• Article •

Training birdsong recognition using virtual reality

María José ARIAS [1]， Gustavo CORRALES [1]， Carlos ARCE-LOPERA [1*]

1. *Department of Technology and Information, Universidad Icesi, Cali,* 760046, *Colombia*

*** Corresponding author，** caarce@icesi.edu.co

Received: 28 March 2021 **Accepted: xxx**

Abstract Background Commonly, Species Monitoring is performed in mega-biodiverse environments by using bioacoustics methodologies where the species are more likely to be heard than seen. Furthermore, since bird vocalizations are reasonable estimators of biodiversity, their monitoring is of great importance in the formulation of conservation policies. However, birdsong recognition is an arduous task that requires dedicated training to achieve mastery; this training is costly in terms of time and money due to the lack of accessibility of relevant information in field trips or even on specialized databases. Immersive technology based on virtual reality (VR) and spatial audio may improve Species Monitoring by enhancing information accessibility, interaction, and user engagement. **Methods** This study used spatial audio, a Bluetooth controller, and a Head-mounted Display (HMD) to conduct an immersive training experience in VR. Participants moved inside a virtual world using a Bluetooth controller while their task was to recognize targeted birdsongs. We measured the accuracy of the recognition and the user engagement according to the User Engagement Scale. **Results** Experimental results revealed significantly higher engagement and accuracy for participants in the VR-based training system when compared to a traditional computer-based training system. All four dimensions of the user engagement scale received high ratings by the participants suggesting that VR-based training provides a motivating and attractive environment to learn demanding tasks through appropriate design, exploiting the sensory system and the virtual reality interactivity. **Conclusions** The accuracy and engagement of a VR-based training system were significantly highly rated when tested against traditional training. Future research will focus on developing a variety of realistic ecosystems and their associated birds to increase the information of newer bird species in the training system. Finally, the proposed VR-based training system must be tested with additional participants and for a greater duration to measure information recall and recognition mastery among users.

Keywords Human computer interaction; Virtual environment; Birdsong; Audio training; User Engagement

1 Introduction

Colombia ranks among the twelve most megadiverse nations and as second on global diversity [1]. Also, Colombia is the most avian-diverse country thanks to its near two thousand different species of birds [2]. However, the list of endangered species grows every year due to the reduction of natural forest lost to illegal deforestation driven by rapid urbanization, the rise of extractive industries, and farming [3]. Since the existence of birds is an optimal ecological indicator, their identification is of great importance for diversity. Presently, bioacoustics is the most used methodology to monitor birds because most of them are easier to hear than to see in the field [4].

Very few professionals develop the ability to recognize bird sounds for several species of birds, adding up to the difficulties of monitoring avian activity through sounds. Furthermore, teaching volunteers implies additional costs and long learning processes that may take more than a year. Indeed, learning to identify birds to tag them is no easy task since it requires constant practice [4]. There are two traditional methods to acquire the required knowledge, one real and one virtual. The first method consists of guided sighting hikes to perform visual and acoustic observations of the birds. This traditional method has to be supported by expert guides in the field, which represent additional costs. The second training method comprises active learning through the remote study of photographs and sound databases to identify bird species. Unfortunately, both methods are limited in accessibility and interactivity [5]. The high associated costs reduce accessibility. For example, field trips are costly and depend on the availability of expert guides and the appropriate conditions of the environment to visit, which can result in time and money waste if not handled appropriately. Also, access to specialized birdsong databases is rather restricted and expensive. Moreover, both training methods depend strongly on expert guides that are not often experienced as educators and fall short when sharing knowledge or motivating trainees; this translates into reduced interactivity.

These challenges motivated the idea to create an immersive birdsong recognition training process. The aim was to create a systematic interactive training that would allow individuals interested in birds to learn and identify them in an interactive virtual reality environment. The system must focus on improving information accessibility and interactivity using inexpensive equipment that can be purchased and used easily by anyone.

1.1 Immersive technology for training

The notion of immersive technology refers to any technology that blurs the boundary between the real and virtual world by immersing the user fully or partially in a virtual experience [6]. In particular, Virtual Reality (VR) is based on a three-dimensional environment generated by a computer that completely simulates a new version of the physical world. Users who explore this type of environment interact with the virtual world as if they were exploring it in reality using different display and input devices. Therefore, virtual environments can provide a rich, interactive, engaging context that can support experiential learning. Moreover, VR's ability to engage users by employing enjoyable multimodal experiences that immerse them in focused tasks can lead to higher mastery and retention of knowledge [7]. Although virtual reality has been around for over 40 years, just recently it has become available for consumer use. Various devices to display immersive technology have been introduced using expensive Head-mounted Displays (HMD), such as the HTC or the Oculus, or economic optical adapters that use available mobile devices to display the related images. The most economically accessible and adaptable way to use VR technology is by using cardboard-based adapters, such as the Google cardboard, which uses cheap DIY materials in its build. By using this type of optical adapters, the training system will be more accessible for general use.

However, previous research on VR training has reported negative responses associated with immersive technology use. For example, researchers found that some HMDs users experienced motion sickness, physical discomfort, and/or cognitive overload [8,9]. Therefore, the appropriate design of the immersive experiences must take into account the limits of perceptual comfort and avoid overstimulating the users. Also, researchers have suggested that for training purposes, immersive experiences should incorporate different stages of learning tasks [10]. This way trainees will seek to accomplish increasingly difficult tasks experiencing a sense of self-progress that may enhance the learning process and user engagement. On the other hand, positive outcomes of the usage of immersive technology for training are the improvement of effectiveness and task performance, the increase of positive attitude towards the learning task, the reduction of task completion time, and an increased in the accuracy rate [9–13]. In spite that the widespread use of immersive technology is at an early stage of development, designing training tasks that take advantage of the experiential learning potential of VR can achieve systematically positive results when traditional training is highly inaccessible.

1.2 Spatial audio for HMD-based VR

Spatial audio refers to the 360-degree sound reproduction around the listener. Spatial audio resembles what we hear in real life and has the potential to enhance the immersion of the listener into virtual experiences. Specifically, for HMD-based VR experiences, spatial audio enhances the interaction by using a dynamic binaural audio synthesis that is based on head-related transfer functions (HRTFs), virtual loudspeakers, and headset orientation data [14]. However, HRTFs are highly personalized, needing individualized calibrations using expensive equipment. To avoid this limitation, generic HRTFs are usually employed to render good-enough spatialized sounds in VR [15]. Audio synthesis can provide realistic immersive experiences with standard headphones, a mobile phone, and a cardboard headset. This way, the experience is enhanced because the virtual world images and other visual stimuli can be associated with the location and distance of the corresponding sound source.

2 Methods

We conducted a between-subjects study that aimed at comparing the engagement and accuracy for birdsong recognition of a traditional training system against a training system based on VR. Participants were assigned to one of two groups, both groups had the same learning goals. The learning steps were as follows: first, participants heard a birdsong associated with a particular bird; then, they heard another audio file with different sound sources and tried to identify the time when the targeted bird sang. The test audio file was labeled by an expert to provide a ground truth where all birds that sang were correctly recognized. Finally, all participants responded to the User Engagement Scale test to assess their perceived engagement during the training activity. Participants were between 18 and 54 years old and had normal visual and auditory perception. The University Committee on Human Research approved the study while the subjects voluntarily agreed to participate by given verbal informed consent after the experimenter explained the project in general terms.

2.1 Equipment

Participants wore an economical optical adapter (VR box virtual reality glasses) where a smartphone was introduced as their HMD. Also, they used a 7.1 headphone (VanTop technology and innovation co. Ltd, Shenzhen, Guangdong) to reproduce the spatial audio. Participants interacted with the virtual world using a standard Bluetooth controller connected to the smartphone (Figure 1). The standard Bluetooth controller had four buttons and a multidirectional stick that was used to navigate the VR environment. The four buttons were programmed as triggers for bird recognition in the VR-based environment. The developed VR mobile application was designed for the operating system Android. The mobile phone used in the training test was a standard Android phone (A51, Samsung, South Korea), with 1080 x 2400 (FHD+) screen resolution, 4 GB RAM, and with an Octa-Core (2.3GHz) processor.

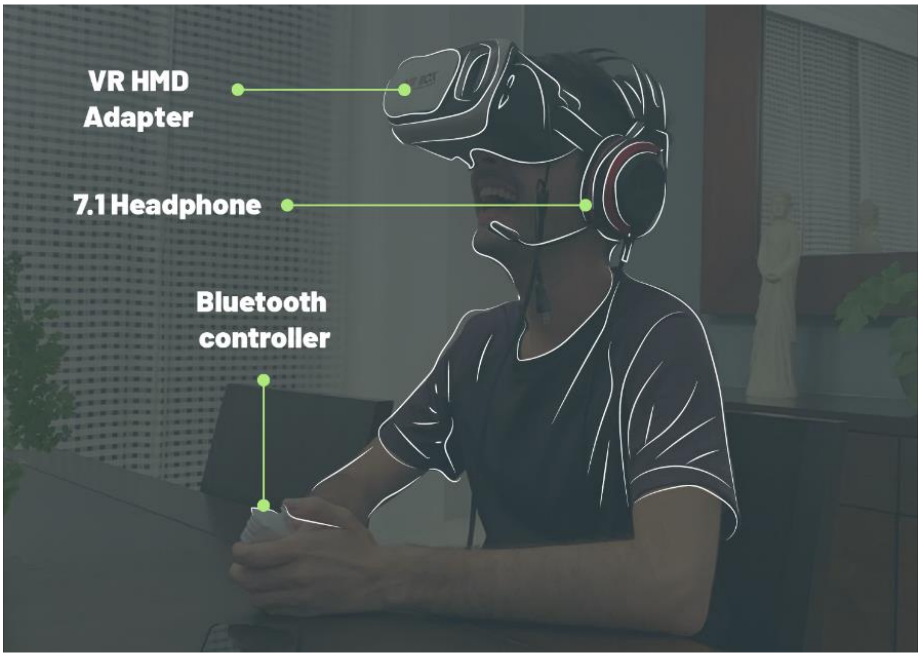

Figure 1 User interacting with the VR Experience.

2.2 Training task

The training task started by presenting information corresponding to a particular bird, the target bird. This information included the scientific name, the common name, an image of the bird, and audio samples of the specific bird singing. This learning phase was identical for both experimental groups, the traditional training group, and the VR-based immersive training group. Participants could remain at this learning stage until they were confident in their ability to identify the birdsong. Once they decided that they could recognize the birdsong, participants entered the next phase of the task, the test. For both conditions, the duration of this phase was identical (278 seconds) corresponding to the reproduction of an audio file with the song of the target bird and other birds several times. By controlling the duration of the test, the influence of time on the task was avoided. For the traditional training, a simple computer interface was designed and implemented in Java to allow participants to play an audio file and recognize the onset time when the birdsong was reproduced. This labeling decision was performed using the space bar of the computer. The audio reproduced was a recording of a single microphone, the most common condition in traditional training.

In contrast, for the VR immersive training, participants listened to a 3D sound based on a six-microphone array recording of the same environment. Moreover, subjects visualized a virtual forest that stretched in all directions (Figure 2). The task was to move into the virtual world and recognize the time when a targeted bird sang. This temporal labeling was performed using the Bluetooth controller. The graphics and tasks were designed and implemented in Unity. Each participant was able to look around the environment by moving the head and was allowed to move in any direction to follow the sound sources. The 3D spatial sound was reproduced to permit the correct identification of the source of the sounds in terms of location and distance. The VR participants were free to move around the VR environment active area, set to 60 meters by 60 meters. Participants were told to pay attention to birdsongs and move towards the sound sources. However, there was no guarantee that participants correctly followed this instruction as they could move to any location in the VR environment without limitations. Therefore, there was the possibility of going to locations where the target birdsongs were harder to identify due to low loudness.

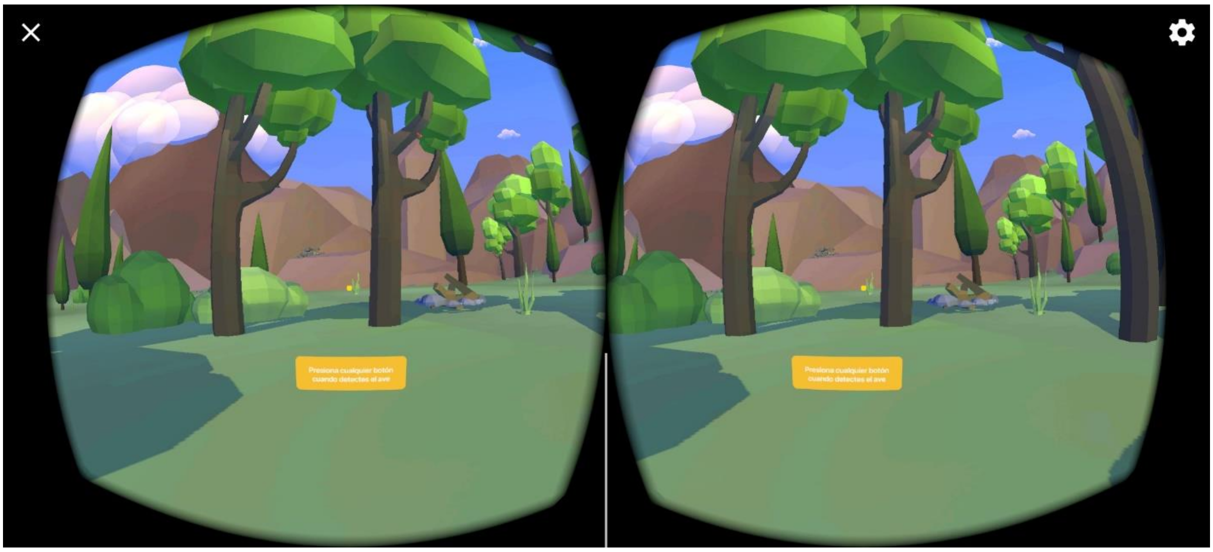
Presiona cualquier botón cuando detectes el ave
Presiona cualquier botón cuando detectes el ave

Figure 2 View of virtual environment.

2.3 Data collection

Their age and the number of times participants listened to the audio of the birdsong in the learning phase were recorded for each participant. In the test phase, the accuracy of each participant was registered by comparing their onset time identification with the ground truth provided by the expert. The accuracy was calculated using a grading system based on how long it took the participant to recognize the audio. If the participant recognized the bird while the song of the bird played, the participant scored 100, the maximum score. If the recognition was in the next two seconds after the end of the bird song, the participant also scored 100. If the selection was 3 seconds off, the score became 85. At 4 seconds off, the score was 80. At 5 seconds off, the score was 75. Finally, at 6 and 7 seconds off, the score was 70 and 65, respectively. At the end of the test trial, the accumulated score was divided by the number of times the participant taught the bird was singing. For example, if a participant recognized the birdsong five times in the test trial but got it right only three times (e.g., scored 75 points for each good recognition event), the overall score was (75x3)/5=45 points.

After each test trial, all participants responded to the User Engagement Scale - Short Form (UES-SF) [16] to assess their perceived engagement with the respective training activity. The UES-SF divides "engagement" into four dimensions; focus attention, perceived usability, aesthetic appeal, and endurability. The dimensions are measured using three questions rated on a five-point Likert scale: (1) Strongly disagree; (2) Disagree; (3) Neither agree nor disagree; (4) Agree; (5) Strongly agree. The focus attention dimension is related to the feeling of being lost in the experience and losing track of time. The perceived usability dimension tested the perceived control of the participants in the interaction and their levels of frustration. The aesthetic dimension accounted for the perceived appeal of the interface concerning the senses. Finally, the endurability dimension measured the reward and satisfaction perception of the participant. Table 1 describes the questions corresponding to each dimension.

Table 1 UES-SF questions used to measure perceived engagement

UES-SF Dimension	*Question*
Focus Attention (FA)	FA1: I lost myself in this experience FA2: I was absorbed in this experience FA3: The time I spent using the application just slipped away
Perceived Usability (PU)	PU1: I felt frustrated while using this application PU2: I found this Application confusing to use PU3: Using this Application was taxing
Aesthetic Appeal (AA)	AA1: This Application was attractive AA2: This Application was aesthetically appealing AA3: This Application appealed to my senses
Endurability (E)	E1: Using the Application was worthwhile E2: My experience was rewarding E3: I felt interested in this experience

2.4 Statistical analysis

For each measured parameter, descriptive statistics were calculated testing significant differences through a two-sample t-test at a confidence level of 95%. There is a controversy when deciding which type of statistical procedure is appropriate for analyzing 5-point Likert items, such as the UES-SF [17–19]. For instance, the Wilcoxon Signed-Rank test is recommended as it is a non-parametric procedure used as an alternative to the t-test when the distribution of the difference between two samples' means cannot be assumed to be normally distributed. However, the recommendations to choose a non-parametric procedure over the t-test, particularly with small sample sizes and Likert scale data, appear to be groundless, even when the t-test assumptions are violated [18]. Nevertheless, running both statistical analyses (t-test and Wilcoxon Signed-Rank test) produced similar results. We decided to present the t-test values as they are the most commonly used when reporting differences when using the user engagement scale [20,21]. Additionally, the Pearson correlation coefficient for each parameter was calculated to measure the linear relationship between all variables.

3 Results

Eighteen participants completed the study. Eight of them were female, and the average and standard deviation of the participants' age was 28.0 ± 10.8 years old.

Summary results describing the age, the number of repetitions in the learning phase, and the accuracy points for all participants depending on the group, traditional or VR-based training, are presented in Tables 2 and 3, respectively.

When comparing the age for the nine participants in the traditional training group (M=25.9 years old; SD=15.9 years old) against the age of the nine participants in the VR-based training group (M=30.2 years old; SD=11.6 years old), no significant differences were shown (t(8)=-0.79, p=0.45 >0.1). Similar results were obtained for the number of repetitions in the learning phase, where both training groups average was not significantly different, t(8)=0.2, p=0.84 >0.1. However, a significant difference t(8)=-3.9, p<0.01 was shown for the average accuracy points between groups. The accuracy of the participants that used the VR-based training was significantly higher (M=71.1 points; SD=13.4 points) than the accuracy of the participants in the traditional training (M=34.8 points; SD=10.3 points). Moreover, Pearson correlation analysis revealed no significant relation between the age of the participant and its accuracy (r(18)=0.001, p=0.99>0.1). Additionally, correlation analysis between the age of the participant and the number of repetitions in the learning phase was not significant (r(18)=0.16, p=0.51>0.1). Finally, no significant correlation (r(18)=0.35, p=0.15>0.1) was shown between the number of repetitions in the learning phase and the accuracy points. Figure 2 shows the plot of the accuracy points in the vertical axis against the number of repetitions in the horizontal axis.

Table 2 Individual results for the traditional training group

User ID	Age	# of Repetitions in Learning Phase	Accuracy Points
User 2	18	2	26.7
User 4	18	3	30.0
User 7	24	1	40.0
User 10	24	2	25.4
User 11	22	7	75.0
User 12	22	1	35.0
User 16	18	5	26.0

User 17	44	3	25.0
User 18	43	2	30.0
Total	**25.9 (15.9)**	**2.9 (1.9)**	**34.8 (10.3)**

Total values are presented as mean (±SD).

Table 3 Individual results for the VR-based training group

User ID	Age	# of Repetitions in Learning Phase	Accuracy Points
User 1	21	6	87.7
User 3	23	2	77.8
User 5	45	1	40.0
User 6	21	2	70.4
User 8	24	2	65.0
User 9	27	3	76.0
User 13	25	1	67.8
User 14	54	3	77.0
User 15	32	4	78.0
Total	**30.2 (11.6)**	**2.6 (1.6)**	**71.1 (13.4)**

Total values are presented as mean (±SD).

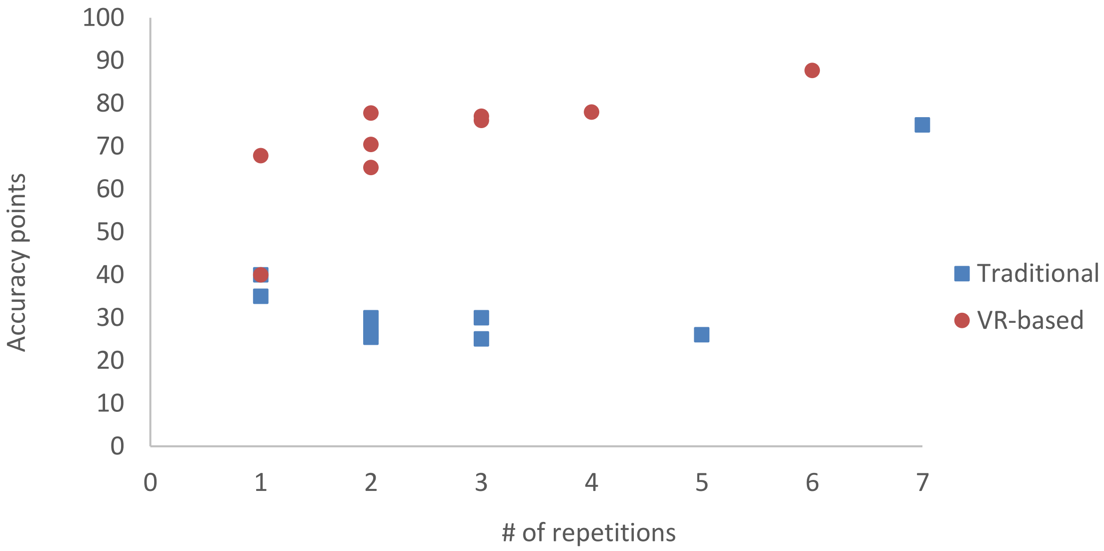

Figure 3 Accuracy points versus number of repetitions in the learning phase. Blue squares represent the participants' results for the traditional training while the red circles show the results of the participants in the VR-based training.

Table 4 UES-SF results for both trainings

UES-SF Dimension	*Question*	*Traditional Training*	*VR-based Training*
Focus Attention (FA)	FA1	2.8 (1.1)	4.4 (0.5)
	FA2	2.6 (1.2)	4.7 (0.5)

	FA3	2.3 (1.5)	4.1 (0.8)
Perceived Usability (PU)	PU1	2.6 (1.4)	3.1 (0.6)
	PU2	3.6 (0.5)	3.7 (0.5)
	PU3	2.0 (1.3)	3.2 (0.7)
Aesthetic Appeal (AA)	AA1	3.3 (0.5)	4.6 (0.5)
	AA2	3.6 (0.9)	4.8 (0.4)
	AA3	3.0 (1.0)	3.9 (0.8)
Endurability (E)	E1	3.7 (0.5)	4.0 (0.7)
	E2	3.0 (0.9)	4.7 (0.5)
	E3	2.8 (1.5)	4.6 (0.5)

UES-SF Ratings: Values are presented as mean (±SD).

The UES-SF results revealed that the overall engagement in the VR-based training (M=4.1; SD=0.8) was rated significantly higher (t(107)=9.8;p<0.001) than the engagement in the traditional training (M=2.9; SD=1.1). For the focus attention dimension, participants rated significantly higher (t(26)=6.3;p<0.001) the VR-based experience (M=4.4; SD=0.6) compared with the traditional training (M=2.5; SD=1.2). Similarly, the perceived usability was significantly highly rated (t(26)=3.7; p<0.005) for the VR-based training (M=3.3; SD=0.6) compared with the traditional training (M=2.7; SD=1.1). The aesthetic appeal was rated significantly higher (t(26)=5.3; p<0.001) for the VR-based training (M=4.4; SD=0.7) when compared to the traditional training (M=3.3; SD=0.8). Finally, the endurability dimension was rated significantly higher (t(26)=5.0; p<0.001) for the VR-based training (M=4.4; SD=0.6) when compared to the traditional training (M=3.1; SD=1.1). Table 4 and Figure 4 show the results of the UES-SF discriminated by questions and training group.

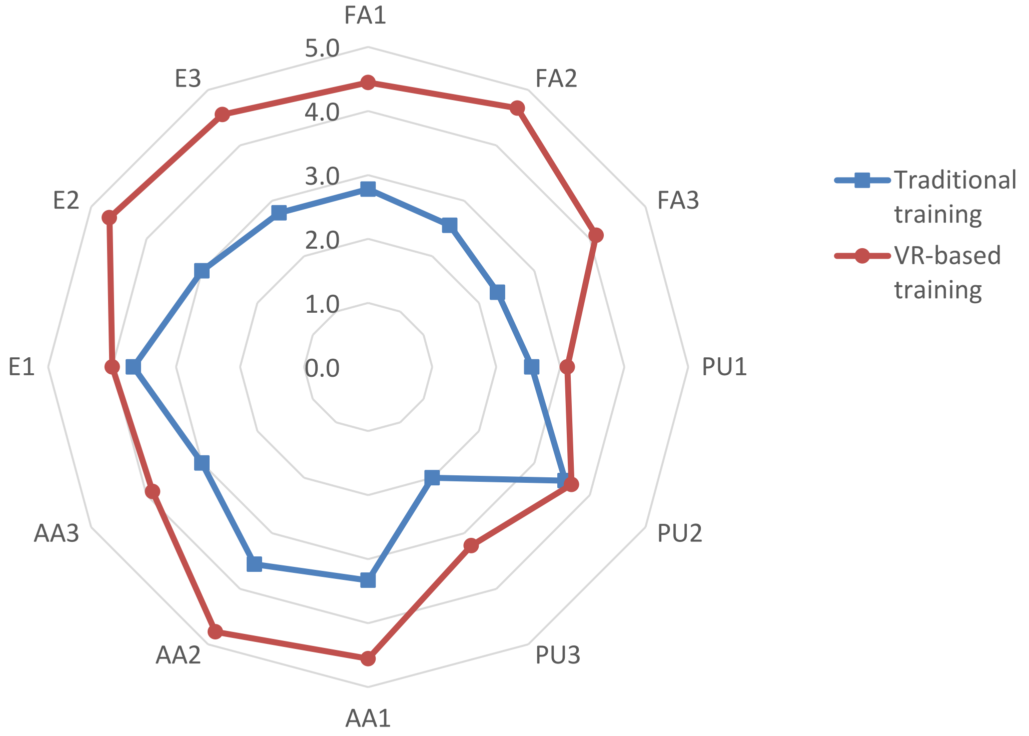

Figure 4 **Radial plot of the results of the UES-SF for both training groups. Red circles represent the results for the VR-based training while blue squares are the results for the traditional training.**

4 Discussion

A VR-based training system for the recognition of birdsongs was designed and implemented using easy-to-find and low-cost equipment. It was characterized mainly by the reproduction of spatial audio and the enhanced interactivity with the virtual world. To test the training, two groups of participants were recruited and assigned to either a traditional version of the training or to the VR-based training. Both types of training followed the same steps; first, a self-controlled learning phase, followed by an accuracy test. Then, the last step consisted of a user engagement survey based on the UES-SF. The age of participants was not significantly different between the two groups. Our experimental results revealed that age was not a determinant factor for accuracy in the training, e.g., older participants had similar accuracy points compared to young users. This finding was surprising for the authors because previous research has documented a degrading relationship between age and performance in VR environments [22,23]. However, it is possible that as designed as a short time experience, our training did not frustrate or fatigue users to the point of seeing any effect on physical endurance that may explain the decrease in performance associated with age. Nevertheless, we recognize the need to expand the number of subjects to include more participants (older and younger) to be able to generalize this result.

Concerning the first step of the interaction, the self-controlled learning stage, the number of repetitions that the participants decided to reproduce the audio was not correlated with the accuracy points obtained in the test phase. This surprising result also contradicts former research that showed that previous repetition in a recognition task positively impacts the subsequent recognition performance [24]. Particularly, Figure 3 shows two participants that can be considered as outliers of their respective groups that could follow that trend. For the traditional training group, user 11 repeated seven times the recording and scored 75 points for accuracy which is similar to the results of the VR-based training group. Similarly, belonging to the VR-based training group, user 5 only reproduces the learning audio once and scored only 40 points for accuracy. Both users were outliers in their respective groups because the other participant's accuracy was not dependent on the number of times they repeated the audio. This difference may arise due to the different motivations for participation. User 11 was intrinsically highly motivated to perform well at the task and enjoyed the learning phase, according to comments conveyed by the participant to the experimenter. On the contrary, user 5 revealed the lack of desire to achieve high scores preferring to enjoy the relaxing atmosphere of the VR-based training interaction. Those insights revealed the importance of designed experiences for already motivated users. Apart from those two outliers, the rest of the users had similar repetitions and accuracy performances within their respective groups.

The significant difference in the accuracy between both types of training may indicate that VR-based training is more effective than traditional training. Indeed, according to the UES-SF results, participants perceived the VR-based training as significantly more engaging than the traditional method in all four UES dimensions. Greater focus attention may lead to increased performance at recognition activities because participants avoid distractions and concentrate on the specific task [25,26]. The usability was perceived as significantly higher for the VR-based training when compared to the traditional computer-based training. This result validates our design of a VR-based experience that transforms boring traditional tasks, like practicing sound recognition, with an immersive and interactive experience where users have more control and less frustration. Also, the aesthetic appeal of the VR-based training was rated significantly higher. Spatial sound use in an immersive and interactive 3D environment appeals to the multimodal sensory system of participants [14]. Finally, the endurability and satisfaction perception in the VR-based system was significantly higher than the traditional training, possibly indicating that the system motivated users to interact and that the design of the task did not overstimulate them. However, it remains unclear if these results are dependent on the novelty perception of the participants for the VR-based system. Prolonged exposure to the VR-based training system may reveal if the novelty wears off and if the accuracy and engagement may be affected [27].

In conclusion, when our VR-based training system was tested against a traditional training system, the experimental results revealed higher user engagement and accuracy, validating the experience design. Future research will focus on developing a variety of realistic ecosystems with their associated birds to increase the information of newer bird species in the training system. Emphasis will be on reproducing accurately the spatialized sounds for which carefully calibrated and personalized HRTFs will be measured for each participant. Furthermore, a gamification module with cooperative and social interaction may enhance the experience and further motivate participants to continue with the training. Finally, the training system must be tested with additional participants and for a longer duration to measure information recall and recognition mastery among users.

References

1. Myers, N., Mittermeier, R. A., Mittermeier, C. G., da Fonseca, G. A. B. & Kent, J. Biodiversity hotspots for conservation priorities. *Nature* **403**, 853–858 (2000).

2. Baptiste, M. P. *et al.* Global Register of Introduced and Invasive Species - Colombia. doi:10.15468/yznr8v.

3. Krause, T. Reducing deforestation in Colombia while building peace and pursuing business as usual extractivism? *Journal of Political Ecology* **27**, (2020).

4. Kvsn, R. R., Montgomery, J., Garg, S. & Charleston, M. Bioacoustics Data Analysis – A Taxonomy, Survey and Open Challenges. *IEEE Access* **8**, 57684–57708 (2020).

5. Venier, L. *et al.* Comparison of semiautomated bird song recognition with manual detection of recorded bird song samples. *Avian Conservation and Ecology* **12**, (2017).

6. Suh, A. & Prophet, J. The state of immersive technology research: A literature analysis. *Computers in Human Behavior* **86**, 77–90 (2018).

7. Alrehaili, E. A. & Osman, H. A. A virtual reality role-playing serious game for experiential learning. *Interactive Learning Environments* **0**, 1–14 (2019).

8. Goh, D. H.-L., Lee, C. S. & Razikin, K. Interfaces for accessing location-based information on mobile devices: An empirical evaluation. *Journal of the Association for Information Science and Technology* **67**, 2882–2896 (2016).

9. Munafo, J., Diedrick, M. & Stoffregen, T. A. The virtual reality head-mounted display Oculus Rift induces motion sickness and is sexist in its effects. *Exp Brain Res* **235**, 889–901 (2017).

10. Ibáñez, M., Di-Serio, Á., Villarán-Molina, D. & Delgado-Kloos, C. Support for Augmented Reality Simulation Systems: The Effects of Scaffolding on Learning Outcomes and Behavior Patterns. *IEEE Transactions on Learning Technologies* **9**, 46–56 (2016).

11. Frank, J. A. & Kapila, V. Mixed-reality learning environments: Integrating mobile interfaces with laboratory test-beds. *Computers & Education* **110**, 88–104 (2017).

12. Loup-Escande, E. *et al.* Contributions of mixed reality in a calligraphy learning task: Effects of supplementary visual feedback and expertise on cognitive load, user experience and gestural performance. *Computers in Human Behavior* **75**, 42–49 (2017).

13. Ke, F., Lee, S. & Xu, X. Teaching training in a mixed-reality integrated learning environment. *Computers in Human Behavior* **62**, 212–220 (2016).

14. Hong, J. Y., He, J., Lam, B., Gupta, R. & Gan, W.-S. Spatial Audio for Soundscape Design: Recording and Reproduction. *Applied Sciences* **7**, 627 (2017).

15. Berger, C. C., Gonzalez-Franco, M., Tajadura-Jiménez, A., Florencio, D. & Zhang, Z. Generic HRTFs May be Good Enough in Virtual Reality. Improving Source Localization through Cross-Modal Plasticity. *Front. Neurosci.* **12**, (2018).

16. O'Brien, H. L., Cairns, P. & Hall, M. A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form. *International Journal of Human-Computer Studies* **112**, 28–39 (2018).

17. de Winter, J. F. C. & Dodou, D. Five-Point Likert Items: t test versus Mann-Whitney-Wilcoxon (*Addendum added October 2012*). *Practical Assessment, Research, and Evaluation* **15**, (2019).

18. Meek, G., Ozgur, C. & Dunning, K. Comparison of the t vs. Wilcoxon Signed-Rank Test for Likert Scale Data and Small Samples. *Journal of Modern Applied Statistical Methods* **6**, (2007).

19. Harpe, S. E. How to analyze Likert and other rating scale data. *Currents in Pharmacy Teaching and Learning* **7**, 836–850 (2015).

20. Nguyen, D. & Meixner, G. Gamified Augmented Reality Training for An Assembly Task: A Study About User Engagement. in *2019 Federated Conference on Computer Science and Information Systems (FedCSIS)* 901–904 (2019). doi:10.15439/2019F136.

21. Ruan, S. *et al.* QuizBot: A Dialogue-based Adaptive Learning System for Factual Knowledge. in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* 1–13 (ACM, 2019). doi:10.1145/3290605.3300587.

22. Coxon, M., Kelly, N. & Page, S. Individual differences in virtual reality: Are spatial presence and spatial ability linked? *Virtual Reality* **20**, 203–212 (2016).

23. Arino, J.-J., Juan, M.-C., Gil-Gómez, J.-A. & Mollá, R. A comparative study using an autostereoscopic display with augmented and virtual reality. *Behaviour & Information Technology* **33**, 646–655 (2014).

24. Kaplan, A. D. *et al.* The Effects of Virtual Reality, Augmented Reality, and Mixed Reality as Training Enhancement Methods: A Meta-Analysis. *Hum Factors* 0018720820904229 (2020) doi:10.1177/0018720820904229.

25. Huang, H.-M., Rauch, U. & Liaw, S.-S. Investigating learners' attitudes toward virtual reality learning environments: Based on a constructivist approach. *Computers & Education* **55**, 1171–1182 (2010).

26. Huang, T.-L. & Liao, S.-L. Creating e-shopping multisensory flow experience through augmented-reality interactive technology. *Internet Research* **27**, 449–475 (2017).

27. Rutter, C. E., Dahlquist, L. M. & Weiss, K. E. Sustained Efficacy of Virtual Reality Distraction. *The Journal of Pain* **10**, 391–397 (2009).