Spatialised Audio in a Custom-Built OpenGL-based Ear Training Virtual Environment

Karsten Pedersen, Vedad Hulusic, Panos Amelidis, Tim Slattery
Bournemouth University
{pedersenk, vhulusic, pamelidis, tslattery}@bournemouth.ac.uk

***Abstract*—Interval recognition is an important part of ear training - the key aspect of music education. Once trained, the musician can identify pitches, melodies, chords and rhythms by listening to music segments. In a conventional setting, the tutor would teach a trainee the intervals using a musical instrument, typically a piano. However, this is expensive, time consuming and non-engaging for either party. With the emergence of new technologies, including Virtual Reality (VR) and areas such as edutainment, this and similar trainings can be transformed into a more engaging, more accessible, customisable (virtual) environments, with addition of new cues and bespoke progression settings. In this work, we designed and implemented a VR ear training system for interval recognition. The usability, user experience and the effect of multimodal integration through addition of a perceptual cue - spatial audio, was investigated in two experiments with 46 participants. The results show that the system is highly acceptable and provides a very good experience for users. Furthermore, we show that the added spatial auditory cues provided in the VR application give users significantly more information for judging the musical intervals, something that is not possible in a non-VR environment.**

***Index Terms*—virtual reality, ear training, spatial audio, OpenGL**

I. Introduction

Ear training is a set of skills by which musicians learn to use their hearing to identify all music elements such as pitches, musical intervals, melodies, chords and rhythms. It plays a vital role in the training of musicians. The peak of developing aural skills through ear training is the transcription and transmission of music entirely by ear [1]. In addition, ear training enhances the pleasure of music listening and sharpens a musician's ears for the study, comprehension, performance, and creation of music [2].

Ear training can take many forms but usually it starts by learning to recognise various melodic passages. In order for a musician to determine the notes in a melody, they must have the ability to distinguish and recognise musical intervals. Improving pitch recognition skill by way of ear training provides the means for musicians and music students to learn the relationships of the musical pitches and to attain good listening skills. Traditionally the training process requires a knowledgeable teacher (or partner) to play the music patterns and to assess the provided answers. Nowadays, there exist a variety of free and commercial software, made available as stand-alone, browser-based applications or for use on mobile devices, designed for ear training. G.U.I.D.O. [3], was the first ear-training "software" developed in the mid 1970s using the PLATO mainframe to provide programmed instruction for the recognition of intervals, melodies, chords harmonies and rhythms for college music students [2]. EarMaster [4] launched in 1996 is still a very popular ear training software, which includes a beginner's course, general workshops, Jazz workshops and customised exercises. EarMaster covers all topics of ear training such as interval comparison, interval scale, chord identification, chord progression, chord inversion identification, rhythm clap back and rhythm error detection.

While Virtual Reality (VR) - encompassing technologies to perceive visual, auditory and potentially haptic sensory input from a virtual environment - is being used in many domains, including training and simulation, there are still no known VR-based systems for ear training. Current VR applications for music have a strong emphasis on making or enjoying music, but there is a lack of focus on utilising the possibilities provided by this technology for music education purposes and, in particular, for ear training.

It has been shown in other fields that detection, discrimination, and localisation are typically performed more reliably and faster when bimodal cues are available [5]. Building on this knowledge, we assume the tone and interval recognition performance can be enhanced by spatial localisation of the sound origin, and its collocation with its visual counterpart, i.e. the piano note. To validate this and test the system we conducted two experiments looking at usability, user experience and the effect of multimodal integration of visual and spatial auditory information on interval recognition performance.

There are three main contributions of this work.

- The novel VR system for ear training using the spatial audio delivered in the Virtual Environment (VE) as an additional auditory cue;
- The validation of the system through two user studies with 46 participants, showing very high acceptability rate and equally good user experience with the system, as well as confirmation of the multimodal cue integration on interval recognition;
- The implementation of the system using only open-source tools, making it licence free and very light, allowing for usage on any VR system, including low-price, stand alone devices.

II. Related work

There have been many VR music systems created in recent years. A series of such applications has been made by

Music Technology Lab at MIT, some of which are available online at MIT's Music Technology Lab website [6]. *Ari* is an immersive storytelling experience in which users with very little VR and/or music experience are able to interact with the environment and experience a song through the space around them. *Drumhead* aims to introduce aspiring musicians to drumming, using a variety of popular songs, in an interactive way and in VR. *Orchestral Explorations* uses VR to allow users to easily and intuitively experience classical music in depth. *VR Sandbox* is an exploration of different techniques and interactions with music composition in VR focusing on immersion and enjoyment.

Harmonix Music VR [7] by Harmonix allows the user to enjoy listening to music in a VR environment. Through different 'worlds' the user can experience a variety of music jamming sessions. Other VR for music systems exist, such as *SoundStage VR* for music jamming, *Exa* for making music and collaborating with others and *AliveInVR* for controlling a Digital Audio Workstation. A thorough review of technology and software for virtual and augmented reality in music has been conducted by Sefarin et al. [8].

A. Interaction in VR

Virtual reality allows users to experience the virtual world and interact with it using various interaction techniques. To move through and interact with the VE, a few interaction modes are used: locomotion, selection, manipulation and scaling. In addition, menu interaction is used for performing other actions that are difficult or not possible to complete through previously mentioned interaction modes [9]. Menu selection and interaction techniques in 3D VEs are not as established and standardised as for 2D spaces and systems. One approach used in the literature is to utilise standard 2D menus in 3D space either as a heads-up display (HUD) or as floating menus (in 3D space) [10]. Alternatively, more natural 3D, physical paradigms could be used in forms of spin/ring menus. However, the most natural type of menus in 3D VEs can be achieved through diegetic interfaces. These interfaces exist as part of the game world and are visible to both the player and the player's character [11].

B. Multimodal Integration and Spatialised Audio in VR

Another important aspect of any VR experience is the addition of a high-fidelity audio through sound generation, reproduction and propagation [12]. Spatial audio is not only the main contributor to enhanced immersion in VR experiences, but also has a an effect on spatial perception in VR and spatial visual processing [13]–[15]. Malpica et al. have shown that crossmodal effects, such as the effect of spatial audio on visual perception, which exist when used with conventional displays, are also present in VR [13]. In their study, Yong and Wang. [14] have shown that there is a higher spatial recall accuracy when spatial auditory cues are present in VR. Additionally, the results from a study by Høeg et al. [15] indicate that having binaural audio reduces reaction time in a visual search task.

While there is a plethora of VR applications used to either play, compose or enjoy music, and research on capturing, generating and reproducing spatial audio in and for VR, to the best of authors' knowledge, there are no studies looking at how spatial audio can be utilised for ear training in VR.

III. Virtual Reality Ear Training Systems

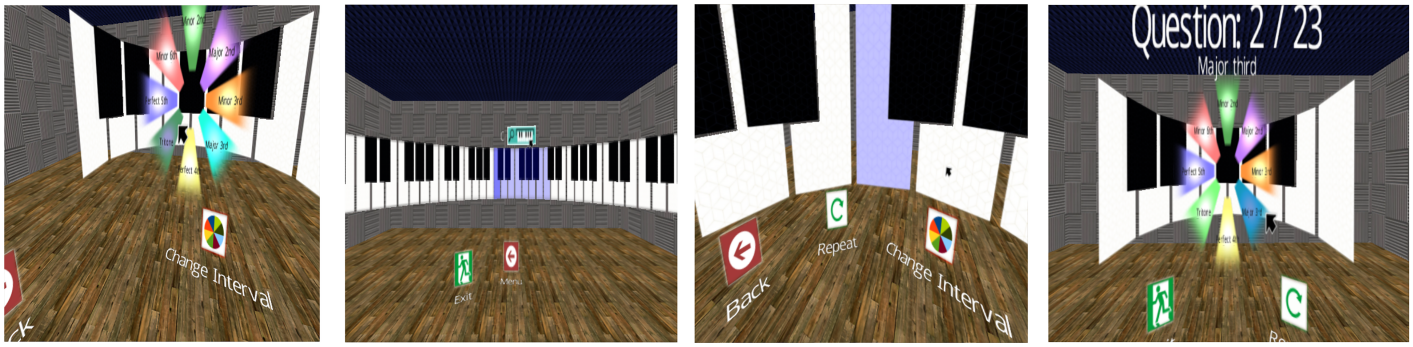
The main aim of the presented system is to provide musical interval training. It consists of two subsystems: training and test. The former is used to teach and train users music intervals, while the latter helps them self evaluate the progress. The main idea in this research is to investigate how multimodal integration through addition of an extra perceptual cue - the spatial audio, helps in interval recognition through spatial identification of the sound origin.

A. Design Considerations

Traditionally, the process of ear training focuses on teaching students to identify the most basic elements of music (i.e. intervals, simple melodies, simple triads, scales, and simple rhythm). There are a variety of approaches to ear training but usually exercises include simple dictation [16]. When it comes to interval identification, the instructor plays (on the piano) intervals in both ascending and descending order for the students to identify. The intervals involved are comprised of two notes played either subsequently (melodic) or simultaneously (harmonic). After hearing an interval two or three times, students are asked to write it on score paper or to say what interval they heard. When training students in identifying specific intervals, instructors often play a known melody (i.e. from a pop song) which starts with the interval to identify, allowing students to correlate and memorise the quality of the interval with a piece of music.

In our VR system, the music instructor (as well as the whole experience of ear training) is substituted by a VR system, playing music intervals in ascending order. In the training phase participants can familiarise themselves with the system and the intervals. In the first experiment (see Section IV-A) there were two intervals belonging to the perfect consonant intervals group (perfect fifths and perfect fourths), one belonging to the imperfect consonant intervals group (major sixth), and one belonging to the dissonant intervals group (major seventh). These intervals were selected to allow the user to train with a variety of both consonances and dissonances. In the second experiment (see Section IV-B) we used eight intervals, from minor second to minor sixth, including minor third and triton that were used in the test condition, see Fig. 1.

The system is designed to be used in a sitting position. While the selection and menu interaction in the first experiment were performed by using HTC Vive controller and a laser pointer, it was decided to use a computer mouse in Experiment 2. The main reason for this is to allow the user to keep the mouse pointer located at the same place and easily replay the interval while being able to rotate their head - something that was not as easy when holding the Vive controller. As spatial audio was a key component of the system, it was important to utilise the 3D space, intrinsically provided in



(a) Interval selection. (b) Octave selection. (c) Interval training. (d) Test mode.

Fig. 1: Screenshots from the training and test sequence.

VR. In addition, to achieve a desired multi-sensory experience, the keys and their corresponding audio clips (sound sources) were positioned around the player in a semicircle. To enhance sound localisation on a horizontal 2D plane which could help with note and thus interval identification, the horizontal angle per note had to be maximised. At the same time, we did not want the user to rotate (the head) more than 90° to each side. Therefore, for both training and testing, we decided to use a semicircular keyboard with 13 keys, representing one octave, see Fig. 2.

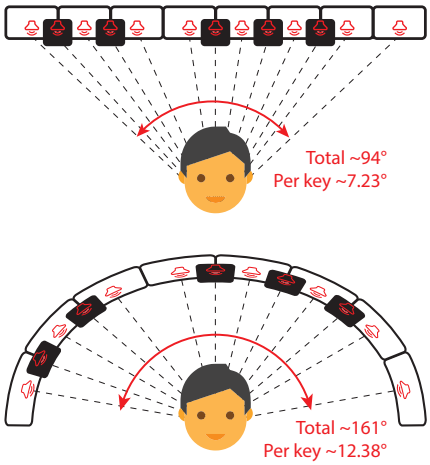


Fig. 2: The auditory angle for the flat and semicircled piano keyboard. The angles are approximated as the keyboard layout varies depending on the octave.

B. System Development

1) System version 1 - Unity: The system was initially developed using the Unity Engine (v. 2018.2.16f1) as it allows a straightforward VR integration providing a highly optimised rendering pipeline, rapid iteration capabilities and cross compatibility with many different platforms. The design of the system was based on a state driven approach with three states: 'Menu', 'Training' and 'Testing'. The 'Menu' state was used for all menu based functionalities within the system. The 'Training' state is where the user gets complete control over most of the system's features, selecting intervals and octaves and being able to hear the sounds from the keyboard in real time. The 'Test' state allows the user to test their skills inside the software. The test involves the user listening to an interval and guessing which interval was played.

The virtual piano used within the 'Test' and 'Training' states was designed to utilise the Stereo Panning feature within Unity's Audio System. In both modes (training and test), the piano keys were placed in a 180 degree semicircle around the user to create an enclosure, which would allow the audio to separate into left and right channels. The interaction with the system was controlled by the Vive controller, only requiring a single (physical) controller. Selection was implemented using the laser pointer and ray casting, allowing easy detection of the object pointed at and immediate visual feedback to the user. While in the 'Test' mode, the system records multiple data per interaction, including the user ID, condition (Mono/Stereo Panning), interval, lower key in the interval, user response (Repeat/Correct/Incorrect) and a timestamp, and saves it into a '.csv' file.

2) System version 2 - OpenGL: Looking to the future, rather than using existing consumer APIs such as the SteamVR implementation of OpenVR or the Oculus SDK, it was instead deemed important to utilise a vendor neutral API. OpenHMD was the chosen API to interface with the VR hardware, Fig. 3. Not only does this experimental open-source API form a large proportion of the underlying platform for Monado (an implementation of Khronos's future OpenXR API [17]) but it also provides much needed flexibility and compatibility when it comes to hardware support. For example, the SteamVR runtimes require a fairly recent set of graphics card drivers in order to access the direct mode functionality from Vulkan [18]. OpenHMD instead simply turns on the HMDs screen using largely undocumented "magic bytes" and allows for output onto it in a standard way just like any other monitor. This allows for a much wider range of graphics cards to be supported, including those as old as an integrated Intel GMA 965 from a 2003 Lenovo Thinkpad; the use of open-source graphical drivers, such as the Nouveau drivers for NVIDIA GPU hardware; and the use of older NVIDIA hardware on modern operating systems where the vendor no longer provides a compatible proprietary driver.

The reduced requirement on performance and features is important because with the increasing popularity of mobile hardware, there is less guarantee that hardware will always

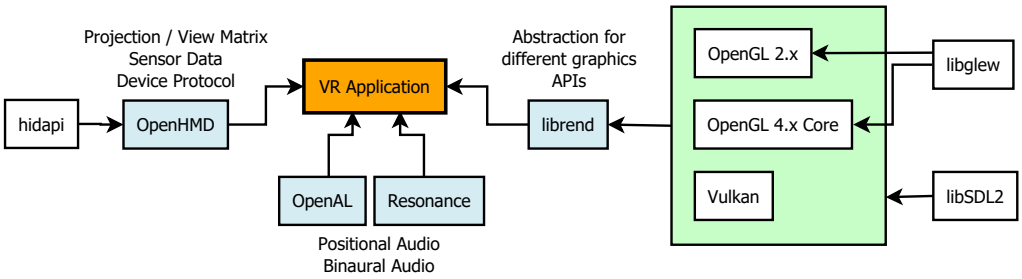


Fig. 3: A dependency diagram showing the architecture of the system with its components.

become more powerful; often a focus is instead on reducing energy requirements and increasing battery life. A future aim for our tool could very much involve running on mobile devices as a portable teaching aide. Requiring a powerful machine simply to satisfy arbitrary runtime requirements would severely limit our ability to achieve that. Avoiding the Oculus SDK was particularly crucial because Oculus had dropped support for important operating systems such as Linux in their driver since the Developer Kit 2 (DK2) revision of the hardware.

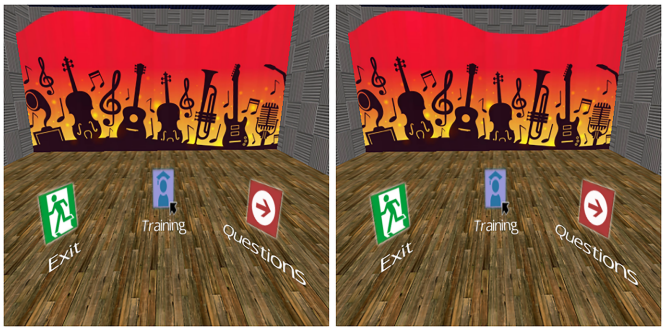


Fig. 4: Screenshot of the research tool when in debug mode. The two viewports can be seen rendering the same scene but from subtly different viewpoints.

Interestingly, OpenHMD provides little else other than the protocol required to turn on the headset and generating a view matrix from the sensors relating to the orientation for the individual viewports (Fig. 4). Unlike commercial solutions it merely provides hardware access rather than a complete framework, providing opportunity for exploring new ideas during the parts of bespoke implementation. The rest of the setup is left to the developer such as implementing multiple cameras to render to the specific viewpoint textures, sending it through a warp shader to counteract the warping from the lenses and finally rendering these textures to a single screen. Again this freedom allows for a more elegant integration with existing engines because it enforces no specific design pattern or framework.

The tool has performed well, with no negative feedback from the users regarding jerkiness or lagging. The machine used to provide the experiment station had fairly modest specifications such as an Intel Core i3 and an NVIDIA Quadro K400 and the Nouveau graphics driver. Some initial tests have shown hardware such as the Raspberry Pi 3 as adequate for our tool to run. The main reason for this is that no heavy off-the-shelf engine was used to implement the tool. The complexity of the shaders were also kept to a minimum, mostly to highlight keys and little else to avoid distracting the user. The software was developed using modern C++ producing fairly optimised native binaries to be executed directly on the processor rather than interpreted or 'JITed' which is fairly typical in some of the common alternatives such as Unity using Mono [19] (an open-source implementation of .NET) as a scripting engine which can unfortunately have a negative impact to performance [20].

Due to the fact that there is very little in terms of vendor middleware required as dependencies for our tool and underlying custom engine, and it being developed in a very portable way using entirely open-source or bespoke software and tools, it is very possible to port the solution to a wide variety of platforms such as Android, iOS and Emscripten (using WebVR) and even more exotic platforms such as FreeBSD and ARMv7 Linux. It is also made up of fairly modular components so the sound system (OpenAL) can be replaced with an alternative such as Google's Resonance if a future platform requires us to do so, Fig. 3. Likewise, if a platform does not provide access to OpenGL, then the underlying renderer can be replaced leaving the rest of the application and engine untouched. This bespoke technology is called Rend and provides a layer of abstraction above the underlying graphics API. In our case both OpenGL and Vulkan have been successfully integrated with initial ports to DirectX underway. It is important to note that OpenGL has been a primary focus of this research tool over Vulkan simply because relatively few hardware configurations can support Vulkan yet. Fig. 5 provides an overview of the types of wrapped functionality for the graphics abstraction technology.

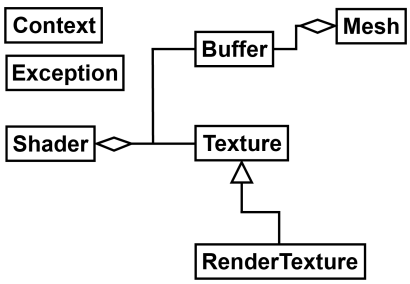


Fig. 5: Simple class diagram showing the extent of Rend. This abstraction layer has been kept fairly trivial in order to make porting it to other graphics APIs as simple as possible.

Through previous prototypes of the system, it was noted that

the use of the traditional consumer VR game controllers was fairly unwieldy. Considering the requirement that the users were moving their head around in order to hear the audio coming from various relative directions, it was often awkward to keep the laser pointer hovering over the individual buttons. It was instead desirable to utilise the stability of a traditional mouse resting on a table to select the buttons instead. This was much less likely to drift whilst the focus and direction of the user was elsewhere. The handling of the mouse within the 3D world was fairly novel compared to traditional VR systems; instead it would circle the user's position, creating a more natural feel than a traditional 2D mouse cursor. When it would intersect 3D buttons, it would rest on top rather than keep its original depth. This functionality was implemented to avoid a strange depth artefact where the cursor would appear in the distance even though it would display on top of the button. This also gave the user a richer feel of the 3D scene, almost as if they were touching the buttons.

The architecture of the spatial audio tool and the underlying custom engine follows that of the Component Entity System (CES); a fairly typical design for modern computer games but it even finds use within industrial and military simulations [21]. It favours composition over inheritance and allows for a much more dynamic approach to controlling multiple objects at runtime.

By addressing the implementation with the unique technical aim of a more flexible and modular approach utilising individual components as opposed to a large self-contained framework it allowed for a greater understanding of the underlying technology and algorithms involved in the areas surrounding VR. In addition, it has also yielded a positive outcome including a viable prototype satisfying all the original objectives. This functionality provided by the system can be seen in Fig. 6.

IV. User Study

The user study was conducted on both systems (see Sections III-B1 and III-B2) via two separate experiments. The main aims of the studies were to investigate the usability and user experience (UX) of the system, its effectiveness, and the effect of multimodal integration through addition of spatial audio on interval recognition. The latter was investigated in a different way in the two experiments as described in the following sections.

A. Experiment 1: Mono Versus Stereo Panning

1) Design: The user study has been conducted focusing on three aspects: usability, UX and system efficiency. The latter was observed by looking at the training and test length, number of interval repeats and user score. In addition the effect of the spatial auditory cue was analysed. The experiment was conducted with two user groups, one of which has been exposed to non-panned mono sound ('Mono' condition) and one to stereo panned audio ('SP' condition). Besides the basic demographics data, the participants were asked for their music education level and VR experience. For evaluating both the usability and UX two commonly utilised questionnaires were used: System Usability Scale (SUS) [22] and the Game Experience Questionnaire [23]. The system efficiency was evaluated by analysing the in-game data from the custom made game analytics. Our research hypothesis was that there will be better performance when the sound is stereo panned (SP).

2) Participants: 27 participants volunteered for the study with no inclusion/exclusion criteria. Out of 26 participants, 23 were male and 3 female, with the age ranging from 18 to 50 (with an average age of 26.46). One participant did not want to disclose their gender identity and age. 12 participants were assigned to the Mono condition, while the other 15 were trained and tested with the 3D audio. On the scale $0-3$ (0-none, 1-basic, 2-moderate, 3-high), the average report music education was 1.4. Similarly, the VR experience was rated 1.3 on average.

3) Apparatus: All the experiments took place in a dedicated, quite test room. The system was run on a VR-Ready MSI Stealth Pro GS73 VR laptop. The visual stimuli were displayed on an HTC Vive head-mounted display (HMD), while the audio was delivered through Sennheiser HD-25-ii headphones. Participants used standard HTC Vive controller for interaction with the system.

4) Procedure: Upon entering the experimentation room, the participants were given the participant information sheet and participant agreement form to read and sign. They were then given detailed instructions on how to use the system and asked if they had any questions about the nature of the experiment and their task. This was followed by a demo session, which used the same interface and both modes: training and test. However, the available intervals were different from those used in the main study. After finishing with the demo, they were offered a short break.

In the main study, the participants were first trained on four music intervals: perfect 4th, perfect 5th, major 6th and major 7th. They were told they could take up to 10 minutes for the training, although they were not interrupted if they took longer. In the training, they could select the interval and the octave (from the whole 88-key piano) in which they want to be trained. The could repeat the interval played from the same key, play the interval from another key within the octave or shift the octave by one key up or down. They were then asked to take the test in which they were hearing the same intervals (as in the training) and had to recognise and select the one they had heard. Each of the four intervals were played in three octaves (low, middle and high) twice, in a random order, resulting in 24 intervals per participant. They could see the starting key but not the second key of the interval. Depending on the condition (Mono or SP) the sounds they were hearing were either played uniformly to both ears, or with a stereo panning based on the key position in the VE. They could replay each interval up to five times and had to select one of the four intervals from the 3D menu. Once selected they had to press the Next button, which allowed them to take a short break whenever they needed it.

B. Experiment 2: 3D Audio With the Origin Offsets

1) Design: To address the issues and findings from the first experiment, and to further improve the system's performance

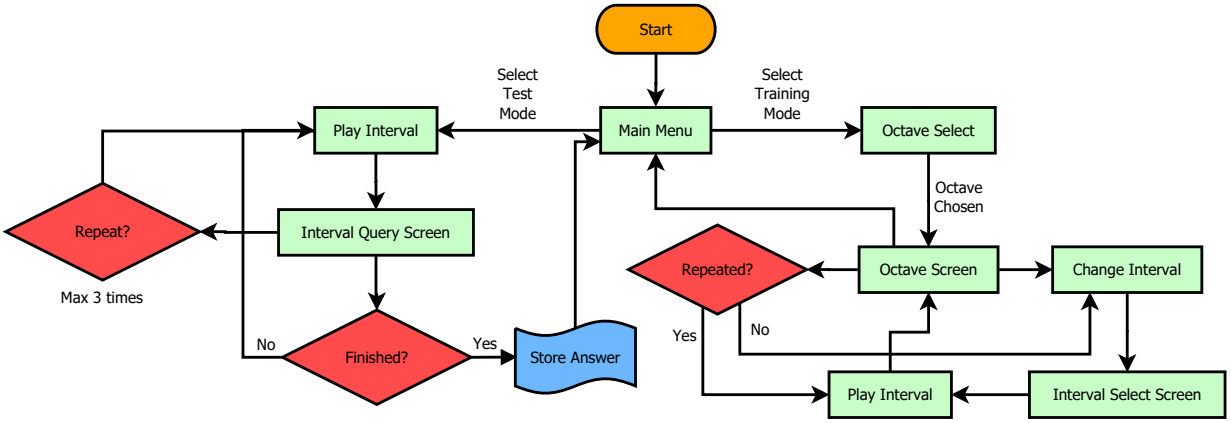


Fig. 6: A flow diagram demonstrating the interaction between the user and the system.

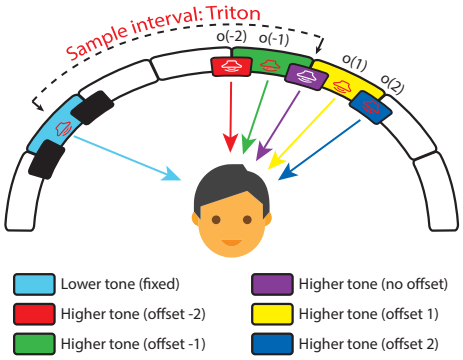


Fig. 7: The illustration of the interval offset implementation in the system on example of the 'triton' interval. The sounds/notes in each five conditions are the same, but the sound origin (location in 3D space) of the second note changes.

and platform/hardware independence, a new system was developed (see Section III-B2) and the experiment redesigned. Firstly, we wanted to test the system only with musically educated people that either actively play a music instrument or sing in a choir. Secondly, we wanted to isolate the effect of spatial auditory cue on interval recognition. Therefore, in this experiment all the participants were exposed to the enhanced 3D spatial audio using OepnAL HRTF-based rendering. For each of the two tested intervals the audio of the second (higher) note could originate from one of the five locations: the corresponding piano key, one key to the left/right (offset $= \pm 1$) or two keys to the left/right (offset $= \pm 2$), see Fig. 7. The left (lower) note was randomly selected and was not influenced by the offset. This allowed us to see how the addition of a spatial auditory cue influences interval recognition. Our research hypothesis was that user errors in interval judgements would tend to be in the direction of the offset—indicating their impact in multimodal perception.

The participants have completed the same questionnaires as in the first experiment and provided the demographics data with one extra piece of information - 'the experience with interval recognition'. The data collected during the gameplay was: test interval, low and high note of the interval, the offset, user's answer, number of repeats and the timestamp.

2) Participants: In this experiment 19 participants volunteered (M=9, F=10). The age ranged from 18 to 48 (with an average age of 26.5). In this experiment only one condition was considered - 3D audio. The inclusion criteria for the participation was a high level of musical education. Therefore, the participants were either music students or members of the local orchestra or choir. On the scale $0-3$ (0-none, 1-basic, 2-moderate, 3-high), the average reported music education was 2.5. Similarly, the average experience with interval recognition on the same scale was 1.7. Finally, the VR experience was rated 0.8 on average, closest to 'used once' option in the questionnaire.

3) Apparatus: This experiment took place in a dedicated, quite test room. The system was run on a mid-range desktop machine. The visual stimuli were displayed on an HTC Vive Pro head-mounted display (HMD), while the audio was delivered through Sony WH-1000XM3 noise-cancelling headphones. Participants used a computer mouse to interact with the system.

4) Procedure: After reading the participant information sheet and signing the participant agreement form the participants were given detailed instructions on how to use the system and asked if they had any questions about the nature of the experiment and their task. This was followed by a training session, where they could listen to any of the eight intervals on any part of the 88-key piano keyboard, see Fig. 6. After finishing with the training, they were offered a short break.

During the test session, the participants were informed that the first three trials are for familiarisation purposes and will not be evaluated. This was followed by 20 test trials. There were two intervals tested: minor third and triton. Each interval was played twice in all five previously described conditions (offset $= 0, \pm 1, \pm 2$). All intervals were played in the same octave (C4 - C5) but from a random starting note/key. The participants could replay each interval twice and had to select one of the eight intervals from the pie menu.

V. Results

A. Usability Study

Usability testing was performed using the SUS scale [22]. The questions and the corresponding responses from the

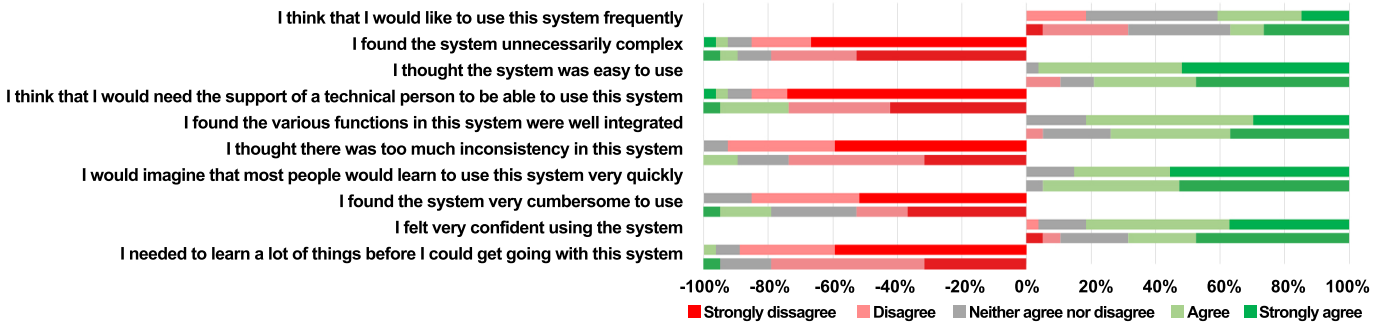


Fig. 8: Questions from the usability questionnaire (SUS) and the distribution of the user responses. The top bars for each question represent the scores from Experiment 1 and the bottom from Experiment 2.

study are presented in Fig 8, where the top bars for each question represent the results of Experiment 1 and the bottom of Experiment 2. Using the score calculation as suggested in [22], the overall SUS score for the initial system (used in Experiment 1) was found to be 80.74, whereas for the new system (Experiment 2) was 72.66, see Fig. 9. This confirms that both systems were well designed and accepted by the users, even though there are certain elements which could be further improved. Furthermore, this shows that the 'light' system in Experiment 2, implemented using a vendor neutral open-source API, that can theoretically run on any hardware, has been rated similarly to that using a consumer API and a game engine, requiring significantly higher processing power. Having slightly lower score in Experiment 2 might be due to the fact that these participants had slightly lower experience in VR (1.3 in Exp1 versus 0.8 in Exp2) and therefore needed more learning and adaptation time.

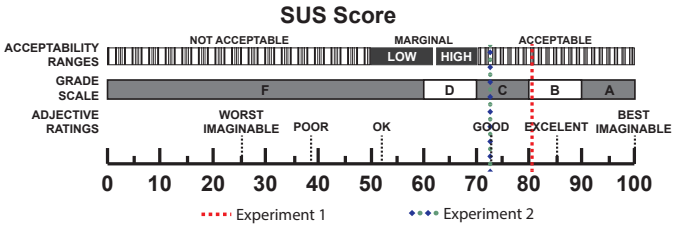


Fig. 9: Grade rankings for SUS scores as proposed by Bangor et al. [24]. Red dashed line represents the score obtained in Experiment 1, whereas the blue-green line depicts the score from Experiment 2.

B. User Experience (UX)

In this study we used a subset of 17 questions from the Core module of the Game Experience Questionnaire (GEQ) [23], covering all seven UX components. The questions were evaluated on a 5-point Likert scale. The results for all the questions and components from both experiments are presented in Table I. The results from both experiments are similar, with minor differences across the components. Although the participants in the second experiment were less experienced with VR systems, they expressed higher competence. In addition, they felt more immersed and occupied with the task (flow) than the participants in Experiment 1. Finally, the participants using the new system felt slightly more frustrated and bored. This could be due to the sample's music education level (1.4 in Exp1 versus 2.5 in Exp2), finding the interval recognition too easy due to the interval repetition.

In addition to the presented usability and UX question, two additional questions were added to the questionnaire for the SP audio user group in Experiment 1 and all the participants in Experiment 2, Table II. The results of these two questions indicate that the spatial audio cue, for the SP group, was helpful in interval recognition (3.42), but not enough to rely entirely on this cue (2.53). Similarly, in Experiment 2, these scores were higher (3.67 and 2.79 respectively) meaning there was even more inclination towards relying on the spatial auditory cue.

Finally, at the end of the questionnaire, there was an open-ended question "Is there something you would change, add or remove from the system, or anything you would like to comment". In Experiment 1, 19 participants (70.4%) responded to this question. The first issue raised is about the user position in the VE, suggesting it moves further from the keys. This was a design consideration, but also a necessity, in order to have as wide auditory angle per key as possible, see Fig. 2. The other issue reported by three participants was the implementation limitation, where users upon clicking on a button, had to remove the laser pointer away from it in order to click on it again. The third element that was criticised is the poor visual appearance of the system. Namely, three users wrote that the system should be more aesthetically pleasing. Finally, two users wrote positive feedback, saying that everything was great. In Experiment 2, 11 participants (57.9%) responded to the question. One participant reported a noticed issue with the system. Namely, during the session, this user's orientation in the VE was slightly offset in one direction. Nevertheless, this and one more participant reported that it was fun to use the system. Another participant reported "I would like to do the test in the octave I practised at. For interval recognition I would like to hear the notes separately and then together". One participant wrote that they haven't done theory for 3 years and the system made it easy to "pick it back up". Two participants reported they wanted to "zoom out" as the keys were too big. Other three participants reported that the system "worked really well", and added "I just need some practice with intervals and this is a perfect way to help me learn", "I think the spatial cue is a very interesting thing and probably

TABLE I: UX mean score values per question and per component on a 1-5 scale, 1 being 'Not at all' and 5 being 'Extremely'. The first column represents the order of the question as found in the questionnaire.

No	Component	Question	Experiment 1		Experiment 2	
			Score (Q)	Score (C)	Score (Q)	Score (C)
2	Competence	I felt skillful	3.11	3.15	3.32	3.39
11		I felt successful	3.19		3.47	
3	Immersion	I was interested in the task	4.44	3.88	4.53	3.95
7		It was aesthetically pleasing	3.33		3.16	
15		I found it impressive	3.85		4.16	
5	Flow	I was fully occupied with task	4.70	4.04	4.63	4.32
8		I forgot everything around me	3.37		4.00	
13	Tension/ Annoyance	I felt irritable	1.56	1.59	1.32	1.61
16		I felt frustrated	1.63		1.89	
14	Challenge	I felt challenged	3.85	3.31	4.00	3.21
17		I had to put a lot of effort into it	2.78		2.42	
6	Negative affect	I found it tiresome	1.59	1.48	1.89	1.55
10		I felt bored	1.37		1.21	
1	Positive affect	I felt content	3.81	4.10	3.84	4.04
4		I thought it was fun	4.41		4.16	
9		I felt good	4.00		4.00	
12		I enjoyed it	4.19		4.16	

very helpful for people with not that much experience" and " I think it would be a useful tool for interval training".

TABLE II: Questions used to evaluate the presence of spatial auditory cue with the corresponding mean score values.

Question	Score (Exp 1)	Score (Exp 2)
The spatial audio cue helped me in recognising the interval	3.42	3.67
I relied on the audio spatial cue (its origin) for recognising the interval	2.53	2.79

C. System Effectiveness and Multimodal Integration

During the testing, various user and game data was collected. These include the demographic data, users' music education level, VR experience, game score, number of repeats per interval, training and test times. In this section the results of the data analysis will be presented.

1) Experiment 1: Since there were two user groups (Mono and 3D) and multiple independent variables, the analysis of covariance (ANCOVA) was utilised. The dependent variable (DV) was the user score, the fixed factor was the condition (Mono, 3D) and the covariates were music education level, VR experience, number of repeats, training time and test time. The result of the Levene's test was not significant ($p = .191$), hence the assumption of homogeneity has been met. This means that the relationship between the dependent variable and the covariates was similar in each of our treatment groups. The test of between-subject effects revealed that the music education level significantly predicts the score ($p < .05$). Even though the group mean value for the 3D condition ($\mu_{3D} = 14.27$) was slightly higher than the group mean for the Mono group ($\mu_{Mono} = 12.66$), the effect of spatial audio, i.e. the fixed factor, was not statistically significant ($p = .39$). These results were further confirmed by looking at the partial correlations between the score (DV) and all the other independent variables (IV) for which the bivariate (Pearson) correlation was found as significant, while controlling the effect other IVs. the results of the test confirmed that the only significant correlation exists between the score and music education ($r = .466, p < .05$).

Finally, factorial repeated-measures ANOVA was utilised to test for the effect of interval on user scores for both conditions (Mono and 3D). The within-subject factor was interval (major 6th, major 7th, perfect 4th and perfect 5th), while the between-subject factor was the condition. The scores were computed as frequencies of correct user responses, i.e. number of correct responses per interval (out of six trials). The Mauchly's test confirmed that the assumption of sphericity was met ($W = .924, p = .495$), meaning that the variances across conditions are similar and there is no need to correct the F-ratio for this effect. The results of the main test show that the user performance when identifying the interval was not affected by the interval, $F(1, 24) = .983, p = .331$. The pairwise comparison for the main effect of interval did not show significant effect between either interval pair ($p > .05$). The mean values with corresponding confidence intervals for both conditions are displayed in Fig. 10.

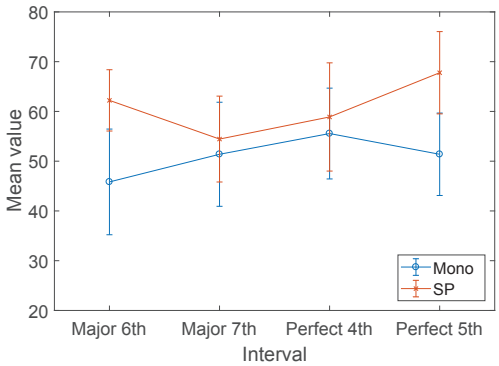


Fig. 10: Mean values with corresponding confidence intervals for frequencies of correct user responses for both conditions.

2) Experiment 2: One of the benefits of using VR over other computer applications for musical interval training is the ability within VR to provide the user with the perception of sounds at different locations. The purpose of the offset

manipulation in Experiment 2 which shifted the perceived location of the second key in the interval, was to examine the extent to which users can take advantage of such localised sound information. If users were to make their interval judgements in Experiment 2 using only the information about the location of sounds, then the offset manipulation would be mirrored in their answers. That is, a minor third interval with a plus 2 offset would be judged as a perfect four (two semitones wider). To assess this, we transformed the user's answers by subtracting the correct interval from them so that a value of zero represented a correct answer. If users were responding based exclusively on sound locations, then our offset variable would perfectly predict this new transformed response variable with a slope equal to one. However, it is also possible that humans are incapable of taking advantage of the spatial auditory cues available from the VR system when judging intervals. If this were the case then the slope of the offset value should not be predictive of users judgement (i.e slope not significantly greater than 1). The data were analysed using a linear mixed effects (LME) model, constructed using the 'lme4' package, version 1.1-21 [25] in R 3.6.0. The model predicted the transformed response variable from the offset value. The model also included a random effect for users which allowed each user to have their own y-intercept within the model. The R code for the model is:

$$model =$$
$$lmer(response\_minus\_interval \sim offset + (1|user), data = s)$$

Model predictors with t-values greater than 2 are interpreted as statistically significant (see Table III).

TABLE III: LME coefficients. Predictors with $|t-values|> 2$ are statistically significant.

Predictor	b	SE	t-value
Intercept	0.547	0.121	4.517
Offset value	0.176	0.058	3.027

The value of the intercept indicates that participants tended to judge the intervals to be larger than they were. The slope for the offset value was statistically significant indicating that users were utilising the location information provided by the VR system when judging the size of intervals, Fig. 11. Due to the overall bias to judge intervals as being larger than they were, negative offsets resulted in judgements that were, on average, closer to the correct interval. This is an important finding as conventional ear training cannot provide these spatial auditory cues which can be beneficial when undergoing interval training.

VI. Limitations

While the new system implementation has several advantages over the conventional scenarios where consumer APIs and proprietary game engines are used, there are still some limitations that will be addressed in the future. Some of the main shortcomings are the visual appearance of the system and user's relative position, combined with the limited field of view. A few users reported that they found the piano keys

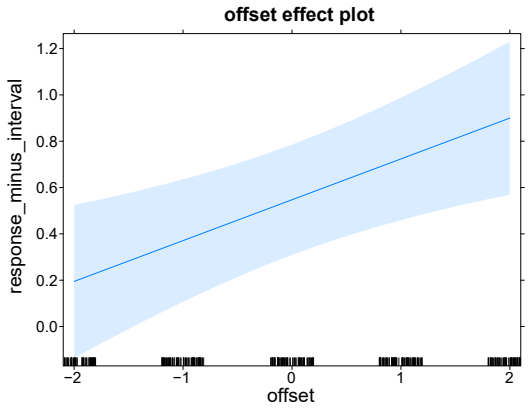


Fig. 11: The effects plot of the model. The shading around the plotted lines indicate the variability in the effect.

too big and wanted to move away from them. However, the reason for such layout was the maximisation of the angle between each key which is a paramount for key position localisation and thus tone identification and interval recognition. Another element that could be considered as a limitation is the number of intervals used in the study. The more experienced participants realised there were only two intervals repeating and found that irritable and boring. However, this was the design decision to minimise the test and VR exposure time. Finally, there was an issue with the user orientation in the system which caused the slight angular drift towards one side during the experiment. Although this did not affect the systems functionality, it could have had a negative effect on user experience.

The use of within-subject experimental procedures was crucial for learning more about human capabilities with the VR tool. By manipulating the VR experience of users, we were able to assess the impact of the manipulated variables on human performance with the VR tasks. However, such experiments are designed to answer specific questions and their ability to explore the differences between users in how they respond the within-subject experimental manipulation is very limited and can require samples too large to be feasible. For instance, a power simulation conducted on our second study indicated that we would need over 100 subjects in order to achieve 80% power to find an interaction between our offset manipulation and individual difference variables such as amount of musical training.

VII. Conclusions and Future Work

Pitch recognition is of paramount importance in music education. It helps musicians in identification of intervals, chord qualities, rhythmic patterns, and audition harmonic and melodic phrases. In addition, it sharpens a musician's ears for studying, understanding, performing and creating music. Finally, it intensifies the pleasure of music listening for both musicians and non-musicians. In this work we designed a novel VR ear training system. To the best of authors' knowledge, this is the first such system to explore multimodal cue integration by using VR technology and spatial audio. Initially, the system has been developed in Unity game engine and,

while working well, it was significantly hardware dependent, requiring a powerful machine and/or VR headset simply to satisfy arbitrary runtime requirements. The second implementation of the system uses OpenGL and a vendor neutral API - OpenHMD, that interfaces with the VR hardware. Not only does this open-source API affects the portability, but it likewise improves the performance and maintainability of the system.

The result from the user study with 46 participants conducted in this work show high levels of acceptance and immersion of the system, as well as a very positive effect on the user. The usability and user experience scores across the two experiments are reasonably consistent with both system implementations. This is a particularly valuable finding, demonstrating the rationale for considering using a lower level approach in developing VR systems with open-source tools whenever possible. This approach provides much higher portability and accessibility of the system, while also enhancing performance.

In the first experiment, music education level was found as the only independent variable to have a significant effect on the user performance, i.e. score. This was expected as people that are trained and/or educated musicians had probably undergone such training(s). This further informed the decision on designing Experiment 2. From the first experiment, there was no strong evidence for the effect of stereo panning on interval recognition. As it was a between participant experiment, it would require more participants, a longer training period, or repetitive testing (longitudinal study) to evidence a difference. In addition, the default Unity spatial sound capabilities are limited, providing only a stereo panned audio, instead of a true 360 audio simulation.

The second experiment which used a within participant design showed that the spatial cues provided by the VR application gave users significantly more information for recognising the musical intervals. This spatial information cannot be obtained in non-VR computer applications and therefore represents a real advantage over such applications. However, at this point there is not enough evidence on how to harness this effect to enhance the performance of interval recognition. Therefore, we believe that this work makes a solid body of contribution to knowledge in the area and opens new questions on how this or other perceptual cues can be utilised in this or similar applications.

In the future, we will improve the limitations mentioned in Section VI, add the remaining intervals and include other music instruments. We will also allow the users to select the way the intervals are played: melodic (ascending or descending), or harmonic. Other desirable extensions would be the addition of different training elements, such as pitch matching and rhythm exercises. We will discuss the system design further with musicians, music teachers and students, and try to identify other relevant key performance indicators that could be auto-assessed by the system. Once these improvements are made, we would like to conduct a larger user study with several user groups, including musicians, music students and 'music-naïve' users. Finally, we will propose incorporating it into the music school(s) curriculum.

ACKNOWLEDGEMENTS

The authors would like to thank all the participants who volunteered in the user study. This research was partially supported by the NVIDIA Corporation with the donation of the Titan Xp GPU.

REFERENCES

[1] R. H. Woody, "Playing by ear: Foundation or frill?" *Music Educators Journal*, vol. 99, no. 2, pp. 82–88, 2012.
[2] C. Loh, "Mona listen: A web-based ear training module for musical pitch discrimination of melodic intervals," in *E-Learn: World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education*. Association for the Advancement of Computing in Education (AACE), 2004, pp. 2026–2032.
[3] F. T. Hofstetter, "Guido: An interactive computer-based system for improvement of instruction and research in ear-training," *Journal of Computer-Based Instruction*, vol. 1, no. 4, pp. 100–106, 1975.
[4] "Earmaster 7," https://www.earmaster.com/, 2019, accessed: 2019-03-07.
[5] J. Miller, "Divided attention: Evidence for coactivation with redundant signals," *Cognitive psychology*, vol. 14, no. 2, pp. 247–279, 1982.
[6] "Mit music technology lab," https://musictech.mit.edu.
[7] "Harmonix music vr," Harmonics, http://www.harmonixmusic.com/games/harmonix-music-vr/, 2017, accessed: 2020-03-02.
[8] S. Serafin, A. Adjorlu, N. Nilsson, L. Thomsen, and R. Nordahl, "Considerations on the use of virtual and augmented reality technologies in music education," in *2017 IEEE Virtual Reality Workshop on K-12 Embodied Learning through Virtual & Augmented Reality (KELVAR)*. IEEE, 2017, pp. 1–4.
[9] M. R. Mine, "Virtual environment interaction techniques," *UNC Chapel Hill CS Dept*, 1995.
[10] D. A. Bowman and C. A. Wingrave, "Design and evaluation of menu systems for immersive virtual environments," in *Virtual Reality, 2001. Proceedings. IEEE*. IEEE, 2001, pp. 149–156.
[11] E. Selmanović, S. Rizvic, C. Harvey, D. Boskovic, V. Hulusic, M. Chahin, and S. Sljivo, "Improving accessibility to intangible cultural heritage preservation using virtual reality," *Journal on Computing and Cultural Heritage (JOCCH)*, vol. 13, no. 2, pp. 1–19, 2020.
[12] S. Serafin, M. Geronazzo, C. Erkut, N. C. Nilsson, and R. Nordahl, "Sonic interactions in virtual reality: state of the art, current challenges, and future directions," *IEEE computer graphics and applications*, vol. 38, no. 2, pp. 31–43, 2018.
[13] S. Malpica, A. Serrano, M. Allue, M. Bedia, and B. Masia, "Crossmodal perception in virtual reality," *Multimedia Tools and Applications*, vol. 79, no. 5, pp. 3311–3331, 2020.
[14] S. Yong and H.-C. Wang, "Using spatialized audio to improve human spatial knowledge acquisition in virtual reality," in *Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion*, 2018, pp. 1–2.
[15] E. R. Hoeg, L. J. Gerry, L. Thomsen, N. C. Nilsson, and S. Serafin, "Binaural sound reduces reaction time in a virtual reality search task," in *2017 IEEE 3rd VR workshop on sonic interactions for virtual environments (SIVE)*. IEEE, 2017, pp. 1–4.
[16] B. Benward and J. T. Kolosick, *Ear training: a technique for listening*. WCB/McGraw-Hill, 1996, vol. 1.
[17] "Openxr overview - the khronos group inc," https://www.khronos.org/openxr/, 2019, accessed: 2019-11-07.
[18] "Cross-process sharing and direct mode with vulkan," https://www.khronos.org/assets/uploads/developers/library/2019-vulkanised/05_Cross-Process-Sharing-And-Direct-Mode-On-Vulkan-May19.pdf, 2019, accessed: 2019-11-07.
[19] M. Smith, A. Maiti, A. D. Maxwell, and A. A. Kist, "Using unity 3d as the augmented reality framework for remote access laboratories," in *International Conference on Remote Engineering and Virtual Instrumentation*. Springer, 2018, pp. 581–590.
[20] A. Jangda, B. Powers, E. D. Berger, and A. Guha, "Not so fast: analyzing the performance of webassembly vs. native code," in *2019 {USENIX} Annual Technical Conference ({USENIX}{ATC} 19)*, 2019, pp. 107–120.
[21] D. D. Hodson and J. Millar, "Application of ecs game patterns in military simulators," in *Proceedings of the International Conference on Scientific Computing (CSC)*. The Steering Committee of The World Congress in Computer Science, Computer ..., 2018, pp. 14–17.
[22] J. Brooke *et al.*, "Sus-a quick and dirty usability scale," *Usability evaluation in industry*, vol. 189, no. 194, pp. 4–7, 1996.

[23] W. IJsselsteijn, Y. De Kort, and K. Poels, "The game experience questionnaire," *Eindhoven: Technische Universiteit Eindhoven*, 2013.
[24] A. Bangor, P. Kortum, and J. Miller, "Determining what individual sus scores mean: Adding an adjective rating scale," *Journal of usability studies*, vol. 4, no. 3, pp. 114–123, 2009.
[25] D. Bates, M. Maechler, B. Bolker, S. Walker, R. Christensen, H. Singmann *et al.*, "lme4: linear mixed-effects models using "eigen" and s4. 2017," 2019.

Karsten Pedersen is a lecturer in the Creative Technology Department at Bournemouth University, UK. He is currently working towards his PhD where he is looking at the facilitation of platform agnostic software development. This research was pioneered whilst working for 3 years in the games industry as a software programmer. His main areas of interest include systems level architecture and development, network and multiplayer and code correctness using C and C++. Contact him at pedersenk@bournemouth.ac.uk.

Vedad Hulusic is a Senior Lecturer at Bournemouth University, UK. He has a PhD in Engineering from the University of Warwick, UK and a first degree in Computer Science from the University of Sarajevo, Bosnia and Herzegovina. He has a long-standing interest in serious games, cultural heritage, VR/AR technologies, assistive technology, HDR imaging, image and video quality assessment, computer graphics, and cross-modal interaction in which he has been a published author. He is an IEEE member and a Senior Fellow of the Higher Education Academy. Contact him at vhulusic@bournemouth.ac.uk.

Panos Amelidis is currently a Lecturer in Music and Audio Technology at Bournemouth University, UK. He received the PhD degree from the Music Technology and Innovation Research Centre, De Montfort University, UK. His research focuses on music, and sound-art as a vehicle for cultural heritage awareness, construction of narrative in music, analysis of audio-only games and virtual reality systems for music education. Contact him at pamelidis@bournemouth.ac.uk.

Tim Slattery is a Principal Academic at Bournemouth University, UK. He earned his PhD in Cognitive Psychology from the University of Massachusetts, Amherst. He is interested in eye-movements, computer vision, cognition, and human-computer interaction. Before coming to Bournemouth University in 2015 he was an Assistant Professor at the University of South Alabama where he won an early career researcher award in 2014. Tim is a Psychonomic Society Member, Experimental Psychology Society Member and a Senior Fellow of the Higher Education Academy. Contact him at tslattery@bournemouth.ac.uk.