Virtual Reality Ear Training System: A study on Spatialised Audio in Interval Recognition

Connor Fletcher, Vedad Hulusic, Panos Amelidis
Creative Technology Department
Faculty of Science and Technology
Bournemouth University
United Kingdom
{cfletcher, vhulusic, pamelidis}@bournemouth.ac.uk

***Abstract*—Ear training is a vital element in music education, analogous to taking dictation in written language. It provides musicians with a crucial skill used to identify pitches, melodies, chords and rhythms. Traditionally, the training is conducted by a tutor using a musical instrument, typically a piano. However, with new technologies emerging, several computer applications to facilitate this aspect of music education have been developed. Nevertheless, none of them utilised the VR technology, that proved to be successful in various scenarios, including educational systems, simulations, etc. In this work, we designed and developed a virtual reality ear training system for interval recognition and investigated its usability and user experience and the effect of spatialised audio in a 3D virtual environment on user performance. The results showed that the system has been successfully designed and provides users with a great experience when using it.**

***Index Terms*—virtual reality, ear training, spatial audio, training, usability, user experience**

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 musicians ears for the study, comprehension, performance, and creation of music [2].

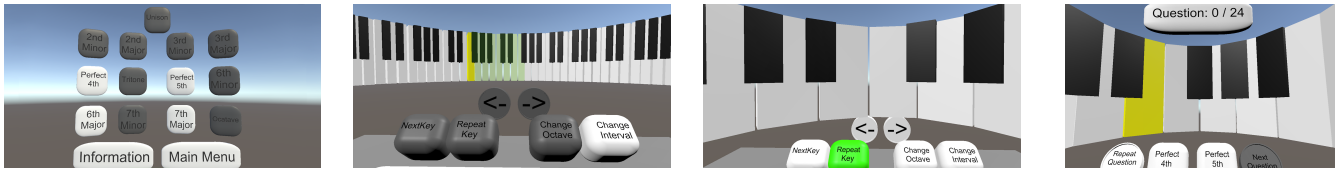
Ear training can take many forms but usually it starts by learning to recognise various melodic passages. In order for a musician to determine the notes in a melody, they must have the ability to distinguish and recognise musical intervals. Improving pitch recognition skill by way of ear training provides 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. However, nowadays, there is a variety of free and commercial software, made available as standalone, 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].

While Virtual Reality (VR) 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 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. For example, *Drumhead* [4] introduces aspiring musicians to drumming, using a variety of popular songs, in an interactive way and in VR. Here the user can see incoming notes as cylindrical shapes guiding where and when he needs to hit which part of the drums while at the same time the song is playing back along. *VR Sandbox* [5], is an exploration of different techniques and interactions with music composition in VR. *Harmonix Music VR* [6] by Harmonix allows the user to enjoy listening to music in a VR environment.

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 [7].

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 [8] either as heads-up display (HUD) or as floating menus (in 3D space) [9]. Alternatively, more natural 3D, physical paradigms could be used in forms of spin/ring menus [10]. 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], [12].

Main contribution of this work is the novel VR system for ear training. One of the main components of the system is the spatial audio delivered in the Virtual Environment (VE), that might help trainees when learning music intervals through associating physical locations to notes heard. The system has been validated with 27 participants, showing very high acceptability rate and equally good user experience with the



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

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

system. Finally, differences in user performance when rating different intervals are identified.

II. Virtual Reality Ear Training Systems

The main aim of the presented system is to train non-musicians and musicians music intervals. 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.

A. Design Considerations

Traditionally, the process of ear training in music education institutions focuses on teaching students to identify the most basic elements of music (i.e. intervals, simple melodies, simple triads, scales, and simple rhythm). There is a variety of approaches to ear training but usually exercises include simple dictation [13]. When it comes to intervals identification, the instructor plays (on the piano) intervals for the students to identify in both ascending and descending order. The intervals involved are comprised of two notes played either subsequently (melodic) or simultaneously (harmonic) [2]. This can also include stacks of two or more notes in which case they would be known as chords. After hearing an interval two or three times, students are asked to write it on score paper or to tell what interval they heard.

In our pilot study, 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 are required to familiarise themselves with the system and four melodic intervals; two 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) [14], [15]. These intervals were selected to allow the user to train in variety of both consonances and dissonances.

The system is designed to be used in a sitting position. Selection and menu interaction are performed by using HTC Vive controller and a laser pointer. As spatial audio was a key component of the system, it was important to utilise the 3D space, intrinsically provided in 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. Therefore, for both training and testing, a piano keyboard with 88 or 13 keys (representing one octave) was displayed, see Fig. 1.

B. System Development

The system was 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 are 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 is controlled by the Steam Vive controller, only requiring a single (physical) controller. Selection is 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.

III. User Study

The main aims of this study were to investigate the usability and user experience (UX) of the system, as well as its efficiency and the contribution of spatial audio on interval recognition. 27 participants volunteered for the study (M=23, F=3) with an average age of 26.46. One participant did not want to disclose their gender identity and age. 12 participants were exposed to non-panned mono sound (’Mono’ user group), while the other 15 were trained and tested in the stereo panned audio condition (’SP’ group). The system was run on a VR-Ready MSI Stealth Pro GS73 VR laptop. The visual stimuli were displayed on a 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. Participants were

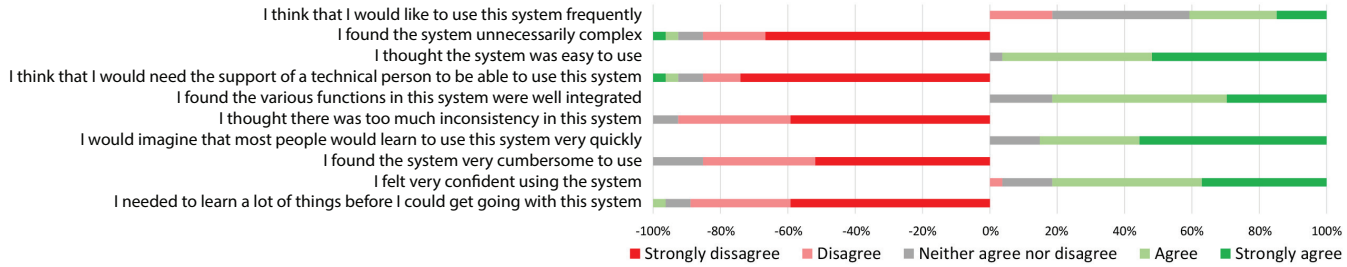


Fig. 2: Questions from the usability questionnaire (SUS) and the distribution of the user responses.

given detailed instructions on how to use the system, followed by a demo session, which used the same interface and both modes: training and test, but different intervals from those used in the main study.

In the main study, the participants were first trained on four music intervals: perfect 4th, perfect 5th, major 6th and major 7th (Fig. 1c). 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 (Fig. 1a) and the octave from the whole 88-key piano (Fig. 1b) in which they want to be trained. They 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 four intervals 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, Fig. 1d. Once selected they had to press the Next button, which allowed them to take a short break whenever they needed it.

IV. Results

A. Usability Study

Usability testing was performed using the SUS scale [16]. The questions and the corresponding responses from the study are presented in Fig 2. Using the score calculation as suggested in [16], the overall SUS score for our system was found to be 80.74, which corresponds to 'acceptable', 'grade B' and 'good', see Fig. 3. This confirms that the system was well designed and accepted by the users, even though there are certain elements which could be further improved.

B. User Experience (UX)

In this study we used a subset of 17 questions from the Core module of the Game Experience Questionnaire (GEQ) [18], covering all seven UX components. The questions were evaluated on a 5-point Likert scale. The results for all the questions and components are presented in Table I.

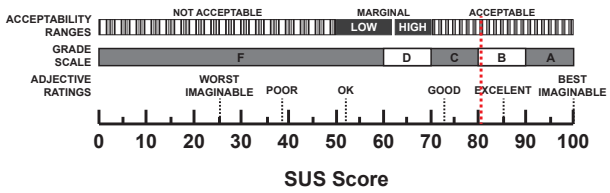


Fig. 3: Grade rankings for SUS scores as proposed by Bangor et al. [17]. Red dashed line represents the score for the evaluated ear training system.

In addition to the presented usability and UX question, two additional questions were added to the questionnaire for the SP audio user group, 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 as much to rely entirely on this cue (2.53).

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	Score (Q)	Score (C)
2	Competence	I felt skillful	3.11	3.15
11		I felt successful	3.19	
3	Immersion	I was interested in the task	4.44	3.88
7		It was aesthetically pleasing	3.33	
15		I found it impressive	3.85	
5	Flow	I was fully occupied with task	4.70	4.04
8		I forgot everything around me	3.37	
13	Tension/ Annoyance	I felt irritable	1.56	1.59
16		I felt frustrated	1.63	
14	Challenge	I felt challenged	3.85	3.31
17		I had to put a lot of effort into it	2.78	
6	Negative affect	I found it tiresome	1.59	1.48
10		I felt bored	1.37	
1	Positive affect	I felt content	3.81	4.10
4		I thought it was fun	4.41	
9		I felt good	4.00	
12		I enjoyed it	4.19	

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". 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. 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. Finally, two users wrote positive feedback, saying that everything was great.

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

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

C. The Training Performance and System Effectiveness

Since there were two user groups (Mono and SP) 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, SP) and the covariates were music education level, VR experience, number of repeats, training time and test time. 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 SP condition ($\mu_{SP} = 14.27$) 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, has not been found as significant ($p = .39$).

In addition, a factorial repeated-measures ANOVA was utilised to test for the effect of interval on user scores for both conditions (Mono and SP). The within-subject factor was interval, 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 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 ($p > .05$).

V. Conclusions and Future Work

In this work we designed a novel VR ear training system. To the best of authors' knowledge, this is the first system of this kind, using VR technology and spatial audio. The result from the user study conducted in this work show high levels of acceptance and immersion of the system, as well as a very positive effect on user. The only significant effect on the user performance, i.e. score, had the music education level. This was expected as people that are trained and/or educated musicians had probably undergone such training(s). From this study, we don't have strong evidence on the effect of stereo panning on interval recognition, as it would require longer training periods an repetitive testing (longitudinal study). In addition, the default Unity spatial sound capabilities are limited, providing only a stereo panned audio, instead of a true 360 audio simulation.

The main limitations of the system were the UI selection; visual appearance of the system; user relative position, combined with the limited field of view; and limited spatial audio functionality, providing only stereo panning and not full 360 audio. In the future, we will improve all these aspects, add remaining intervals and include other music instruments. We would also like to add both ascending and descending intervals, the option for the user to be trained in melodic and harmonic interval recognition, pitch matching and rhythm exercises. In addition, other sound effects, for feedback and user interaction will be added. 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] "Drumhead - practice your virtual rhythmic skills in this vr drumming game," MIT Music Technology Lab, https://musictech.mit.edu/blog/drumhead, 2018, accessed: 2019-03-07.

[5] "Vr sandbox - exploring different techniques and interactions with music composition in virtual reality," MIT Music Technology Lab, https://musictech.mit.edu/sites/default/files/documents/zahray_uap.pdf, 2017, accessed: 2019-03-07.

[6] "Harmonix music vr," MIT Music Technology Lab, http://www.harmonixmusic.com/games/harmonix-music-vr/, 2017, accessed: 2019-03-07.

[7] M. R. Mine, "Virtual environment interaction techniques," *UNC Chapel Hill CS Dept*, 1995.

[8] R. H. Jacoby and S. R. Ellis, "Using virtual menus in a virtual environment," in *Visual Data Interpretation*, vol. 1668. International Society for Optics and Photonics, 1992, pp. 39–49.

[9] 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.

[10] D. Gerber and D. Bechmann, "The spin menu: A menu system for virtual environments," in *Virtual Reality, 2005. Proceedings. VR 2005. IEEE*. IEEE, 2005, pp. 271–272.

[11] A. R. Galloway, *Gaming: Essays on algorithmic culture*. U of Minnesota Press, 2006, vol. 18.

[12] E. Selmanovic, S. Rizvic, C. Harvey, D. Boskovic, V. Hulusic, M. Chahin, and S. Sljivo, "Vr video storytelling for intangible cultural heritage preservation." 2018.

[13] B. Benward and J. T. Kolosick, *Ear training: a technique for listening*. WCB/McGraw-Hill, 1996, vol. 1.

[14] C. S. Myers, "Theories of consonance and dissonance," *Brit. J. Psychol*, vol. 1, no. 1905, pp. 315–316, 1905.

[15] R. Kamien and A. Kamien, *Music: an appreciation*. McGraw-Hill New York, NY, 1988.

[16] J. Brooke *et al.*, "Sus-a quick and dirty usability scale," *Usability evaluation in industry*, vol. 189, no. 194, pp. 4–7, 1996.

[17] 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.

[18] W. IJsselsteijn, Y. De Kort, and K. Poels, "The game experience questionnaire," *Eindhoven: Technische Universiteit Eindhoven*, 2013.