed these pairs to be significantly different ($p < 0.001$). Overall, we can notice that techniques based on hand-tracking were perceived as the most difficult for the Unimanual condition, while techniques based on the tracked controller were considered less difficult. For the Bimanual condition, HP and GP were considered the most difficult for keeping the beat, for both musicians and non-musicians. However, we found a significant difference for GP between hand conditions. We did not find any significant difference for musical expertise.

5.3. Total workload

Fig. 10 presents the scores obtained from the raw NASA TLX questionnaires for each technique. As shown in Table 14, we found a significant main effect for both Unimanual and Bimanual conditions as well for musical expertise ($p < 0.001$). We also found an interaction effect between the conditions technique and hand ($p < 0.001$). Table 15 shows the results of the pairwise comparison for the condition hand.

For the Unimanual condition, GP showed the lowest workload for musicians and non-musicians (M = 46.7, SD = 15.5; M = 50.6, SD = 19.4). The highest workload was found for HP for participants with

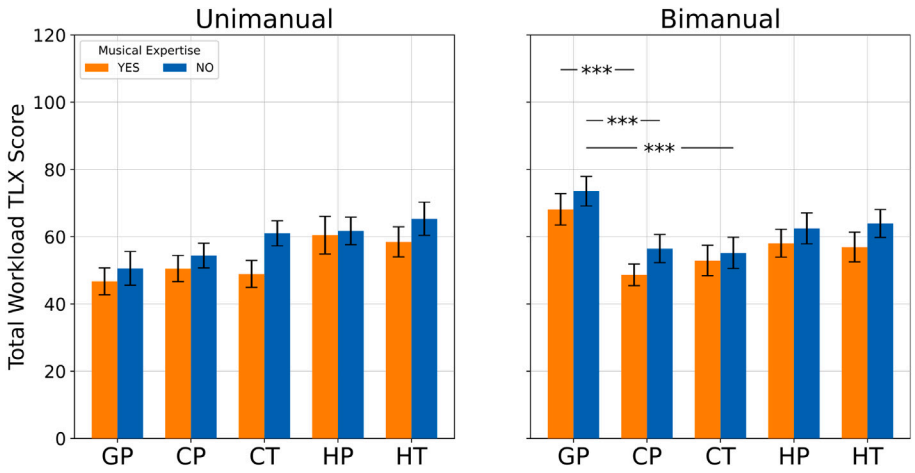

Fig. 10. The results of the total workload of the TLX questionnaire. Error bars represent the standard error. Only the most significant pairs ($p < 0.001$ ***) are shown. For details regarding the statistical analysis refer to Tables 15 and 17.

Table 14
Results of the statistical analysis for the TLX questionnaire, with main effect and interactions among factors.

TLX								
		Total workload	Mental demand	Physical demand	Temporal demand	Performance	Effort	Frustration
Condition	Factor	Main effect	Main effect	Main effect	Main effect	Main effect	Main effect	Main effect
Unimanual	Technique	F(4,112.762) = 9.8 ***	F(4,112.898) = 3.2 *	F(4,113.141) = 17.0 ***	F(4,112.509) = 8.4 ***	F(4,112.533) = 8.3 ***	F(4,113.008) = 6.0 ***	F(4,113.008) = 6.0 ***
	Musical expertise	–	–	–	–	–	–	–
	Technique – Musical expertise	–	–	–	–	–	F(4,113.008) = 3.5 **	–
Bimanual	Technique	F(4,113.055) = 15.8 ***	F(4,113.343) = 17.1 ***	F(4,113.955) = 13.8 ***	F(4,112.91) = 11.3 ***	F(4,113.337) = 14.3 ***	F(4,112.742) = 11.7 ***	F(4,112.996) = 10.9 ***
	Musical expertise	–	–	–	–	–	–	–
	Technique – Musical expertise	–	–	–	–	–	–	–
Musicians	Technique	F(4,126.93) = 5.8 ***	F(4,126.94) = 6.6 ***	F(4,127.30) = 6.2 ***	F(4,126.60) = 5.0 ***	F(4,126.47) = 12.7 ***	F(4,127.11) = 8.8 ***	F(4,126.74) = 6.9 ***
	Hand	F(1,126.95) = 5.3 *	F(1,126.96) = 6 *	F(4,127.31) = 6.6 *	–	F(1,126.49) = 6.5 *	–	F(1,126.76) = 4.9 *
	Technique – Hand	F(4,126.93) = 7 ***	F(4,126.94) = 2.9 *	F(4,127.30) = 6.3 ***	F(4,126.60) = 6.2 ***	–	F(4,127.11) = 3.9 **	F(4,126.74) = 2.6 *
Non-musicians	Technique	F(4,126) = 4.1 **	F(4,126) = 5.6 ***	F(4,126) = 10.4 ***	F(4,126) = 4.5 **	F(4,126) = 5.6 ***	–	F(4,126) = 6 ***
	Hand	F(1,126) = 5.3 *	F(1,126) = 8.7 **	F(1,126) = 7.6 **	–	F(1,126) = 4.5 *	–	–
	Technique – Hand	F(4,126) = 9.6 ***	F(4,126) = 3.6 **	F(4,126.11) = 11.2 ***	F(4,126) = 4 **	F(4,126) = 5.4 ***	F(4,126) = 7.8 ***	F(4,126) = 3.8 **

Table 15
The table presents the results of the pairwise comparisons for the total workload of the NASA TLX questionnaire regarding the hand condition. The color orange refers to participants with a musical background, and blue to participants without a musical background. Dark green indicates the values with the lowest means of the group, while light green indicates the techniques with the highest mean.

NASA TLX – Total workload					
Condition hand					
		Low mean	High mean	p-value	
Unimanual		GP	HP	0.0104	*
		CT	HP	0.0487	*
		GP	HT	0.056	*
		GP	HT	0.0027	**
Bimanual		CP	GP	0.0001	***
		CT	GP	0.0021	**
		CT	GP	0.0001	***
		CP	GP	0.0003	***

musical background (M = 60.4, SD = 3.3), and HT for participants without musical background (M = 5.1, SD = 3.1). Pairwise comparisons found a statistical difference between pairs GP-HP ($p = 0.0104$) and GP-HT ($p = 0.0027$). For the Bimanual condition, CP and CT were perceived as having the lowest workload for musicians and non-musicians respectively (M = 48.6, SD = 12.5; M = 55.2, SD = 17.9). GP exhibited the highest workload for participants with and without musical expertise (M = 68, SD = 4.6; M = 73, SD = 4.3). Pairwise comparisons revealed a significant difference between pairs CP-GP ($p \leq 0.0001$) and CT-GP ($p = 0.0001$). We can further notice that GP exhibited the lowest workload for Unimanual, but has the highest for Bimanual. As shown in Table 17, pairwise comparisons revealed a statistical difference for GP between the two hand conditions for both musicians and non-musicians ($p \leq 0.0001$). However, we did not find any significant difference between participants with and without musical expertise.

5.3.1. Workload subscales

Figs. 11 and 12 show the raw values for each individual subscale of the NASA TLX and the results of the pairwise comparisons between techniques, for the Unimanual and Bimanual conditions, respectively. Results of the pairwise comparison are presented in Table 16 for the hand condition, and in Table 17 for the musical expertise condition. We will briefly discuss each subscale. We report only the most significant pairs.

Mental Demand: No significant difference between techniques was found for Unimanual condition. For Bimanual, we found a significant statistical difference ($p \leq 0.001$) between CT and GP for both musicians and non-musicians being the techniques with lowest and highest mental demand respectively.

Physical Demand: For the Unimanual condition, the lowest physical demand was found for musicians to be GP and the highest HP ($p = 0.0021$). Even if GP was also ranked lowest for non-musicians, CT showed the highest physical demand ($p \leq 0.001$). For the Bimanual condition, we found similar results for both participants with and without musical expertise, with CP being the lowest and GP the highest in terms of score for musicians ($p = 0.0002$) and non-musicians ($p = 0.0001$). Pairwise comparisons revealed a difference for GP between Unimanual and Bimanual (musicians: $p = 0.0001$; non-musicians: $p = 0.0001$).

Temporal Demand: In the Unimanual condition, musicians ranked HP as the technique perceived with the highest temporal demand and CT with the lowest ($p = 0.0288$), whereas non-musicians reported a difference between HT and CP ($p = 0.0494$). In the Bimanual condition, both categories of subjects reported GP to have the highest temporal

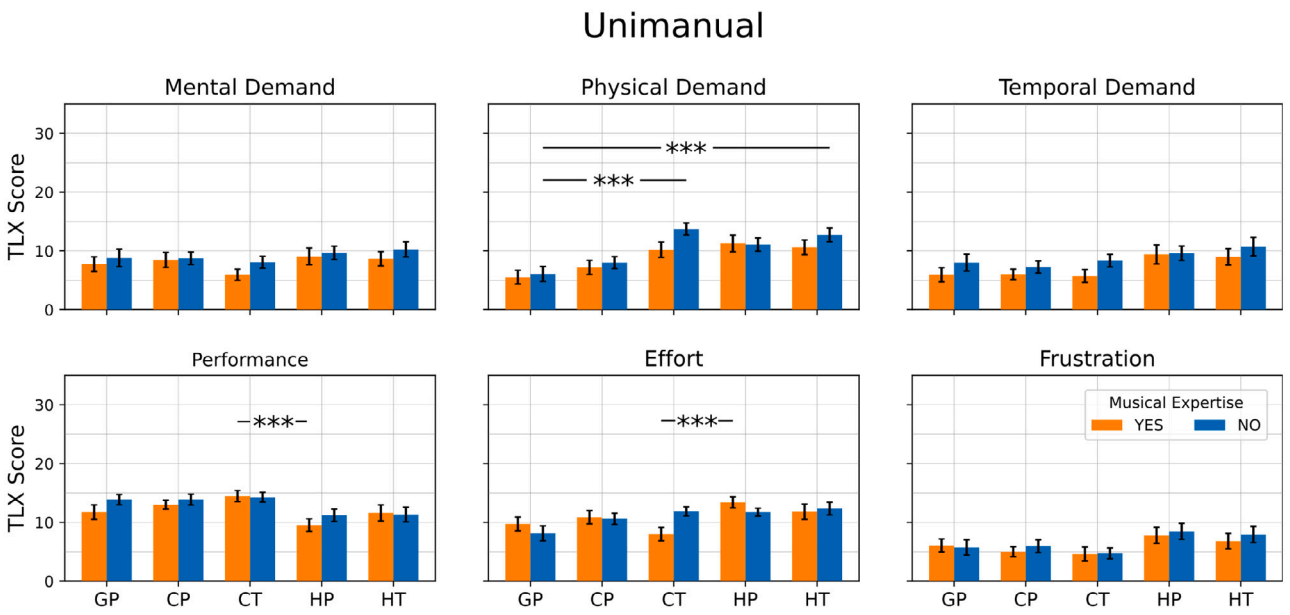

Fig. 11. The averaged responses for each subscale of the TLX questionnaire for the Unimanual condition. Error bars represent the standard error. Only the most significant pairs ($p < 0.001$ ***) are shown. For details regarding the statistical analysis refer to Tables 16 and 17.

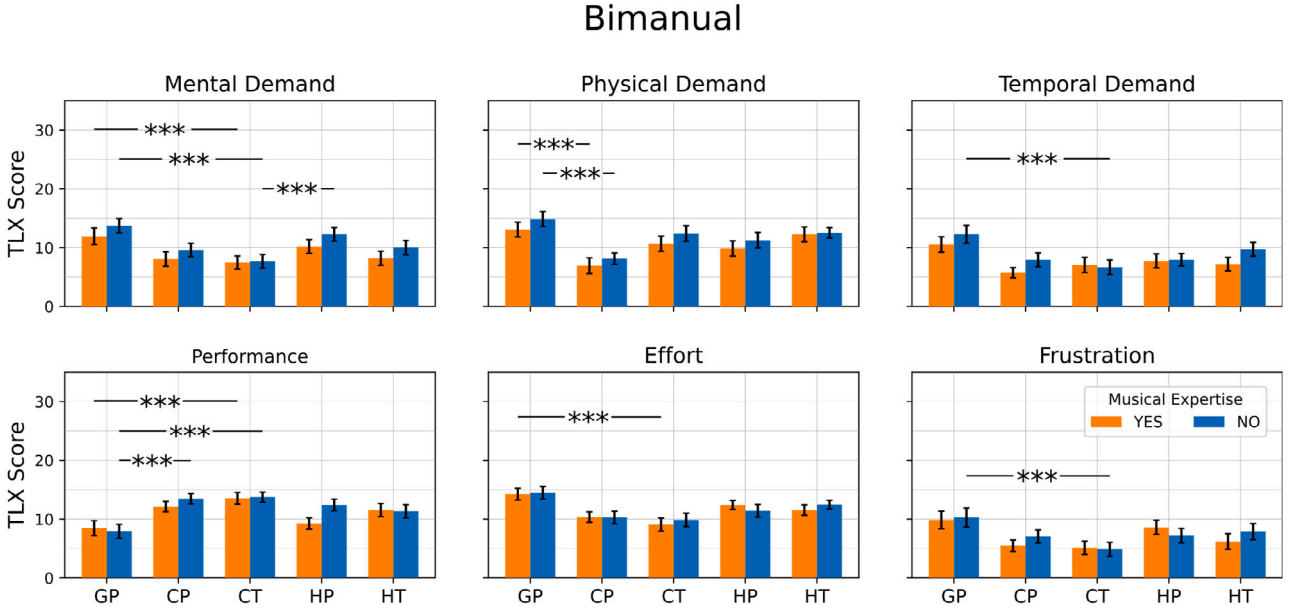

Fig. 12. The averaged responses for each subscale of the TLX questionnaire for the Bimanual condition. Error bars represent the standard error. Only the most significant pairs ($p < 0.001$ ***) are shown. For details regarding the statistical analysis refer to Tables 16 and 17.

demand, but the lowest was found for musicians in CP ($p = 0.0014$) and non-musicians in CT ($p = 0.0001$).

Performance: Differently from the previous subscales, in this question the technique ranked lowest is the one perceived to perform poorly compared to the one ranked highest. For musicians in Unimanual condition, CT showed to perform better compared to HP ($p = 0.0007$). No differences were found for non-musicians. In the Bimanual condition, CT is the technique ranked highest for both musicians and non-musicians, while GP is the technique ranked lowest. The pairs showed a significant difference for musicians ($p = 0.0005$) and for non-musicians ($p \leq 0.001$).

Effort: For Unimanual condition, CT and HP are the techniques ranked lowest and highest for musicians ($p = 0.0008$), while for non-musicians are GP and HT ($p = 0.0243$). For Bimanual condition, pairwise comparison showed a statistical difference between CT and GP for both musicians ($p = 0.0243$) and non-musicians ($p = 0.0022$).

Frustration: For Unimanual, CT was ranked as the technique with the lowest frustration, while HP was ranked the highest. Pairwise comparisons showed a statistical difference between such pairs for participants with ($p = 0.0483$) and without musical expertise ($p = 0.0091$). For Bimanual condition, GP was the technique that showed the highest level of frustration for musicians and non-musicians alike. The pairwise comparison revealed a statistical difference between this technique and the ones ranked lowest, such as CP for musicians ($p = 0.0014$) and CT for non-musicians ($p = 0.0001$).

Table 16
The table presents the results of the pairwise comparisons for the averaged responses to each subscale of the TLX questionnaire regarding the hand condition. The color orange refers to participants with a musical background, and blue to participants without a musical background. Dark green indicates the values with the lowest means of the group, while light green indicates the techniques with the highest mean.

NASA TLX

Condition hand

	Menta demand				Physical demand			
		Low mean	High mean	p-value		Low mean	High mean	p-value
Unimanual						GP	HP	0.0021 **
						GP	HT	0.0121 *
						GP	CT	0.0306 *
						GP	CT	<0.0001 ***
						GP	HT	0.0001 ***
						CP	CT	0.0014 **
						GP	HP	0.0097 *
						CP	HT	0.0186 *
Bimanual		CT	GP	0.0009 ***		CP	GP	0.0002 ***
		CP	GP	0.008 **		CP	HT	0.0026 **
		HT	GP	0.0123 *				
		CT	GP	<0.0001 ***		CP	GP	<0.0001 ***
		CT	HP	0.0006 ***		CP	HT	0.029 *
		CP	GP	0.0032 **		CP	CT	0.0393 *
		HT	GP	0.0123 *				
	Temporal demand				Performance			
		Low mean	High mean	p-value		Low mean	High mean	p-value
Unimanual		CT	HP	0.0288 *		HP	CT	0.0007 ***
		GP	HP	0.0347 *				
		CP	HT	0.0494 *				
Bimanual		CP	GP	0.0014 **		GP	CT	0.0005 ***
						HP	CT	0.0068 **
						GP	CP	0.0379 *
		CT	GP	0.0001 ***		GP	CT	<0.0001 ***
		CP	GP	0.0064 **		GP	CP	0.0001 ***
		HP	GP	0.0064 **		GP	HP	0.0036 **
	Effort				Frustration			
		Low mean	High mean	p-value		Low mean	High mean	p-value
Unimanual		CT	HP	0.0008 ***		CT	HP	0.0483 *
		GP	HT	0.0243 *		CT	HP	0.0091 **
						CT	HT	0.0483 *
Bimanual		CT	GP	0.0003 ***		CP	GP	0.0014 **
		CP	GP	0.0172 *				
		CT	GP	0.0022 **		CT	GP	0.0001 ***
						CP	GP	0.0064 **
						HP	GP	0.0064 **

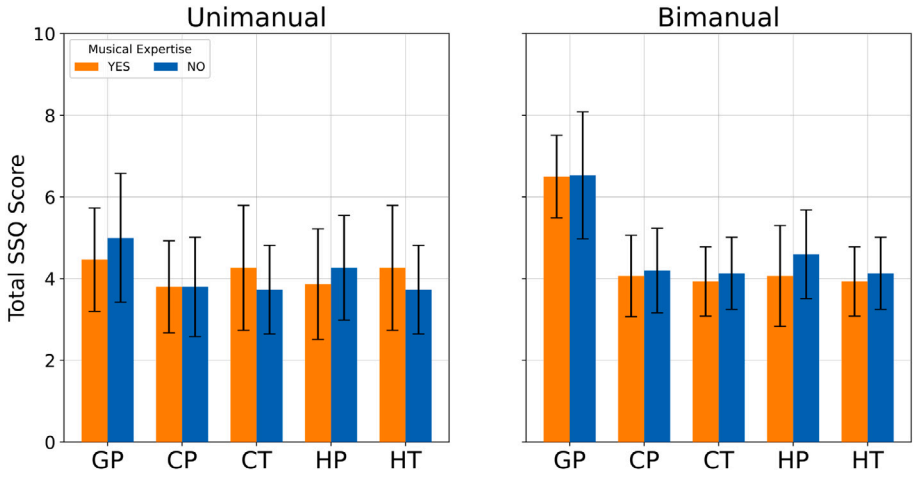

Fig. 13. The total score of the SSQ questionnaire. Error bars represent the standard error. Only the most significant pairs ($p < 0.001$ ***) are shown.

5.4. Comfort

Fig. 13 shows the raw results of the SSQ, while Table 18 presents the main effect and interactions for each subscale. Regarding total sickness, we found an effect for interaction technique, for the Bimanual condition ($p < 0.001$), and for non-musicians ($p < 0.05$). Overall, we found no difference between techniques and musical expertise, and between Unimanual and Bimanual conditions. Surprisingly found that

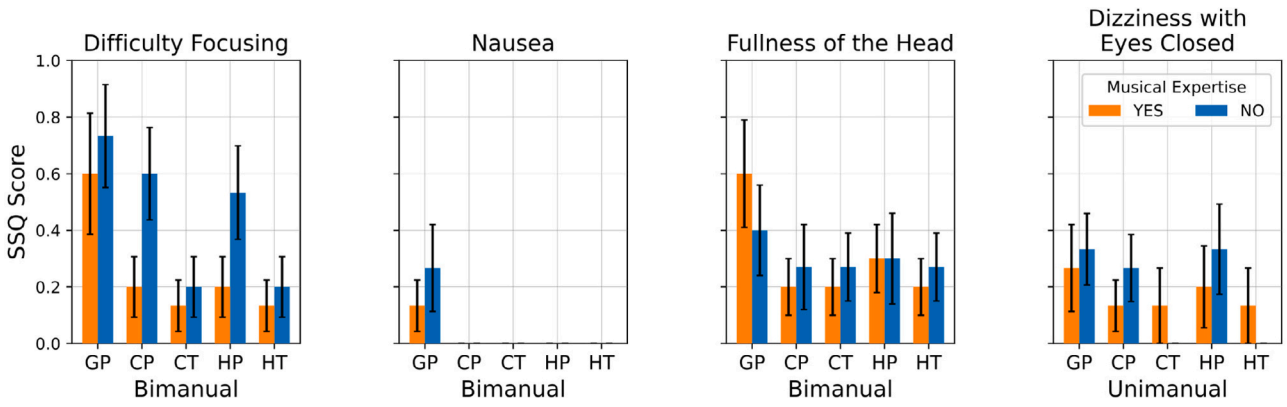

Fig. 14. Results of three questions of the SSQ that showed relevant results after the pairwise comparison. Error bars represent the standard error. Only the most significant pairs ($p < 0.001$ ***) are shown. For details regarding the statistical analysis refer to Table 19.

Table 17
The table presents the results of the pairwise comparisons for the total workload and the averaged responses to each subscale of the TLX questionnaire regarding the musical background condition. The color orange refers to participants with a musical background, and blue to participants without a musical background. Dark green indicates the values with the lowest means of the group, while light green indicates the techniques with the highest mean.

NASA TLX				
Condition musical expertise				
		Unimanual	Bimanual	*p*-value
Total workload		GP	GP	<0.0001 ***
		GP	GP	<0.0001 ***
Mental demand		GP	GP	0.012 *
		GP	GP	0.0018 **
Physical demand		GP	GP	<0.0001 ***
		GP	GP	<0.0001 ***
Temporal demand		GP	GP	0.0011 **
		GP	GP	0.0262 *
Performance		GP	GP	0.0164 *
		GP	GP	0.0001 ***
Effort		GP	GP	0.0043 **
		GP	GP	<0.0001 ***
Frustration		GP	GP	0.0113 *
		GP	GP	0.0036 **

the overall score for each symptom remained in the range between *None* and *Slight*, signaling a very low impact on sickness. This was particularly interesting considering the length of VR exposure and tasks. Interestingly, we found a statistical difference between techniques for symptoms such as "*Difficulty Focusing*", "*Dizziness with Eyes Closed*", "*Fullness of the Head*", and "*Nausea*". Nonetheless, only for participants without musical expertise (see Fig. 14, and Table 19 for the results of the pairwise comparison). The two symptoms related to vision impacted users who wore glasses in their daily lives. However, we do not consider these results to carry any important value for answering our research questions.

5.5. Qualitative results

The answers to the open-ended questions were analyzed using an inductive thematic analysis (Braun and Clarke, 2006). The analysis was conducted by the authors by generating codes, which were further organized into the following themes that reflected relevant patterns. Our analysis revealed two main categories of themes revolving around which techniques were facilitating or not keeping pace.

5.5.1. Techniques that facilitated the timing tasks

- **Large Movements** : Thirteen participants commented that the movement afforded by CT helped them to keep a constant pace (e.g., *"As long as my arm is in constant motion it makes it easy to keep the tempo"*). Participants typically described such a movement as *"large"*. CT was to be more intuitive and direct, because of the ample movements of the hand and the arm used for tapping the pads (e.g., *"Large movements require less precision"*). Such movements seem to be very consistent and suitable for the tasks of synchronizing with the beat as well as keeping a constant pace.
- **Musical Metaphor** : Fourteen participants drew a comparison between the experience with CT and the experience of playing a percussive instrument (e.g., *"Seems like if I was beating a drum"*, *"seems to be very close to a musical instrument"*). Moreover, participants explicitly referred to tools such as *"mallets"* and *"drumsticks"*. This suggests that tapping the pads with CT is perceived to be very musical. Because of the continuous and oscillatory motion, CT not only allows one to effectively perform a downbeat but also to perform an upbeat, which seems to be absent in other techniques (e.g., *"it allows to beat the off-beat"*, *"I can prepare the movement between beats...because of this I feel more in control"*). We should notice that these comments were distributed among musicians and non-musicians. The musical aspect was also reinforced by four participants with adjectives such as *"fun"* and *"engaging"*.
- **Trade-off between perceived precision and physical effort** : Even if CT was considered to be direct and precise, eighteen participants reported that this technique was also perceived as physically demanding (e.g., *"It was the best to keep the tempo, however, it was very tiring"*, *"it was very demanding, but it allowed me to be precise"*). These comments came from both musicians and non-musicians. However, such a trade-off between physical effort and precision was not intended by participants in terms of positive or negative value.
- **Small Movements** : Eighteen participants commented that the button present in the controllers (and used in techniques such as GP and CP) was an important element of reference for keeping the tempo. Therefore techniques such as GP or CP appear to be very convenient when the task requires the participant to synchronize with the beat (e.g., *"After you start to synchronize, you know that the button is there, therefore if, by any reason you miss a beat, it's not difficult to pick up the pace again"*). Specifically, eight participants referred to the action of pressing the button as *"small movements"*. Such kinds of movements were perceived to require less effort, both physical and mental (e.g., *"it was more practical to make use of a small movement -like pressing a button- instead of using a larger one"*).

These clusters of themes regarding large and small movements were not mutually exclusive. Eleven participants who commented on one commented also on the other. The small movements afforded by the

Table 18
Results of the statistical analysis for the SSQ questionnaire, with main effect and interactions among factors.

SSQ										
		Total sickness	Fatigue	Headache	Eye strain	Difficulty focusing	Nausea	Difficulty Concentrating	Fullness of the head	Dizziness with eyes closed
Condition	Factor	Main effect	Main effect	Main effect	Main effect	Main effect	Main effect	Main effect	Main effect	Main effect
Unimanual	Technique	–	F(4,113.140) = 2.7 *	–	–	–	–	F(4,112.647) = 2.6 *	–	F(4,111.76) = 4.3 **
	Musical expertise	–	–	–	–	–	–	–	–	–
	Technique – Musical expertise	–	–	–	–	–	–	–	–	–
Bimanual	Technique	F(4,112.988) = 5.0 ***	–	F(4,112.921) = 3.3 *	F(4,113.55) = 2.6 *	F(4,112.781) = 7.6 ***	F(4,111.092) = 5.0 ***	–	F(4,112.840) = 2.8 *	–
	Musical expertise	–	–	–	–	–	–	–	–	–
	Technique – Musical expertise	–	–	–	F(4,113.55) = 2.6 *	–	–	–	–	–
Musicians	Technique	–	–	–	F(4,126) = 2.2 *	–	–	F(4,126) = 3.2*	F(4,126) = 3.6**	–
	Hand	–	–	–	–	–	–	–	F(1,126) = 4.5*	–
	Technique – Hand	–	–	–	–	–	–	–	–	–
Non musicians	Technique	F(4,126) = 2.4 **	–	–	–	F(4,126) = 3.6 *	F(4,126) = 2.6 *	–	–	–
	Hand	–	–	–	–	–	–	–	–	–
	Technique – Hand	–	–	–	–	–	–	–	–	–

Table 19
The table presents the results of the pairwise comparisons for the results of three questions of the SSQ questionnaire regarding the hand condition. The color orange refers to participants with a musical background, and blue to participants without a musical background. Dark green indicates the values with the lowest means of the group, while light green indicates the techniques with the highest mean.

SSQ					
Condition hand					
		Low mean	High mean	*p*-value	
Unimanual		GP	CT	0.0184 *	Difficulty focusing
		GP	HT	0.0184 *	
Unimanual		GP	CT	0.0351 *	Nausea
		GP	CP	0.0351 *	
		GP	HT	0.0351 *	
		GP	HP	0.0351 *	
Bimanual		GP	CT	0.0172 *	Fullness of the head
		GP	CP	0.0172 *	
		GP	HT	0.0172 *	
Unimanual		GP	CT	0.0322 *	Dizziness with eyes closed
		GP	HT	0.0322 *	
		GP	CP	0.0322 *	
		GP	HP	0.0322 *	

button were felt more useful for synchronization tasks. While large movements of CT were perceived as more intuitive for keeping pace with and without the auditory stimuli. This might suggest that the fewer participants have to think about the execution of a technique, the more such technique will help them to concentrate on the task.

5.5.2. Techniques that did not facilitate the timing tasks

- **"Point and commit" is perceived as an additional task** : Twenty-four participants commented that techniques using the *"point and commit"* paradigm (GP, CP, and HP) were problematic for timing tasks. Participants perceived the action of pointing a target as an additional task that was distracting them from the main one and contributing to a negative performance (e.g., *"It was complex because pointing became 'something more' to think about while trying to keep the pace"*, *"If you don't point correctly, you will miss the beat"*). What contributes to this negative experience was also the insecurity generated by the confirmation mechanism, being it a button or a hand gesture which can also cause errors (i.e., missing a beat), thus decreasing their sense of control and trust (e.g., *"when I press or touch the fingers the ray move out the pad...I felt insecure"*).
- **Head Movements** : When performed with the head – as with GP – pointing results to be problematic. For the Unimanual condition sixteen participants commented that since they were required to look to a fixed point for the entire duration of the task, this created a sense of fatigue and eye strain. Moreover, participants reported that they wanted to move (even slightly) the head to help them keep pace, but since this caused errors they felt constrained (e.g., *"I had to keep my head still and at the same time I had to synchronize with the beat, this is physically demanding. You have to get stiff, and this, in my opinion, affected my performance negatively"*). For the Bimanual condition, fourteen participants commented that moving the head was causing them confusion and irritation. They reported that it was difficult to exactly point the targets, and the movement became difficult to sustain (e.g., *"Extremely uncomfortable because I had to constantly move the head, the result was that I was losing concentration"*). In addition, the use of the buttons with the movements of the head created a problem of coordination between two actions (e.g., *"It was difficult, I had to coordinate the movement of the head with the action of pressing the button"*). The negative experience in the Bimanual condition was reinforced by eight participants through the use of negative adjectives such as *"awful"*, *"worst"*, and *"bothersome"*.
- **Incorrect hand gesture recognition hinders performance** : The problem related to the selection phase of the gesture negatively impacted HP. Sixteen participants reported that the system for gesture recognition was not perceived as constant and reliable. This caused them to lose trust in the technique a difficulty in keeping a constant pace (e.g., *"I became doubtful, I was never sure if the gesture was correctly recognized or not"*, *"I had the feeling that the system was not registering the input correctly"*).
- **Hand tracking increases insecurity in execution** : Sixteen participants commented that they perceived some sort of inconsistency with the hand tracking techniques such as HP (seven participants) and HT (six participants), for the Unimanual condition. According to participants the movements of the virtual hands were perceived as not fluid, and this impacted the performance negatively. Since the movement of the virtual hand was not matching the one perceived by the user, this created a contrast with the timing task (e.g., *"I had in mind the beat I had to follow, and I was trying to understand the delay to adapt but I was not able to do it, was not constant, this creates confusion"*, *"I had the feeling there was some sort of a lag, I found the movement of the virtual hand not realistic in comparison with the movement of the real hand"*). These problems indicate that if participants have to think not only to keep a constant pace but also about the mechanic of the technique or to compensate for some delay this distracts them from the main task.

6. Discussion

In this section, we discuss the results analyzed in our experiments.

6.1. Musical background

Compared to non-VR tapping experiments (Franěk et al., 1991; Scheurich et al., 2018) we found no statistically significant differences between participants with and without a musical background. Nonetheless, we should notice that non-musicians showed larger asynchronies than musicians, even if not statistically relevant. This finding may be due to the novelty of the task participants were exposed to. We analyzed our results concerning the participants' previous level of experience and exposure to VR. However, we found no impact between experience in VR and the performance in the "*synchronization-continuation*" task. These additional results might suggest that interaction techniques for musical tasks must be learned and mastered, as this happens with musical instruments in general. While at the moment, we have to reject H1, it will be important to better investigate the impact of VR input techniques on musical skills, as well as compare selection techniques with standard finger tapping.

6.2. Bimanual tapping

Regarding asynchronies and variances, we found no improvement in performance between Unimanual and Bimanual conditions. However, for techniques such as CT and HT we can notice a slight improvement in the Bimanual condition with respect to their results in the Unimanual condition. Interestingly, these two techniques use collision as a means of selection. Our results suggest that interaction techniques using the method "*point and commit*" might pose more difficulties when used with two alternated hands. For instance, our results show a relevant drop in performance and user experience for GP (Bimanual condition). Moreover, participants considered this technique the worst in terms of user experience. For these reasons, we cannot accept H2.

6.3. What differs?

What our results highlighted is not a strong difference between single techniques but rather between groups of techniques, such as: techniques that make use of tracked controllers and techniques that make of hand-tracking (i.e., CP–CT, and HP–HT); techniques that allow distant interactions with rays and techniques that allow for direct interaction with collision (i.e., GP–CP - HP, and CT–HT).

6.3.1. Tracked controllers versus tracked hands

For both "*synchronization-continuation*" tasks and the questionnaires pairwise comparisons showed a noticeable difference between free-hand (i.e., HP, HT) and controller-based techniques (i.e., GP, CP, CT). Regarding free-hand techniques, one reason behind the lowest performance results observed for HP and HT might be found in the hand-tracking system used in the Quest 2. A study by Abdlkarim et al. (2022) showed that during a reaching task, the hand tracking of the Quest 2 exhibits an average delay of 38.0 ms. Especially with faster hand movements (which were measured with a reference of 160 BPM, IOI = 375 ms), they observed larger offsets between the position of the physical and virtual hands. According to the authors, the reason might reside in the Quest pre-trained machine-learning model for hand tracking. They hypothesize that the Quest 2 receives fewer frames when the hand is in rapid motion. In addition, they reported that hand tracking seems quite sensitive to several factors such as the position of the hands in respect of the headset, self-occlusion, and ambient light. We tried to avoid such issues as much as possible by performing the experiment in the same environment and with constant light conditions. The thematic analysis results revealed that the participants also reported such problems since they noticed the presence of some jitter and delay for HP and HT. Seemingly, such problems had an impact on the user experience, as highlighted in the results of the post-task questionnaires (see Tables 10–12). The results regarding HP and HT show how stability and consistency in hand tracking can hinder not only the performance of timing tasks but also how participants perceive their difficulty. The lack of standardized methods and comparable to measure and compare embedded tracking systems in headsets makes judging difficult. However, it will be important to understand if different hardware and systems differ and especially if new versions improved the issues highlighted in previous works or not. Interestingly, several participants expressed a strong sense of disappointment when using HT. They assumed it would be the most effective and easy because of its apparent realism. Our findings also indicate that these techniques had the highest miss ratio percentage, possibly due to tracking method inconsistencies.

6.3.2. Pointing versus collision

Another significant finding relates to the distinction between methods involving pointing (i.e., GP, CP, HP) and those involving collision (i.e., CT, HT). For instance, when considering musicians' user preferences for keeping the beat, we observed a difference between the HT-HP and CT-CP pairs for both Unimanual and Bimanual conditions (see Table 10). Our findings indicate that interaction methods utilizing the "*point and commit*" mechanism may present timing challenges. This applies to techniques such as CP, HP, and GP. Our thematic analysis revealed that the act of pointing to and selecting a target is interpreted by participants as an additional task that needs to be performed in addition to the main one (i.e., synchronize to a pacing tone). The actions of pointing and committing are carried out in sequence. In contrast to non-musical tasks, the timing between these actions becomes critical in synchronization tasks. This timing discrepancy between the pointing and selecting actions can lead to significant asynchronies and variations.

Another negative effect could have been caused by the "Heisenberg Effect" which is a common drawback observed in many selection tasks utilizing a "point and commit" mechanism. Such an effect becomes particularly noticeable when both actions of pointing and committing are performed on the same surface used for pointing, as seen in CP and HP, in contrast with GP where point is performed with the head and commit with the buttons on the controller. In fact, some participants commented that they preferred pointing with CP instead of HP because the ray could be controlled by a tool that was felt not anchored to their body. Moreover, several participants reported fatigue because they needed to constantly keep their hands in front of the headset.

Conversely, with GP they could leave the controllers in a more comfortable position since such a technique does not enforce mid-air interaction. However, we propose that for GP the source of the problems found in the Bimanual condition was not the pointing mechanism per se, but in the body movements required to select the pads.

6.3.3. Conflicting movements hinders bimanual tapping

Another significant finding is the difference between Unimanual and Bimanual conditions for GP. While in the Unimanual condition, we can observe similar levels of asynchronies and variances compared to CP and CT, along with the lowest workload and physical demand, in the Bimanual condition, GP appeared to be a very problematic technique.

According to participants' feedback, in the Unimanual condition, GP was perceived to be less demanding because they could keep their heads still since there was only one pad to point at. However, in the Bimanual condition, participants had to utilize their heads to move between the target pads. Such a movement was judged as extremely negative and was associated with high physical demand, effort, and sickness scores. Therefore, we propose to interpret the differences between GP in the Unimanual and Bimanual conditions as the consequences of a conflict between the *pointing* action (i.e., head movement) and the *committing* action (i.e., button press) used during the task.

In the Unimanual condition, the only action that requires timing is pressing the button on the controllers. Conversely, in the Bimanual condition, timing becomes a factor not only when the button is pressed, but also when the action of pointing has to be performed. This scenario may have given rise to a potential conflict between the continuous oscillatory movement of the head and the discrete finger movements.

This is evident in two key ways: first, through the significant asynchronies observed during the synchronization phase, which were accompanied by a notable level of participant frustration. Second, in the context of the GP, head movements assumed an active role in the interaction, departing from their conventional role in time perception. Previous research has highlighted the role of head movements in meter and beat perception (Phillips-Silver and Trainor, 2005, 2008), often used by musicians for structural communication and expressiveness (Nusseck and Wanderley, 2009) rather than sound production. This departure from the non-VR use of head movements could have introduced potential confusion and uncertainty regarding which action participants needed to synchronize: the pointing or the committing action.

6.4. Passive and self-haptics do not improve performance

Research on tapping conducted outside of VR has showed that the presence of tactile feedback has a beneficial effect on synchronization tasks with a paced signal (Repp, 2005; Repp and Su, 2005). Therefore, we hypothesized that techniques such as GP, CP, and HP would have led to a higher timing accuracy since they provide a form of passive haptics (the buttons on the controllers) and self-haptics (as provided by the pinch gesture). While this appears to be valid for CP, we cannot find evidence for GP and HP.

The latter is one of the techniques that showed the largest asynchronies and variances (see Figs. 5 and 6–7). Even if some participants appreciated the pinching gesture, several factors could have influenced the results. HP was rated as one of the most difficult techniques to use, especially in the Unimanual condition, as it can be seen in high levels of physical and temporal demand, as well as effort and frustration

(see Figs. 11–12). While, this might have been caused by issues in the hand-tracking system used for gesture recognition of the Quest 2, as described above, we have to consider other possibilities.

A study on tapping showed that when two fingers are touching together to keep pace, the timing of the taps becomes more variable (Keller et al., 2011). According to the authors, this might be linked to how the brain processes tactile information when both the tapping finger and the finger being tapped are active and send signals simultaneously, creating neural overlap and potentially making the task less clear or efficient. Therefore, they suggest that less sensitive body parts must be chosen (i.e., hand wrist).

Regarding GP and CP, they exhibited two of the lowest variance in the synchronization phase for the Unimanual condition (see Fig. 5). As mentioned by several participants, pressing the button on the controller appears to be effective for repeated actions. This was also highlighted in previous research on non-musical VR (Argelaguet and Andujar, 2013). However, if the results of CP were consistent between the two hands conditions, GP showed the largest variance in Bimanual. Because of these reasons, we have to reject H3.

One more piece of evidence can be found in the results regarding CT, a technique that does not provide any form of tactile feedback. Not only CT showed the shortest asynchronies and lowest variances, but participants also considered it as the technique most easy and comfortable to use. Several factors could have contributed to these results.

First, the stability of the tracking could have provided participants with a more consistent experience, which can be seen in the very low percentage of miss ratio.

Second, participants could have used the virtual pads as a visual reference to help them gauge distance and plan their movements effectively, as they were required to make contact between the tracked controller and these pads. Furthermore, when participants touch the pads, they receive a combination of visual and auditory feedback. This might have resulted in a simplified form of pseudo-haptic feedback (Lécuyer, 2009; Turchet et al., 2013), potentially aiding participants in timing tasks. This is evident from the absence of significant differences between CP and HT in terms of variance, observed in both the synchronization and continuation phases.

Third, another reason could be found in the movement afforded by CT. While keeping the beat, participants produced an oscillatory movement composed of a cycle of extension and flexion of both the hand and the arm. This movement was considered by participants to be "musical" since it was associated with the movement performed while hitting a drum. Furthermore, participants reported that they could synchronize not only the "on-beat" but also the "off-beat". Such an oscillatory movement could have played a role in helping participants establish better internal timing, through a cycle of extension and flexion of both hand and arm. Previous research showed the importance of kinaesthetic reafferences to the timing of movement even in contact-free tapping (Aschersleben et al., 2001). However, the same consideration cannot be applied to HT since the difference with CT can be explained by the issues of hand-tracking discussed above.

6.5. Workload

In the realm of 3D UI research within VR applications, workload assessment is a crucial metric, especially in scenarios involving mid-air interactions. Our hypothesis was that these techniques might impose a significant workload, particularly in terms of cognitive and physical demands, which could potentially lead to decreased performance.

In the evaluation 3D UIs and interactions in VR, workload assessment is a crucial, especially in scenarios involving mid-air interactions. Our results of the NASA TLX questionnaire's subscales reveal that CP and CT showed high physical demand and yielded the highest performance values. In canonical HCI and VR research, it is generally believed that a lower workload is desirable and crucial for the success of mid-air interactions (Jerald, 2015; LaViola et al., 2017). However, in the context of musical interactions, this topic has been extensively debated. Some argue for intuitive and easy instrument designs based on HCI principles (Mulder, 2000), while others emphasize the importance of a more ambiguous approach and the creation of physically demanding systems to provide musicians with expressive controls (Ryan, 1991). Our findings do not resolve this debate. However, based on the results of CP and CT, we contend that even in VR, this dichotomy persists. Rather than providing a definitive answer, it offers an avenue for experimentation and different design choices. Therefore, we cannot accept H4.

7. Design guidelines

Based on the reported results, we have developed a set of guidelines to assist in selecting the most effective technique for maintaining the beat in VR. These guidelines are intended for developers and researchers aiming to create user-friendly experiences with optimal performance in various scenarios. Based on our findings, if the objective is to design a VR musical system that caters to a wide range of users, we recommend utilizing techniques such as CT. This technique not only showed low asynchronies and variances but also afforded a movement considered "musical". It also offers stable and consistent tracking, along with direct contact with a virtual surface. However, if the main focus is on reducing workload while still ensuring precision, techniques like CP may be more appropriate. This type of technique could also be employed when the user needs to interact with a virtual instrument from a distance that cannot be physically reached. Nevertheless, it is important to consider the issues related to pointing discussed earlier. Both CT and CP are already commonly utilized in current VR musical applications. Furthermore, when the user needs to generate the beat themselves, techniques like CT should be prioritized. On the other hand, if the main objective is only to synchronize with a beat, techniques such as CP can be employed to enhance the experience.

There are other potential applications for VR systems that focus specifically on precision and synchronization with a metronome or music track, aiming to minimize physical and mental demand. These applications are relevant in the context of music education or for researchers studying sensorimotor synchronization through tapping. In such cases, GP (as in the Unimanual condition) can be considered. However, it is important to ensure that the target is fixed or changes in a way that does not require users to move their heads. Additionally, it may be possible to use two controllers in this configuration. With GP, where controllers are not tracked, buttons could be replaced with force-sensors or foot controllers. This design approach may be particularly beneficial for users with limited mobility or those utilizing low-end hardware like smartphone-based HMD.

When it comes to using hand-tracking, we advise for caution. Hand-tracking relies both on the mounted cameras on the HMD and the internal computer vision algorithms used for tracking and gesture recognition. In the previous section we have discussed this aspect and its drawbacks. Therefore, designers should take the following precautions: (1) ensure that the application is used in a stable environment with adequate lighting, (2) provide clear instructions to users on how to position their hands correctly, (3) design interactive surfaces that are easily accessible, avoiding movements that might obstruct the user's view, (4) avoid using fast tempi that could lead to sudden or jarring movements. Moreover, according to our results, hand-tracking techniques such as HT showed a high level of effort and frustration in the Unimanual condition compared to the Bimanual. Therefore, we can suggest the use of HT with two alternating hands. By following these guidelines, designers can optimize the experience and reliability of hand-tracking interactions.

8. Limitations and future work

Concerning our study, there are several limitations that should be reported. In the "*synchronization-continuation*" task we utilized a metronome as a pacing signal. However, it was acknowledged that tapping on a metronome may not be equivalent to tapping along with a musical piece (Repp, 2006a). Given the exploratory nature of our study, our objective was to gather comprehensive information about the usage of each technique. The metronome was chosen to avoid potential biases associated with using specific melodies or scales. Moreover, to simplify the task and ensure clarity for participants with varying levels of musical expertise, we employed a basic metric structure. However, it will be important to explore whether similar results can be obtained with more complex rhythmic structures, incorporating different accents and tempi.

While we conducted our experiment using five different inter-onset intervals (IOIs), we analyzed the results collectively as normalized values. Our research aimed to provide a comprehensive understanding of the impact of five VR interaction techniques on timing tasks. Therefore, we employed various IOIs to encompass a broader range of tempi, as music is not limited to specific sequences of beats per minute. Future studies should investigate the detailed analysis of performance differences and user experiences associated with each technique for each IOI.

Our study focused on evaluating the precision that participants exhibited with each technique. However, future studies should complement our findings by analyzing the five techniques in a free-form musical task. This might provide some additional clues in the process of better understanding the relationship between expression and precision.

Certain interpretations we made based on our results warrant further exploration, particularly regarding phenomena that are unique to VR. This includes examining the role of visuo-auditory pseudo-haptic effects as an alternative to direct contact and investigating the coordination between the pointing and selection mechanisms.

Another limitation regards the apparatus used in the experiment, such as the Meta Quest 2, which is one of the most widespread VR headsets that allow for "inside-out" tracking. Our results showed that two of the most problematic techniques are the ones based on hand-tracking. Currently, there are no shared metrics for evaluating and comparing hand-tracking hardware and tracking algorithms. Moreover, the core machine-learning model used by Meta is not public and accessible, and it is subject to improvements in future releases of the Interaction SDK. Therefore, future work should focus on more consistent evaluation of hand-tracking hardware and software and more precise bench-marking techniques. While recent studies explored this direction for music applications (Reimer et al., 2023), it will be extremely important to evaluate hand-tracking with timing tasks, such as the one we have employed. Moreover, because of the "Audio-first" characteristics of musical VR systems, it will be fundamental to explore approaches for measuring action-to-sound latency, which have been explored in the area of music technology (Deber et al., 2015; McPherson et al., 2020)

Even if this study has considered five input techniques, several other input techniques have been explored for tasks such as text entry, that have not been also explored systematically in the musical VR. Such techniques include the combination of gaze and hand pinch (i.e., GP + HP) (Pfeuffer et al., 2017), and the use of physical surfaces and body parts capable of providing passive haptics and redirected touch (Dube et al., 2022; Gil and Oakley, 2023). While most of these techniques have been studied in a Unimanual condition, they should be explored also using two hands. In our study, we did not include any hands-free technique such as the use of gaze and dwell time, or eyes blinking (Lu et al., 2020). Those techniques have never been explored systematically in musical VR, especially for timed activities such as synchronizing with a pacing tone.

By conducting additional research in these areas, we can gain deeper insights into the specific nuances and potential benefits of different interaction techniques in virtual reality. Finally, further limitations of this study are represented by the relatively low number of participants involved, and the gender imbalance. A highest number of participants and a more balanced gender distribution would confer the results with a higher level of generalizability.

9. Conclusion

In this paper, we conducted a comparative analysis of five primary selection techniques for musical VR. Our experiment aimed to gain insights into the performance and user experience associated with each technique when performing timing tasks. The collective findings of our study indicate that different 3D selection techniques have an impact on both task performance and user experience in the context of timing tasks. Our analysis revealed that factors such as tracking stability, selection metaphor, and the use of one or two hands play crucial roles in selecting an appropriate input technique for music in VR. Building upon our research, there are numerous possibilities and challenges that can be explored further regarding the influence of different input techniques. For instance, including haptic feedback and investigating the congruence and incongruence of audio-visual stimuli could provide valuable insights. We hope that our study will stimulate further research in the largely unexplored domain of utilizing VR technology for music creation. By delving deeper into this topic, we can uncover new avenues for innovation and enhance the potential of VR as a medium for musical expression.

CRediT authorship contribution statement

Alberto Boem: Conceptualization, Methodology, Formal analysis, Investigation, Software, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Luca Turchet:** Conceptualization, Data curation, Methodology, Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: ALBERTO BOEM reports financial support was provided by Ministry of Education and Merit.

Data availability

Data will be made available on request.

Acknowledgments

We acknowledge the support of the MUR PNRR PRIN 2022 grant, prot. n. 2022CZWWKP, funded by Next Generation EU. We would like to thank Matteo Tomasetti for the helpful feedback while testing the experiment, and Chiharu Seki for the support on the illustrations.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ijhcs.2024.103231.

References

Abdlkarim, D., Di Luca, M., Aves, P., Yeo, S.-H., Miall, R.C., Holland, P., Galea, J.M., 2022. A methodological framework to assess the accuracy of virtual reality hand-tracking systems: A case study with the oculus quest 2. BioRxiv 2022-2002.

Argelaguet, F., Andujar, C., 2013. A survey of 3D object selection techniques for virtual environments. Comput. Graph. 37 (3), 121–136.

Aschersleben, G., Gehrke, J., Prinz, W., 2001. Tapping with peripheral nerve block: A role for tactile feedback in the timing of movements. Exp. Brain Res. 136, 331–339.

Bååth, R., Strandberg, T., Balkenius, C., 2011. Eye tapping: How to beat out an accurate rhythm using eye movements. In: NIME. pp. 441–444.

Bachynskyi, M., Palmas, G., Oulasvirta, A., Weinkauf, T., 2015. Informing the design of novel input methods with muscle coactivation clustering. ACM Trans. Comput.-Hum. Interact. 21 (6), 1–25.

Balasubramaniam, R., Wing, A.M., Daffertshofer, A., 2004. Keeping with the beat: movement trajectories contribute to movement timing. Exp. Brain Res. 159, 129–134.

Bardos, L., Korinek, S., Lee, E., Borchers, J., 2005. Bangarama: Creating music with headbanging. In: Proceedings of the 2005 Conference on New Interfaces for Musical Expression. pp. 180–183.

Batmaz, A.U., Sun, X., Taskiran, D., Stuerzlinger, W., 2019. Hitting the wall: Mid-air interaction for eye-hand coordination. In: Proceedings of the 25th ACM Symposium on Virtual Reality Software and Technology. pp. 1–5.

Bavassi, M., Tagliazucchi, E., Laje, R., 2013. Small perturbations in a finger-tapping task reveal inherent nonlinearities of the underlying error correction mechanism. Hum. Mov. Sci. 32 (1), 21–47.

Bella, S.D., Farrugia, N., Benoit, C.-E., Begel, V., Verga, L., Harding, E., Kotz, S.A., 2017. BAASTA: Battery for the assessment of auditory sensorimotor and timing abilities. Behav. Res. Methods 49, 1128–1145.

Berthaut, F., 2020. 3D interaction techniques for musical expression. J. New Music Res. 49 (1), 60–72.

Berthaut, F., Desainte-Catherine, M., Hachet, M., 2011. Interacting with 3D reactive widgets for musical performance. J. New Music Res. 40 (3), 253–263.

Bilbow, S., 2022. Evaluating polaris - an audiovisual augmented reality experience built on open-source hardware and software. In: NIME 2022.

Bowman, D.A., Hodges, L.F., 1997. An evaluation of techniques for grabbing and manipulating remote objects in immersive virtual environments. In: Proceedings of the 1997 Symposium on Interactive 3D Graphics. p. 35.

Bowman, D., Wingrave, C., Campbell, J., Ly, V., 2001. Using pinch gloves (tm) for both natural and abstract interaction techniques in virtual environments.

Braun, V., Clarke, V., 2006. Using thematic analysis in psychology. Qual. Res. Psychol. 3 (2), 77–101.

Bruck, S., Watters, P.A., 2009. Estimating cybersickness of simulated motion using the simulator sickness questionnaire (SSQ): A controlled study. In: 2009 Sixth International Conference on Computer Graphics, Imaging and Visualization. pp. 486–488.

Bugos, J.A., 2019. The effects of bimanual coordination in music interventions on executive functions in aging adults. Front. Integr. Neurosci. 13, 68.

Çamcı, A., Hamilton, R., 2020. Audio-first VR: New perspectives on musical experiences in virtual environments. J. New Music Res. 49 (1), 1–7.

Cabral, M., Montes, A., Roque, G., Belloc, O., Nagamura, M., Faria, R.R., Teubl, F., Kurashima, C., Lopes, R., Zuffo, M., 2015. Crosscale: A 3D virtual musical instrument interface. In: 2015 IEEE Symposium on 3D User Interfaces. 3DUI, IEEE, pp. 199–200.

Cadoz, C., Wanderley, M., 2000. Gesture-music. In: Wanderley, M., Battier, M. (Eds.), Trends in Gestural Control of Music. Ircam - Centre Pompidou, pp. 71–93.

Caggianese, G., Gallo, L., Neroni, P., 2019. The vive controllers vs. leap motion for interactions in virtual environments: a comparative evaluation. In: Intelligent Interactive Multimedia Systems and Services: Proceedings of 2018 Conference 11. Springer, pp. 24–33.

Çamcı, A., Vilaplana, M., Wang, L., 2020. Exploring the affordances of VR for musical interaction design with VIMEs. In: Proc. Int. Conf. New Interfaces Musical Expression. pp. 1–6.

Choe, M., Choi, Y., Park, J., Kim, H.K., 2019. Comparison of gaze cursor input methods for virtual reality devices. Int. J. Hum.–Comput. Interact. 35 (7), 620–629.

Costa, W., Ananias, L., Barbosa, I., Barbosa, B., De'Carli, A., Barioni, R.R., Figueiredo, L., Teichrieb, V., Filgueira, D., 2019. Songverse: a music-loop authoring tool based on virtual reality. In: 2019 21st Symposium on Virtual and Augmented Reality. SVR, IEEE, pp. 216–222.

Davanzo, N., Avanzini, F., 2020. Experimental evaluation of three interaction channels for accessible digital musical instruments. In: Computers Helping People with Special Needs: 17th International Conference, ICCHP 2020, Lecco, Italy, September 9–11, 2020, Proceedings, Part II 17. Springer, pp. 437–445.

Davanzo, N., De Filippis, M., Avanzini, F., et al., 2021. Netychords: An accessible digital musical instrument for playing chords using gaze and head movements. In: CHIRA. pp. 202–209.

Deber, J., Jota, R., Forlines, C., Wigdor, D., 2015. How much faster is fast enough? User perception of latency & latency improvements in direct and indirect touch. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. CHI '15, Association for Computing Machinery, New York, NY, USA, pp. 1827–1836.

Distance Hand Grab Interaction, Distance Hand Grab Interaction, https://developer.oculus.com/documentation/unity/unity-isdk-distance-hand-grab-interaction/. Accessed: 2023-04-05.

Drum Beats VR, Drum Beats VR, https://store.steampowered.com/app/1015480/DrumBeats_VR/. Accessed: 2023-04-05.

Dube, T.J., Johnson, K., Arif, A.S., 2022. Shapeshifter: Gesture typing in virtual reality with a force-based digital thimble. In: CHI Conference on Human Factors in Computing Systems Extended Abstracts. pp. 1–9.

Dudley, J., Benko, H., Wigdor, D., Kristensson, P.O., 2019. Performance envelopes of virtual keyboard text input strategies in virtual reality. In: 2019 IEEE International Symposium on Mixed and Augmented Reality. ISMAR, IEEE, pp. 289–300.

Electronauts VR Music, Electronauts VR Music, https://survios.com/electronauts/. Accessed: 2023-04-05.

EXA Infinite Instrument, EXA Infinite Instrument, https://store.steampowered.com/app/606920/EXA_The_Infinite_Instrument/. Accessed: 2023-04-05.

Faust, Faust Home, https://faust.grame.fr/. Accessed: 2023-04-05.

Feick, M., Kleer, N., Tang, A., Krüger, A., 2020. The Virtual Reality Questionnaire Toolkit. In: UIST '20 Adjunct, Association for Computing Machinery, New York, NY, USA.

Figueiredo, L., Rodrigues, E., Teixeira, J., Teichrieb, V., 2018. A comparative evaluation of direct hand and wand interactions on consumer devices. Comput. Graph. 77, 108–121.

Fillwalk, J., 2015. ChromaChord: A virtual musical instrument. In: 2015 IEEE Symposium on 3D User Interfaces. 3DUI, IEEE, pp. 201–202.

Franěk, M., Mates, J., Radil, T., Beck, K., Pöppel, E., 1991. Finger tapping in musicians and nonmusicians. Int. J. Psychophysiol. 11 (3), 277–279.

Fujii, S., Schlaug, G., 2013. The harvard beat assessment test (H-BAT): a battery for assessing beat perception and production and their dissociation. Front. Hum. Neurosci. 7, 771.

Gaze and Commit, Gaze and Commit, https://learn.microsoft.com/en-us/windows/mixed-reality/design/gaze-and-commit. Accessed: 2023-04-05.

Gaze Cursor Component, Gaze Cursor Component, https://aframe.io/docs/1.4.0/components/cursor.html. Accessed: 2023-04-05.

Gil, H., Oakley, I., 2023. ThumbAir: In-air typing for head mounted displays. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 6 (4), 1–30.

Gourlay, M.J., Held, R.T., 2017. Head-mounted-display tracking for augmented and virtual reality. Inf. Disp. 33 (1), 6–10.

Hart, S.G., 2006. NASA-task load index (NASA-TLX); 20 years later. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting. Vol. 50, Sage publications Sage CA, Los Angeles, CA, pp. 904–908.

Holland, S., Mudd, T., Wilkie-McKenna, K., McPherson, A., Wanderley, M.M. (Eds.), 2019. New Directions in Music and Human-Computer Interaction. Springer.

Interact with Objects Remotely, Interact with Objects Remotely, https://developer.vive.com/resources/openxr/openxr-mobile/tutorials/unity/hand-tracking/interact-objects-remotely/. Accessed: 2023-04-05.

Jacob, R.J., 1995. Eye tracking in advanced interface design. In: Virtual Environments and Advanced Interface Design. Vol. 258, p. 288.

Jerald, J., 2015. The VR Book: Human-Centered Design for Virtual Reality. Association for Computing Machinery and Morgan & Claypool.

Kapur, A., Lazier, A.J., Davidson, P., Wilson, R.S., Cook, P.R., 2004. The electronic sitar controller. In: NIME. pp. 7–12.

Keller, P.E., Ishihara, M., Prinz, W., 2011. Effects of feedback from active and passive body parts on spatial and temporal parameters in sensorimotor synchronization. Cogn. Process. 12, 127–133.

Kelly, A., Klipfel, K., 2017. Audiovisual playground: A music sequencing tool for 3D virtual worlds. In: Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems. pp. 437–440.

Kim, W., Xiong, S., 2022. Pseudo-haptics and self-haptics for freehand mid-air text entry in VR. Applied Ergon. 104, 103819.

Krause, V., Pollok, B., Schnitzler, A., 2010. Perception in action: the impact of sensory information on sensorimotor synchronization in musicians and non-musicians. Acta Psychol. 133 (1), 28–37.

LaViola, Jr., J.J., Kruijff, E., McMahan, R.P., Bowman, D., Poupyrev, I.P., 2017. 3D User Interfaces: Theory and Practice. Addison-Wesley Professional.

Lécuyer, A., 2009. Simulating haptic feedback using vision: A survey of research and applications of pseudo-haptic feedback. Presence: Teleoper. Virtual Environ. 18 (1), 39–53.

Lorås, H., Aune, T.K., Ingvaldsen, R., Pedersen, A.V., 2019. Interpersonal and intrapersonal entrainment of self-paced tapping rate. PLOS ONE 14 (7), 1–14.

Lu, X., Yu, D., Liang, H.-N., Xu, W., Chen, Y., Li, X., Hasan, K., 2020. Exploration of hands-free text entry techniques for virtual reality. In: 2020 IEEE International Symposium on Mixed and Augmented Reality. ISMAR, IEEE, pp. 344–349.

Lucas Bravo, P.P., Fasciani, S., 2023. A human-agents music performance system in an extended reality environment. In: Proceedings of the International Conference on New Interfaces for Musical Expression.

Lyra VR, Lyra VR, https://lyravr.com/. Accessed: 2023-04-05.

Maestro MIDI, Maestro MIDI Player Toolkit, https://paxstellar.fr/. Accessed: 2023-04-05.

Mäki-Patola, T., 2005. User interface comparison for virtual drums. In: Proceedings of the International Conference on New Interfaces for Musical Expression. Vancouver, BC, Canada, pp. 144–147.

Mates, J., Radil, T., Pöppel, E., 1992. Cooperative tapping: Time control under different feedback conditions. Percept. Psychophys. 691–704.
McPherson, A., Jack, R., Moro, G., 2020. Action-sound latency: Are our tools fast enough? In: Proceedings of the International Conference on New Interfaces for Musical Expression. NIME, Brisbane, Australia, pp. 20–25.
Men, L., Bryan-Kinns, N., 2018. LeMo: supporting collaborative music making in virtual reality. In: 2018 IEEE 4th VR Workshop on Sonic Interactions for Virtual Environments. SIVE, IEEE, pp. 1–6.
Merchant, H., Zarco, W., Pérez, O., Prado, L., Bartolo, R., 2011. Measuring time with different neural chronometers during a synchronization-continuation task. Proc. Natl. Acad. Sci. 108 (49), 19784–19789.
Moore, A.G., Howell, M.J., Stiles, A.W., Herrera, N.S., McMahan, R.P., 2015. Wedge: A musical interface for building and playing composition-appropriate immersive environments. In: 2015 IEEE Symposium on 3D User Interfaces. 3DUI, IEEE, pp. 205–206.
Morimoto, C., Hida, E., Shima, K., Okamura, H., 2018. Temporal processing instability with millisecond accuracy is a cardinal feature of sensorimotor impairments in autism spectrum disorder: analysis using the synchronized finger-tapping task. J. Autism Dev. Disord. 48 (2), 351–360.
Mulder, A., 2000. Towards a choice of gestural constraints for instrumental performers. In: Trends in Gestural Control of Music. Vol. 315, Paris, France: Institut de Recherche et Coordination Acoustique Musique ..., p. 335.
Musescore, Musescore 3.6.2, https://musescore.org/en/3.6.2. Accessed: 2023-04-05.
Naef, M., Collicott, D., 2006. A vr interface for collaborative 3d audio performance. In: Proceedings of the 2006 Conference on New Interfaces for Musical Expression. pp. 57–60.
Numata, A., Terao, Y., Owari, N., Kakizaki, C., Sugawara, K., Ugawa, Y., Furubayashi, T., 2022. Temporal synchronization for in-phase and antiphase movements during bilateral finger-and foot-tapping tasks. Hum. Mov. Sci. 84, 102967.
Nusseck, M., Wanderley, M.M., 2009. Music and motion—how music-related ancillary body movements contribute to the experience of music. Music Percept. 26 (4), 335–353.
O'Boyle, D.J., Freeman, J.S., Cody, F.W., 1996. The accuracy and precision of timing of self-paced, repetitive movements in subjects with parkinson's disease. Brain 119 (1), 51–70.
Oculus Integration, Oculus Integration, https://assetstore.unity.com/packages/tools/integration/oculus-integration-82022. Accessed: 2023-04-05.
OpenXR, OpenXR, https://www.khronos.org/openxr/. Accessed: 2023-04-05.
OSC, OSC - Open Sound Control, https://ccrma.stanford.edu/groups/osc/index.html. Accessed: 2023-09-05.
Park, K., Kim, D., Han, S.H., 2020. Usability of the size, spacing, and operation method of virtual buttons with virtual hand on head-mounted displays. Int. J. Ind. Ergon. 76, 102939.
Patch XR, Patch XR, https://patchxr.com/. Accessed: 2023-04-05.
Pecenka, N., Engel, A., Keller, P.E., 2013. Neural correlates of auditory temporal predictions during sensorimotor synchronization. Front. Hum. Neurosci. 7, 380.
Peters, M., 1989. The relationship between variability of intertap intervals and interval duration. Psychol. Res. 38–42.
Pfeuffer, K., Mayer, B., Mardanbegi, D., Gellersen, H., 2017. Gaze+ pinch interaction in virtual reality. In: Proceedings of the 5th Symposium on Spatial User Interaction. pp. 99–108.
Pfeuffer, K., Mecke, L., Delgado Rodriguez, S., Hassib, M., Maier, H., Alt, F., 2020. Empirical evaluation of gaze-enhanced menus in virtual reality. In: 26th ACM Symposium on Virtual Reality Software and Technology. pp. 1–11.
Phillips-Silver, J., Trainor, L.J., 2005. Feeling the beat: movement influences infant rhythm perception. Science 308 (5727), 1430–1430.
Phillips-Silver, J., Trainor, L.J., 2008. Vestibular influence on auditory metrical interpretation. Brain Cogn. 67 (1), 94–102.
Point and commit with hands, Point and commit with hands, https://learn.microsoft.com/en-us/windows/mixed-reality/design/point-and-commit. Accessed: 2023-04-05.
Poupyrev, I., Ichikawa, T., Weghorst, S., Billinghurst, M., 1998. Egocentric object manipulation in virtual environments: empirical evaluation of interaction techniques. In: Computer Graphics Forum, vol. 17, Wiley Online Library, pp. 41–52.
Pressing, J., Jolley-Rogers, G., 1997. Spectral properties of human cognition and skill. Biol. Cybern. 76 (5), 339–347.
Reimer, D., Podkosova, I., Scherzer, D., Kaufmann, H., 2023. Evaluation and improvement of HMD-based and RGB-based hand tracking solutions in VR. Front. Virtual Real. 4.
Repp, B.H., 2003. Rate limits in sensorimotor synchronization with auditory and visual sequences: The synchronization threshold and the benefits and costs of interval subdivision. J. Motor Behav. 35 (4), 355–370.
Repp, B.H., 2005. Sensorimotor synchronization: A review of the tapping literature. Psychon. Bull. Rev. 12 (6), 969–992.
Repp, B.H., 2006a. Musical synchronization. In: Altenmüller, E., Kesselring, J., Wiesendanger, M. (Eds.), Music, Motor Control and the Brain. Oxford University Press, Oxford, pp. 55–75,
Repp, B.H., 2006b. Rate limits of sensorimotor synchronization. Adv. Cogn. Psychol. 2 (2), 163.
Repp, B.H., 2010. Sensorimotor synchronization and perception of timing: effects of music training and task experience. Hum. Mov. Sci. 29 (2), 200–213.
Repp, B.H., Keller, P.E., 2004. Adaptation to tempo changes in sensorimotor synchronization: Effects of intention, attention, and awareness. Q. J. Exp. Psychol. Sect. A 57 (3), 499–521.
Repp, B.H., Su, Y.-H., 2005. Sensorimotor synchronization: A review of recent research (2006–2012). Psychon. Bull. Rev. 20 (3), 1531–5320.
Reynaert, V., Berthaut, F., Rekik, Y., Grisoni, L., 2021. The effect of rhythm in mid-air gestures on the user experience in virtual reality. In: Human-Computer Interaction–INTERACT 2021: 18th IFIP TC 13 International Conference, Bari, Italy, August 30–September 3, 2021, Proceedings, Part III 18. Springer, pp. 182–191.
Ruspantini, I., D'Ausilio, A., Mäki, H., Ilmoniemi, R.J., 2011. Some considerations about the biological appearance of pacing stimuli in visuomotor finger-tapping tasks. Cogn. Process. 12, 215–218.
Ryan, J., 1991. Some remarks on musical instrument design at STEIM. Contemp. Music Rev. 6 (1), 3–17.
Scheurich, R., Zamm, A., Palmer, C., 2018. Tapping into rate flexibility: musical training facilitates synchronization around spontaneous production rates. Front. Psychol. 9, 458.
Serafin, S., Erkut, C., Kojs, J., Nilsson, N.C., Nordahl, R., 2016. Virtual reality musical instruments: State of the art, design principles, and future directions. Comput. Music J. 40 (3), 22–40.
Shima, K., Tsuji, T., Kandori, A., Yokoe, M., Sakoda, S., 2009. Measurement and evaluation of finger tapping movements using log-linearized Gaussian mixture networks. Sensors 9 (3).
Speicher, M., Feit, A.M., Ziegler, P., Krüger, A., 2018. Selection-based text entry in virtual reality. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. pp. 1–13.
Steed, A., Takala, T.M., Archer, D., Lages, W., Lindeman, R.W., 2021. Directions for 3D user interface research from consumer VR games. IEEE Trans. Vis. Comput. Graphics 27 (11), 4171–4182.
Sugioka, J., Suzumura, S., Kuno, K., Kizuka, S., Sakurai, H., Kanada, Y., Mizuguchi, T., Kondo, I., 2022. Relationship between finger movement characteristics and brain voxel-based morphometry. PLoS One 17 (10), e0269351.
Swinnen, S.P., Wenderoth, N., 2004. Two hands, one brain: cognitive neuroscience of bimanual skill. Trends Cogn. Sci. 8 (1), 18–25.
The Music Room, The Music Room, https://store.steampowered.com/app/431030/The_Music_Room/. Accessed: 2023-04-05.
Turchet, L., Hamilton, R., Çamci, A., 2021. Music in extended realities. IEEE Access 9, 15810–15832.
Turchet, L., Serafin, S., Cesari, P., 2013. Walking pace affected by interactive sounds simulating stepping on different terrains. ACM Trans. Appl. Percept. (TAP) 10 (4), 1–14.
Unity, Unity 2020.3.13, https://unity.com/releases/editor/whats-new/2020.3.13. Accessed: 2023-04-05.
Valbom, L., Marcos, A., 2005. WAVE: Sound and music in an immersive environment. Comput. Graph. 29 (6), 871–881.
van Vugt, F.T., 2020. The TeensyTap framework for sensorimotor synchronization experiments. Adv. Cogn. Psychol. 16 (4), 302.
Virtuoso VR, Virtuoso VR, https://fasttravelgames.com/games/virtuoso. Accessed: 2023-04-05.
Wakefield, G., Palumbo, M., Zonta, A., 2020. Affordances and constraints of modular synthesis in virtual reality. In: Proceedings of the International Conference on New Interfaces for Musical Expression. pp. 547–550.
Wanderley, M.M., Orio, N., 2002. Evaluation of input devices for musical expression: Borrowing tools from hci. Comput. Music J. 26 (3), 62–76.
Wang, Y., Martin, C., 2022. Cubing sound: Designing a NIME for head-mounted augmented reality. In: NIME 2022.
WebXR, WebXR, https://immersiveweb.dev/. Accessed: 2023-04-05.
Wiederhold, B., et al., 2016. Human instruments: Accessible musical instruments for people with varied physical ability. In: Annual Review of Cybertherapy and Telemedicine 2015: Virtual Reality in Healthcare: Medical Simulation and Experiential Interface. Vol. 219, IOS Press, p. 202.
Wing, A.M., Kristofferson, A., 1973. The timing of interresponse intervals. Percept. Psychophys. 13 (3), 455–460.
Wingrave, C.A., Tintner, R., Walker, B.N., Bowman, D.A., Hodges, L.F., 2005. Exploring individual differences in raybased selection: strategies and traits. In: IEEE Proceedings. VR 2005. Virtual Reality. pp. 163–170.
Xu, W., Liang, H.-N., He, A., Wang, Z., 2019. Pointing and selection methods for text entry in augmented reality head mounted displays. In: 2019 IEEE International Symposium on Mixed and Augmented Reality. ISMAR, IEEE, pp. 279–288.
Zanto, T.P., Padgaonkar, N.T., Nourishad, A., Gazzaley, A., 2019. A tablet-based assessment of rhythmic ability. Front. Psychol. 10, 2471.
Zappi, V., Gaudina, M., Brogni, A., Caldwell, D., 2010. Virtual sequencing with a tactile feedback device. In: International Workshop on Haptic and Audio Interaction Design. pp. 149–159.
Zatorre, R.J., Chen, J.L., Penhune, V.B., 2007. When the brain plays music: auditory–motor interactions in music perception and production. Nat. Rev. Neurosci. 8 (7), 547–558.