Auditory Feedback for Navigation with Echoes in Virtual Environments: Training Procedure and Orientation Strategies

Andreasen, Anastassia; Geronazzo, Michele; Nilsson, Niels Christian; Zovnercuka, Jelizaveta; Konovalovs, Kristians; Serafin, Stefania

Published in:
I E E E Transactions on Visualization and Computer Graphics

DOI (link to publication from Publisher):
10.1109/TVCG.2019.2898787

Publication date:
2019

Document Version
Accepted author manuscript, peer reviewed version

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2019.2898787, IEEE Transactions on Visualization and Computer Graphics

Auditory Feedback for Navigation with Echoes in Virtual Environments: Training Procedure and Orientation Strategies

Anastassia Andreasen, Michele Geronazzo, *Member*, *IEEE*,
Niels Christian Nilsson, Jelizaveta Zovnercuka, Kristian Konovalov, and Stefania Serafin

Fig. 1: The virtual environment used for training human echolocation resembled a dark virtual cave (left panel). Test participants performed a navigation task which consists in finding the exit of a tunnel to the opening of the cave (right panel, 1-4 are photographs in sequence of a trial) with different types of unimodal auditory or visual feedback. Real-time auralization was designed within Steam Audio engine and delivered through headphones. An Oculus Rift and Touch controller supported the navigation.

Abstract— Being able to hear objects in an environment, for example using echolocation, is a challenging task. The main goal of the current work is to use virtual environments (VEs) to train novice users to navigate using echolocation. Previous studies have shown that musicians are able to differentiate sound pulses from reflections. This paper presents design patterns for VE simulators for both training and testing procedures, while classifying users' navigation strategies in the VE. Moreover, the paper presents features that increase users' performance in VEs. We report the findings of two user studies: a pilot test that helped improve the sonic interaction design, and a primary study exposing participants to a spatial orientation task during four conditions which were early reflections (RF), late reverberation (RV), early reflections-reverberation (RR) and visual stimuli (V). The latter study allowed us to identify navigation strategies among the users. Some users (10/26) reported an ability to create spatial cognitive maps during the test with auditory echoes, which may explain why this group performed better than the remaining participants in the RR condition.

Index Terms—Human echolocation, navigation, spatial cognition, virtual reality, sonic interactions, spatial audio, binaural synthesis

1 Introduction

In the seminal 1974 paper 'What Is it Like to Be a Bat?' [32], Nagel used the alien and ineffable mental life of bats to question common held views about what it means to be conscious. While we may never fully understand the subjective experience of bats, virtual reality (VR) might be able to provide users with a glimpse of what it is like to be a bat. Moreover, a particular feature of bats' perceptual system has proven to have many applications; namely *echolocation*. Echolocation has for example been applied for military purposes (e.g., radar for boats and planes), in robotics [6], and assistive technologies, where it potentially enables people with visual disabilities to localize objects, orient themselves, and navigate environments [44, 53]. It is meaningful to draw inspiration from bats when creating new echolocation technologies; it is also essential to establish the correct design patterns within audio-visual frameworks for virtual environments (VEs).

This work is a continuation of ongoing research about embodiment and navigation in VEs with echolocation that allows users to acquire spatial information about the surrounding environment. Echolocation has been observed in bats, dolphins, whales, birds, and squirrels [20], who are using ultrasonic sounds to attain spatial acuity. While there is a huge evolutionary gap between human and mentioned animals in navigating with echoes, due to different evolutionary paths, some blind and sighted people are also able to echolocate [27]. However, they are generally less proficient in echolocation tasks, and substantial training is required in order to obtain this ability. In this paper, the main goal is to train auditory navigation by developing learning environments with VR technologies that would allow users to reach performances comparable to the ground-truth condition of navigating with immersive visual feedback. Skills requirements for human echolocation are rather unknown. In [2], we conducted several pilot experiments suggesting that people with musical backgrounds performed better when relying on echolocation for completing a navigation task. Specifically, we hypothesized that these users were trying to separate sound pulses from reflections during a basic localization test, whereas non-musicians may have oriented themselves only using reverberations. However, due to perceptual variability among users and insufficient ability to describe the spatial audio qualities that helped orientation in the environment, the current work aims to discriminate more useful features. Therefore, we considered trained musicians as a starting point, being aware of future opportunities to apply our methodology with other user populations (e.g. based on rhythmic and dancing skills, or knowledge on audio production). The main goal of this paper is to identify relevant properties and specific aspects of human hearing and propioception

• *All authors are with Aalborg University Copenhagen, Department of Architecture, Design, and Media Technology, E-mail: {asta,mge,ncn,sts}@create.aau.dk, {jzovne13,kkonov13}@student.aau.dk.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2019.2898787, IEEE Transactions on Visualization and Computer Graphics

which contribute to the design of better environmental acoustics and auditory feedback for an improved quality of the experience during navigation using echoes. Thus, we aimed at characterizing both users and VEs, identifying peculiarities in navigation strategies; to the best of the authors' knowledge, there is little information about this topic in the scientific literature. Therefore, the novel aspects of the current work include the following challenges:

1. creating design patterns for a training simulator in VE;
2. identifying and designing the process for the training procedure;
3. classifying users' navigation strategies;
4. further analyzing critical aspects that help trained musicians to perform better in the VE by identifying users who are able to more easily create spatial cognitive maps during the navigation task.

The paper is structured as follows: Sec. 2 discusses recent literature related to non-visual navigation with echos and VEs, Sec. 3 provides technical details about auditory feedback design and system implementation, Sec. 4 describes the evaluation phase (participants, groups, test protocol, and data analysis), Sec. 5 reports both quantitative and qualitative results, Sec. 6 discusses key elements from the navigation experience, and finally Sec. 7 summarizes the paper's contributions and discusses potential future work.

2 Background

For navigation with echolocation in an environment, several features should be highlighted, mainly the spatial domain (sound pulse emission) and the time domain (level of details), which will be discussed in this section.

Echolocation is a form of audio navigation in an environment. Thus, the more in-depth discussion of the subject is intended to highlight how this navigation process is supported by cognition. Neuroscientific research has produced evidence suggesting that the human brain creates certain associations related to the connected routes being navigated. Distances and directions are calculated, and the boundaries of the environment are predicted by keeping track of the displacement from the point of origin during spatial navigation [56]. Each time such a route is created, the brain calculates the potential outcome of the chosen direction and the process starts again based on the obtained information about the navigational boundaries in a specific direction [46]. Because environmental terrain might include obstacles, the brain recalculates the most optimal direction to the destination taking into account these obstacles.

2.1 Navigation with auditory information

Spatial perception and cognition rely on multimodal information, and especially navigation within an environment, is based on three complementary mechanisms [55]:

- knowledge about a point in space such as a landmark or a destination, providing people with sensory information about their current position and orientation;
- knowledge about a sequence of points (i.e. a path to a destination, or "route knowledge");
- integrated knowledge about the environment (i.e., cognitive-map like knowledge, or "survey knowledge"), supporting the update of the perceived position and orientation in space.

All these aspects contribute to build the so called *spatial cognitive map* that people create during everyday exploration of an environment by continuously update their self-position and orientation in both egocentric and allocentric spatial representations [10]. In particular for VR scenarios, the auditory channel has been increasingly taken into account, considering the constant evolution of algorithms for spatial audio rendering [31, 42]. Moreover, a recent systematic review of the literature on sonification approaches [14] corroborated the idea that spatial features of sound are particularly effective to sonify quantities related to kinematics (i.e., relative to position and motion).

Studies in the fields of behavioral psychology, physiology, and neuroscience support the dual interleaved representation of space in both goal-dependent and -independent manners. The episodic memory and allocentric view is related to latter representation, while the goal-oriented actions in egocentric view promotes the acquisition of landmarks and route knowledge [11]. Place cells within the hippocampus activate for the rapid associative memory connecting the goal and its environment, guiding navigation towards a desired destination, especially in unfamiliar contexts. Because the cortex encodes human spatial skills, the hippocampus is used for recovering such fine spatial details [28] within the influence of task-dependent context cues in order to reinforce memory [43]. Evidence lending credence to this claim has also been produced using VEs [21].

Spatial features can be partially acquired without intentional, conscious focus resulting in learning with no extra workload [35]. For instance, taxi drivers are really efficient at spatial navigation during their routes, and structural brain changes may occur in their hippocampi due to constant training [57]. This might be viewed as an indication that training is essential in spatial orientation ability, and such an acquired skill could be developed over time. Accordingly, interactive and immersive VEs may increasingly be capable of supporting the sense of environmental, personal, and social presence [43]. In particular, algorithms for spatial audio rendering make it possible to direct users' attention to specific aspects of the VR experience with an enhancement in realism that could have a positive influence on performance, workload, and spatial presence scores [9].

In the past decades, a large body of work has explored sensory substitution devices with a special focus on acoustic navigation-aids for visually impaired people. This body of work includes research on applications providing enhanced spatial auditory cues to while walking in an environment. One study explored the effects of non-speech auditory beacons on navigation performance. The study involved sounds that changed in timbre and position, and showed that practice using the system improved both navigation speed and accuracy [50]. Viad *et al.* focused specifically on systems for memorization of spatial scenes by means of 3D audio and movement cues [48]. Katz and coworkers developed a system exploiting a 3D audio VE to investigate structural properties of spatial representations [25].

However, different navigation aids support different strategies for collecting spatial information regarding the position, size, material, and shape of sound reflecting objects [38]. In particular, orientation strategies are characterized by reference point strategies [22] and cyclic patterns [17]. The former requires walking back and forth from a known location to a target destination (e.g., an object, a wall, or the starting position), and the latter consists of successively visiting all places before returning to the starting position.

Finally, it is worthwhile to notice the importance of providing users with more options and flexibility in terms of the available navigation aids in order to accommodate their individual sensory sensitivities (e.g., their level of visual or hearing impairment, their tolerated cognitive load, and their situational preferences) [30]. In particular, Loomis and colleagues studied the preferences and performances with simple and single channel haptic or audio confirmations [30], spatial language guidance, and spatialized sound indicating the location of the next way point [26]. All these multimodal displays could be combined in order to find the required precision in spatial navigation and localization of objects in space.

2.2 Human Echolocation

Echolocation is a process of achieving visual acuity by continuously perceiving reflected sounds (echos) from the environment [23]. Echoes are the emitted sound pulses, revealed in the environment as reflections and reverberations. Habitually obtained audio information from the environment is used for spatial orientation and thus echolocation. The process of sound acquisition is based on binaural processing, where the interaural time difference (ITD) and interaural intensity difference (IID) of a sound source between the two ears allow a user to localize a sound in space [58]. Head-related transfer functions (HRTFs) describe listener's acoustics in a three dimensional space, allowing for the construction of source-listener spatial relations [58]. Amplitude modulations, reverberation, time and frequency disparities are constantly processed by human hearing [53].

Studies have shown that human echolocators are able to navigate in

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2019.2898787, IEEE Transactions on Visualization and Computer Graphics

environments with a certain degree of precision using self-generated mouth clicks, reminiscent of the tongue clicks used by bats during echolocation [45].Notably, expert echolocators may even be able to estimate distances to objects and distinguish between objects that vary in terms of shape, size, and other material attributes [44]. Clicks emitted by human echolocators tend to have a duration of 3–15 ms with peak frequencies ranging from 3-8 kHz [44]. Blind experts are very good at retrieving spatial information through perception of reflections and reverberation, and their ability to navigate based on audition is sometimes considered to be "the second vision", as the absence of visual cues gives auditory perception a more dominant role in spatial perception [29]. Indeed, based on the findings of neuroimagining studies, it has been suggested that the visual cortex of blind people may be co-opted for echolocation [44]. Notably, in the absence of visual cues, head and body movements also play an important role in relation to auditory perception of space during navigation [29].

It has been noticed that sighted people are able to learn echolocation as well [16, 47, 52]. In the presence of two closely spaced audio signals some information gets suppressed, which influences echolocation quality, especially when visual feedback is present. During multisensory integration processes, vision tends to dominant in the spatial domain, while audition tends to dominate in the temporal domain due to its higher temporal acuity [13]. Reverberant environments influence auditory perception if two coherent signals are sent from different locations with a minor delay. That is, auditory system tends to detect the signal which arrives first while suppressing the second one, which is called the "precedence effect" [51]. Animals, such as bats, that are born with echolocation abilities are able to distinguish spatial information of the second order reflections, while humans suppress it. Thus, for audio navigation we need to learn how to inhibit the precedence effect during echolocation [52].

2.3 Echolocation in VR

VR technology provides the necessary tools for building safe training simulations [36], and makes it possible to control the visual and acoustic setting separately. Thus, for a successful process of learning echolocation it would be feasible to remove all visuals after a certain amount of training and leave the participants to rely only on spatial audio navigation. Nonetheless, large individual differences in spatial abilities, spatial hearing skills, and familiarity with the VR equipment also play a critical role, requiring a longer learning period for obtaining acquired echolocation skills [16].

Since the ability to echolocate is contingent on natural and realistic sound reconstruction in VEs, the following features should be taken into account: the amplitude, the room size and architecture through reverberation (time delay from the sonar signal emission to detect the reflections), the frequency of the objects' reflections, the absorption properties of the materials through reflection from the surfaces, and the positions of the sound sources, users included due to self-produced audio. In general, sound has four-dimensional characteristics, as sounds emitted from a source (e.g., human echolocator) propagates in cardioid directivity patterns with a traveling distance of 2-10 meters, where time is considered to be the fourth dimension (e.g., low frequencies have longer decay time than high ones) [45]. For natural perception of auralization of the VE, orientation with respect to the sound sources is essential. Accordingly, HRTFs help to obtain spatial information from the environment allowing for binaural processing.

In principle, all users should have their individually measured HRTFs to be able to extract such information. However, obtaining individual HRTFs recordings is impractical for most real-world applications due to demanding requirements, such as a special measuring apparatus and a time-consuming procedure. The most common practice in relation to VR systems and applications is to use dummy-head HRTFs (generic HRTFs hereafter) for all listeners, without any personalization procedure. On the other hand, constant exposure to generic HRTF listening and cross-modal (audio-video) learning could improve auditory source localization [7]. However, employing generic HRTFs is always a suboptimal condition that introduces listener-specific degradation in sound localization and immersion with high unpredictability [19].

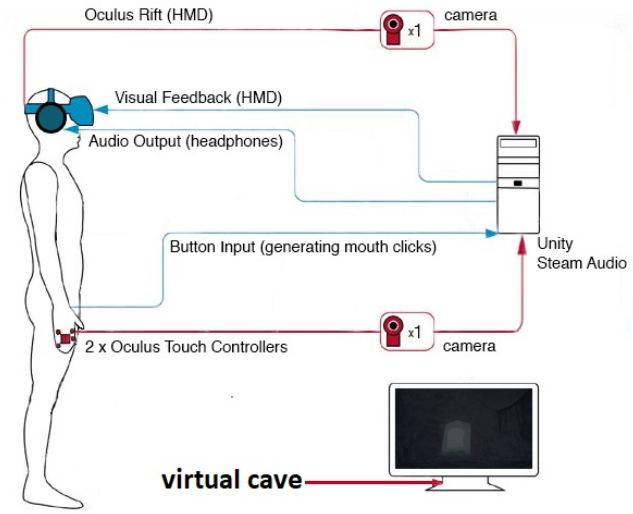


Fig. 2: System hardware/software specification: high-level view.

Evidence of reverberation influencing object detection and navigation with echolocation has been confirmed by the recent scientific literature [37,47] to the extent of small differences in pitch and loudness that can be recognized by both visually-impaired and sighted people. Reverberation provides information about the room volume: the bigger the virtual space is the more reverberation the system should produce. Furthermore, absorption characteristics of the material highly influence the loudness of the reverberation [44].

As a consequence of the described auralization properties of the VE, several approaches could be adopted from sound propagation theories. According to [42], there are two main methods that are used for sound propagation simulation–geometric and numeric. The first method includes physically based models of ray-tracing and imitates sound particles, which is widely used in computer graphics as well, where sound is visualized as spatial propagation of the pulse response from the sound source. The second method is used by directly solving the wave equation, though it is computationally expensive.

The key aspects of the auditory feedback required for a rich perceptual construction of spatial maps of VEs while navigating with spatial audio cues can be summarized as follows:

- binaural rendering through HRTFs;
- time delay from the sonar signal emission for the detection of reflections;
- amplitude (loudness) and frequency of room/object reflections;
- absorption property of materials.

3 System Architecture

In this section, we describe the hardware and software characteristics of our system with more emphasis on the audio architecture. The system is developed with C# in Unity 3D. The echolocation subsystem is realized with the Steam Audio plugin for Unity 3D.

3.1 VR Hardware for Virtual Echolocation

The system architecture consists of an Oculus Rift head-mounted display (HMD) with a resolution of 2160×1200, a refresh rate of 90Hz, and a FOV of 110°. Two Oculus Touch controllers are used for generating the movement of the user's avatar in the VE. Two buttons on the right-hand controller are responsible for the movement of the virtual avatar – "B" for forward movement and "A" for backwards movement. The heading direction is calculated as a direction between the two Oculus Touch controllers (as a normal vector between the controllers that is always pointing in the forward direction). Two Oculus cameras are used for positional tracking of the HMD and controllers.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2019.2898787, IEEE Transactions on Visualization and Computer Graphics

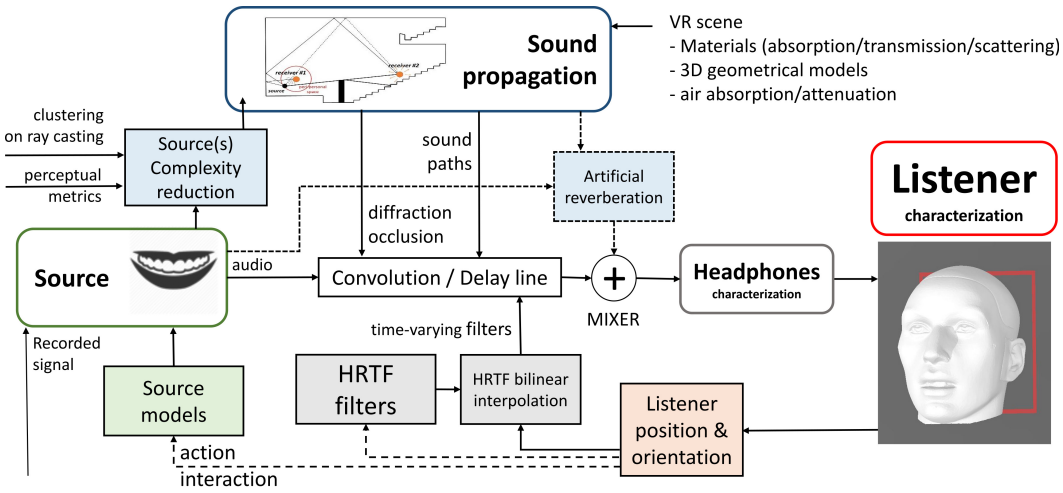


Fig. 3: Simplified block diagram for binaural rendering and auralization.

3.1.1 Navigation Control

To ease task completion, we aspired to make movement through the VE intuitive and effortless. Specifically, we considered different interaction hardware; that is, traditional peripherals (i.e., mouse and keyboard) and VR controllers (i.e., two Oculus Touch controllers). The pilot test involved traditional peripherals while the final navigation test involved the Oculus Touch controllers.

Mouse and keyboard is traditionally considered an effortless means of interacting with computers [34]. However, the mouse was disregarded for the pilot test, as it was likely to cause navigation difficulties when used in combination with the Oculus Rift. Consequently, keyboard buttons (arrows "Up" and "Down") were used for forward and backward movement, while the "space" button – was used for generating the mouth-clicks making echolocation possible. Moreover, the simplicity of the navigation controls was meant to ensure that participants would focus on the auditory feedback rather than introducing unnecessary cognitive load. Accordingly, movement along the ground plane was implemented at a very low and constant pace (Unity speed += 0.1f), allowing the users to hear fine changes of the audio signals.

The pilot test, described in Sec. 4.2.1, revealed that keyboard navigation was not beneficial for HRTF usage as it limited the users' movements. Thus, for the navigation test, described in Sec. 4.2, the "A" and "B" buttons of the right-hand Oculus Touch controller were used for generating forward and backward movement. We chose this control scheme as PC users are familiar with keyboard "Up" and "Down" arrow buttons for forward and backward movement. The direction of movement was controlled by the orientation of the Touch controllers. Decoupling the users' movement and gaze direction increased the likelihood that they would exploit the HRTF during head movements. The "X" button on the left-hand controller was reserved for producing an expert-generated mouth click. When "X" was pressed the sound was generated repeatedly with a constant interval of 1 ms.

3.1.2 Audio Stimulus

The audio stimulus was delivered using a semi-closed Razer SWTOR Gaming Headset. Ad-hoc filters provided by the manufacturer equalized the played-back stimulus in order to reduce spectral variations and coloration due to headphone repositioning which could be difficult to predict also with individual headphone compensation [8,33]. Moreover, it is worthwhile to notice that acquiring individual headphone response is not trivial, requiring listeners to measure user-headphone acoustic coupling before performing the listening test.[1]

[1]Headphone's contribution on immersive spatial audio is a controversial topic. For example, there is no evidence in the scientific literature suggesting degradation in localization due to headphones [40].

The mouth-click was generated in Matlab, based on the frequency derived from studies of blind expert echolocators [45], and was produced by pressing the button "X" of the left-hand controller. The overall system setup is visible in Fig. 2.

3.2 Auditory Feedback for Virtual Echolocation

The Steam Audio engine performed all computations necessary for the auralization, such as the amount and level of reflections, depending on the user location in a scene, and the presence of obstacles/walls in the VE. Steam Audio is a C-based framework that implements physics-based sound behavior in VEs. It is a powerful tool for real-time sound spatialization and propagation, providing high quality real-time frequency dependent audio responses of auditory events that can be designed to create different scenarios and tasks. Figure 3 depicts a typical pipeline of audio-signal processing which is similarly implemented in Steam Audio.

Among the different auralization parameters that allow a flexible real-time acoustic simulation, Steam Audio implements HRTF fast convolution, several options for occlusion and reflections considering different types of raycasting, physics-based attenuation, air absorption, direct and indirect mix level, and indirect artificial binaural reverberation (a similar auralization engine is described by Schissler, Nicholls, and Mehra [39]). It also has a library of pre-made audio shaders representing different settings of frequency dependent sound absorption and transmission. Significant changes in Steam Audio performance was detected when quality and overall amount of rays used in raytracing were increased. Accordingly, the most influential parameters were direct mix Level and indirect mix Level (i.e., the amount of relevant reflections). However, reverberation, reflection and occlusion effects can be either real-time or baked to reduce CPU usage. The large VE created for the current study resembled a real stone cave and was chosen due to its high reverberation capacity and clear acoustic fingerprint. This design decision was to the detriment of computational cost for large spaces.

3.2.1 Sonic Interaction Design for Echolocation in VR

The expert mouth-click was generated with Matlab algorithm described by Thaler *et al.* [45]. It creates a 3ms-long signal simulating the same mouth-click that blind experts use for echolocation (see Fig. 4 for a graphical view of such a signal). The directivity pattern a click would be cardioid with a constant sound propagation level within a 60° cone and gradual level drop outside the cone. The main energy component of the expert click varies in a range of 2-4kHz with an additional peak at 10kHz.

Steam Audio had no controls over the directivity pattern. Particularly, the sound propagation was omnidirectional and no possible changes could be implemented using present settings. Therefore, a natural human body absorption and reflection had to be simulated. For this

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2019.2898787, IEEE Transactions on Visualization and Computer Graphics

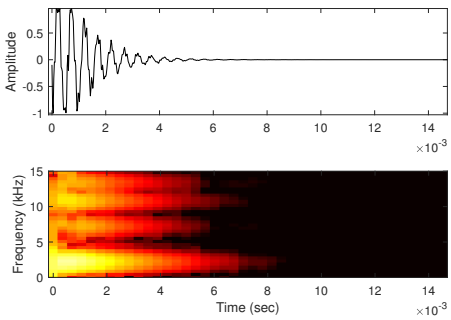
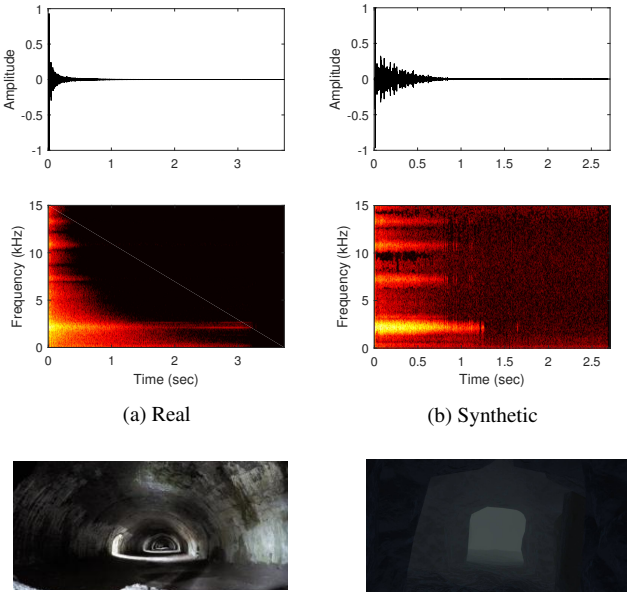


Fig. 4: Expert-generated mouth click: waveform (top) and spectrogram (bottom).



(a) Real (b) Synthetic

Fig. 5: Signal level comparison between real and synthetic clicks in space. (a) and (b) are waveforms and spectrograms of real and synthetic signals, respectively. Bottom pictures depict iconographic representation of the involved RIRs.

purpose, a 3D model of the human body, created using MakeHuman[2], was imported to the scene. An audio source was placed in front of the virtual avatar's lips. A custom audio shader was placed on the avatar's model for imitating the frequency absorption and transmission of the skin, muscles, and bones that would interfere with sound propagation in real life. The Unity "Mesh Renderer" component was deselected for making the 3D model invisible in the scene to limit the computational power dedicated to rendering.

To simulate natural room acoustics the Steam Audio parameters had to be adjusted correctly. With this in mind, we acquired a realistic sound propagation based on analysis of the room impulse response (RIR) of a real cave. RIR shows the response of a system to external alterations [24]. The RIR of the real cave was compared with the synthetic RIR of the VE designed with Steam Audio. The RIR of the real cave was acquired from OpenAIR online impulse response database[3] and convolved with the expert mouth-click generated in Matlab . The Hoffman Lime Kiln stone building with semi-cylindrical shape and several open entrances similar our virtual settings was chosen (real RIR hereafter, see Fig. 5(a) for the reference signal and Fig. 5(b) for a picture of the measurement venue)[4]

Synthetic RIRs were acquired from audio output channels by performing a click in the virtual cave with similar spatial arraignments of the source (avatar's position) and receiver (microphone's position): a source-receiver distance of 3m, and a microphone height of 1.2m.

Henceforth, real and synthetic reference clicks were analyzed in terms of reverberation time (RT60) and direct-to reverberant ratio (DRR). RT60 measures the amount of time necessary for an audio signal to decay by 60 db in a large room according to standard ISO 3382-1:2009 [15]. The adopted algorithm uses reverse cumulative trapezoidal integration to estimate the decay curve, and a linear least-square fit to estimate the slope between 0 dB and -60 dB. DRR is a measure that describes the energy ration between direct sound and reverberations of an audio signal [49], and can be computed from the following equation:

$$DRR = 10 * log10 \left(\frac{\sum_{t=t_0-L}^{t_0+L} x^2}{\sum_{t=t_0+L+1}^{N} x^2} \right) \quad (1)$$

where x is the impulse response of N samples, t_0 is the onset time, and L is the length of the window which contains the early reflection part.

Fig. 5 depicts the final synthetic click in space resulting from a manual tuning of Steam Audio parameters in order to have close RT60 and DRR values compared to the reference measurement. DRRs of the real and synthetic RIRs were -2.29 and -2.72, respectively. The mean RT60 in the real RIRs was 3.38s while in the virtual cave it was 4.95s. This small difference was due to the spatial dimensions of the real and virtual environments not perfectly matching. The relevant final Steam Audio settings can be summarized as follow. Bilinear HRTF interpolation among available HRTFs was applied for directions of direct sound together with physics-based attenuation based on inverse-squared law with distance and frequency-dependent air absorption. In the overall mix, direct level was set to 0.9 on a normalized scale with real-time reflections mix level equal to 0.52 (min. value is 0 and max. value is 16). Reverberation was continuously simulated with real-time rays, and secondary rays and bounces were set to the maximum allowed values. Moreover, frequency dependent transmission occlusion was implemented via a partial method within the radius of 10 meters, performing a raycasting from the source to the listener in order to compute the desired proportion of occluded rays. In particular, the mesh audio shader for the main VE material was set "Rock" which acoustically absorbs all frequencies and has the following absorption and scattering properties (all values are normalized in the range [0 1]):

- *Low frequencies* (band central frequency 400 Hz): absorption equals to 0.13 and transmission equals to 0.015.
- *Middle frequencies* (band central frequency 2.5 KHz). absorption equals to 0.20 and transmission equals to 0.002.
- *High frequencies* (band central frequency 15 kHz): absorption equals to 0.24, and transmission equals to 0.001.
- *Scattering*: 0.05 (mirror-like manner).

4 METHODS

The main goal of the tests was to investigate whether echolocation cues simulating real life experiences in the virtual cave were sufficient for the correct navigation in the VE. Therefore, first the quality of the audio cues provided by Steam Audio were assessed through the pilot test, and subsequently the actual navigation test in the VE was performed. Inspired by previous studies, our test design included a training procedure followed by the actual test. The completion time was between 40 minutes and 1 hours, and the training session lasted between 20 and 30 minutes, depending on participants' needs. For this purpose we compared four conditions. The first three audio conditions were devoid of visual feedback, and the fourth condition included the visual-only feedback that acted as the *ground-truth* for navigation performances in the test:

[2]http://www.makehumancommunity.org/

[3]http://www.openairlib.net/

[4]http://www.openairlib.net/auralizationdb/content/hoffmann-lime-kiln-langcliffeuk . The picture of Hoffman Lime Kiln stone building was collected from OpenAIR website.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2019.2898787, IEEE Transactions on Visualization and Computer Graphics

1 **RF**: Unimodal auditory feedback based on reflection rendering only.

2 **RV**: Unimodal auditory feedback based on late reverberation only.

3 **RR**: Unimodal auditory feedback based on reflections together with reverberation.

4 **V**: Unimodal visual feedback.

Both the pilot test and navigation test included the same training and testing procedure. The difference between the two tests was the hardware used for controlling the virtual movement, as described in Sec. 3.1. It should be stressed that the participants could visually experience the virtual cave in the training sessions and the proposed navigation task could be considered a spatial memory recall task with auditory echos [10]. The immersion in a multimodal learning environment cognitively allowed an integration of modalities that might be enhanced and extended in order to get ideally comparable localization and navigation performances between auditory echos cues and visual-only feedback.

All data was collected using non-probabilistic sampling method, in particular convenience sampling. Both tests followed a within-subject design. We conducted the experimental session at the Multisensory Experience Lab, Aalborg University Copenhagen. For the pilot study, 6 participants (4 males and 2 female) were invited, ages between 23-38 (M = 27.33, SD = 5.75). They had 5 to 30 years of musical experience (M = 13.83, SD = 80.7). For the navigation test, 26 participants (18 males and 8 females) were invited, ages between 19-58 (M = 34.88, SD = 10.3). They had 2 to 40 years of musical experience (M = 15.92, SD = 10.2) and they did not participate in pilot test. The adopted experimental procedures were in accordance with the Declaration of Helsinki (Edition 2013). All participants were informed of the main goal of the research and provided informed consent before the test commenced. All participants reported that they had normal hearing and that they were semi-professional musicians.

4.1 Training Procedure

Generating mouth-clicks is quite challenging, as it is difficult to produce a very short click within a specific range, as described in Sec. 2. Humans normally rely on multisensory integration when orienting themselves in an environment. They primarily rely on vision, but the vestibular, auditory, and somatosensory systems also play a central role [12]. Therefore, it might be challenging for users without visual impairments to primary rely on sounds when navigating in VR without prior training. Additionally, novice users needs to get familiar with the hardware and interaction techniques making virtual navigation possible. Notably, it has been demonstrated that healthy individuals can be trained to echolocate using mouth clicks or finger taps [47]. For these reasons a training session was included.

When the training session started users were immersed into a version of the VE that included both visual and auditory stimuli. The main purpose was to make participants familiar with the VE and hardware (Oculus Touch controllers and HMD), and at the same time to get them acquainted with the clicking sound and its spatialization properties. For this reason, users were asked to walk around the virtual cave and explore how the sonic environment changed in response to the clicking sounds. Specifically, they were encouraged to move close to the walls and hear the sound reflections and reverberations, and move along the virtual corridors and hear the changes in frequencies.

When they reported being familiar with the task, the virtual lights were turned off. While in the dark VE, the users were asked to identify directions to the walls and the corridors, and report if they were close to or far away from a wall. These reports were based only on audio navigation. After a five minute break, the navigation test started. After the test the users filled in a digital questionnaire. The same procedure was applied during the pilot test.

4.2 Spatial Orientation Test

The participants were tasked with finding their way out of the corridor/tunnel to the bigger hall of the virtual cave (see Fig. 6 for a top view map of the testing scenario). The starting positions and rotations were randomized across conditions. As soon as the participants entered the bigger hall, distance travelled, completion time, and amount of clicks were logged for each participant.

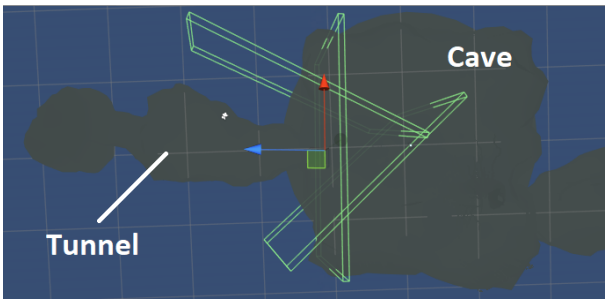


Fig. 6: A top view map of the virtual cave. When colliders were activated datalog file registered distance walked, time and number of clicks.

By pressing the "X" button on the right-hand Oculus Touch controller, participants were able to hear the clicking sounds through the Razer SWTOR Gaming headphones and navigate the VE using echolocation. Time, spectrum, and amplitude evolution of cliks in space provided feedback to the participants in terms of distance to the objects/walls in the VE. The closer participants were to the walls the faster the reflections returned to the participants and vice versa. In terms of reverberation, the amplitude of the sound was higher when coming closer to the walls.

During the navigation test the conductor was taking notes while observing participants. The main reason for observing the participants' behavior was to see if they appeared to employ a specific navigation strategy, if they kept the button generating mouth-clicks pressed all the time or only pressed it occasionally, and if participants appeared aware of the obstacles/walls, which would indicated that they were able to correctly distinguished between the audio signals .

After each condition, the participants were asked to fill a short questionnaire. Because the main research objective was to determine how well the participants could orient themselves using echolocation, all the notes and questionnaire items were related to navigation quality. The questionnaire contained four questions:

1. How informative do you think this condition was?
2. Which frequency helped you to navigate the most through the environment (low, middle, high)?[5]
3. Describe how the chosen frequency helped you to navigate.
4. Describe which strategy you used (how did you orient yourself in the environment, how did you walk, how much did you press the button, if you pressed it during walking).

At the end of the test, the participants were asked to indicate which of the three audio conditions they found the most informative in terms of navigation.

4.2.1 Pilot Test

The technical goal for this test was to examine if the implemented mouth-click and RIR were perceived as natural by the participants in the designed environment. Furthermore, the overall goal of the pilot test was to examine if the participants were able to navigate with keyboard controls by hearing the different audio signals and echoes in the VE. With exception of the varying navigation controls, described in Sec. 3.1.1, the same procedure was used for both the pilot test and the final navigation test. The results of the latter test are presented in Sec 5. The data obtained from the pilot test were compared using one-way ANOVAs to determine if there were statistically significant differences between the four conditions (RR, RF, RV and V). Pairwise comparisons of conditions were performed using Turkey HSD test with 95% confidence level.

[5]It is worthwhile to notice that all participants were semi-professional musicians and some of them had audio production skills, thus resulting in precise

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2019.2898787, IEEE Transactions on Visualization and Computer Graphics

Conditions	Group with spatial cognitive map (CM)	Group without spatial cognitive map (N)
RR **Reflection** **Reverberation**	wall detection, informative reverb decay, spatial cognitive map - walls distance and direction, sound spectrum was clear, high frequency preference, head and body movements present	wall detection, sound volume change when closer to the walls, reverb tail contained information about the space, high frequency preference, head and body movements absent
RF **Reflections only**	wall detection, sound spectrum was clear, spatial cognitive map - walls distance and direction, artifact detection, high and middle frequency preference, head and body movements present	wall detection, high frequency preference, head and body movements absent
RV **Reverberation only**	wall detection, not informative reverb decay, clear feeling of space, spatial cognitive mapping - direction, high frequency preference, head and body movements present	volume change when closer to the walls, artifact detection, informative reverb and decay, high frequency preference, head and body movements absent

Table 1: Summary of main participants' trends in the qualitative evaluation for each group divided by auditory feedback condition.

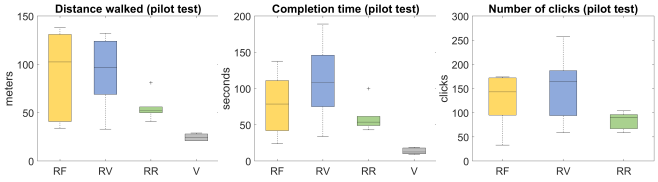


Fig. 7: Boxplots summarizing the results of the pilot test related to walked distance (left), completion time (middle), and number of clicks (right).

In relation to completion time, the test showed statistical significance between the four conditions [F (3, 20) = 7.192, p=0.02].The pairwise comparisons revealed significant differences between V and RF (p = .029) and V and RV (p = .001). RR got the smallest mean in seconds (M = 60.19, SD = 20.46) among audio-only conditions. A statistically significant difference was also found between the four conditions [F (3, 20) = 6.727, p=.003] in regards to walked distance. Pairwise comparison identified significant differences between V and RF (p = .006) and V and RV (p = .006). Moreover, RR revealed the smallest mean in meters (M = 55.33, SD = 13.58) among audio-only conditions. The biggest number of clicks was performed during the RV condition (M = 154.17, SD = 70.26), while the smallest number of clicks was detected in the RR condition (M = 84.17, SD = 17.45).[6] This might be viewed as an indication that RR requires less time and distance in regards to task completion compared to RV and RF conditions. The results from the pilot test are presented in Fig. 7.

Finally, participants reported reduced and constrained head movements due to the forced posture while using keyboard controls. From their comments, we implemented a more embodied interaction with the Touch Controller buttons as described in Sec. 3.1.1.

5 Results

Subsections 5.1, 5.2, and 5.3 present the results pertaining to qualitative data, behavioral measures, and preference rankings gathered during final navigation test.

5.1 Qualitative participants characterization

At the beginning of the test some demographic data was collected in order to get an overview of the 26 participants' background, The majority of the test participants were playing a piano (12/26), guitar (7/26), some of them were singers (3/26), a couple of them were composers (2/26), one was playing violin and one was playing clarinet, the rest had experience with sound engineering (10/26) and had some experience playing instruments as well.

The characterization of participants were performed by analyzing their comments and searching for the following key elements with reference to Sec. 2.1:

- *Route knowledge*: Identification of wall positions (left/right from the user).
- *survey knowledge*: Finding the direction of the walls and orienting accordingly.
- *Landmark/destination knowledge*: Identification of the opened way (when being in the middle of the corridor) and choosing the free way and, thus the right route to the goal.

The self-reporting of all such elements meant that a participant was able to create a cognitive map (CM) of the space and goal. These participants were assigned to a special group, labelled CM hereafter. In contrast, participants who did not report forming spatial cognitive map were assigned to group N. Both groups' behavioral trends can be found in Table 1. These behavioral trends are based on the derived observations and comments described through the three audio conditions (RR, RF, and RV). The observations indicated whether the participants belonging to group CM and N were able to detect obstacles/walls and moved with their head and body during the navigation test. Self-reported data is presented as preferred and informative audio frequency, reverberation information about the size of the environment, sound spectrum, and reported spatial cognitive mapping.

CM Group The CM group consisted of 8 males and 2 females (M =34.88, SD = 10.29) with musical background (M = 15.92, SD = 10.22). All these participants had higher education and were working with computer technologies: 6 people from this group had a sound engineering background, 1 had a lighting design technology background, 2 had a media technology background, and 1 had an engineering background. Of the 10 participants, 8 reported using VR technology on a regular basis, and the one participant with an engineering background had never experienced VR.

Regarding the synthesized mouth click, this group used different strategies: from keeping the button pressed all the time (*dynamic* strategy hereafter) to stopping and hearing the sound (*static* strategy hereafter). In the latter case, they pressed the button once and tried to move with their body and the head. The two participants who scored the highest in all the audio conditions were observed to move a lot with their body and the head in all directions. This appears to have helped them to perceive the HRTF localization cues better and allowed them to orient themselves faster than the rest of the group. These two participants also held the button pressed during walking but released

reports on time-spectral quality of the auditory feedback.

[6]No clicks were triggered in visual condition. Accordingly, the V condition is not displayed in Fig. 7 (right).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2019.2898787, IEEE Transactions on Visualization and Computer Graphics

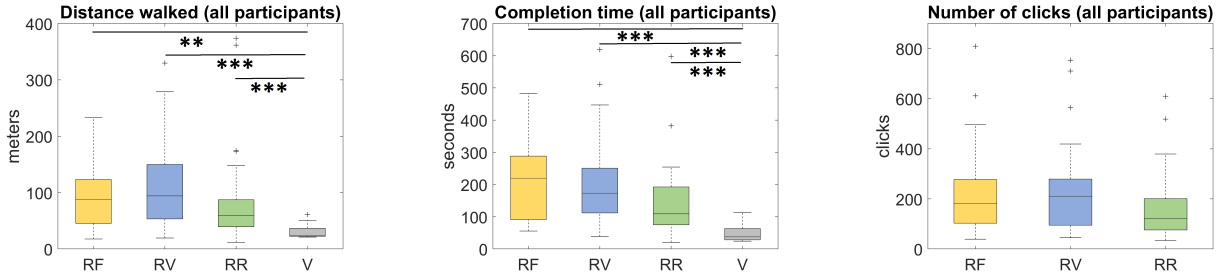


Fig. 8: Boxplots summarizing the results related to walked distance (left), completion time (middle), and number of clicks (right) for all participants ($n = 26$) across conditions. Asterisks and bars indicate, where present, a significant difference (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$ at paired comparison).

it when they decided to stop. While stationary they would move their heads around and occasionally press button to instigate clicks.

N Group The group consisted of 9 males and 7 females (M =37.25 ; SD = 10.13) with musical background (M = 18, SD = 10.9). The participants had varying backgrounds, but the majority worked within IT (9/16). Half of the participants in this group (8/16) had experience with VR technology.

The participants in the N group used different strategies: walking along the walls (*wall* strategy hereafter), trying to identify the direction based on the sound volume (if they stood in front of the wall the sound got louder and got back faster, meaning that they could better identify reflections), pressing button continuously or not. In general, participants in this group reported that they got lucky finding the exit to the bigger hall, as they had problems with distinguishing between the sound signals in general and did not hear reflections very well.

5.2 Behavioral measures

The following analyses of data obtained from the three behavioral measures (distance walked, completion time, and number of clicks) were performed:

- To determine if there were any statistically significant differences between the four conditions (RF, RV, RR, and V), the data obtained from all participants ($n = 26$) were analyzed.
- As suggested, a subset of the participants ($n = 10$) were believed to deliberately form spatial cognitive maps while navigating the VE (group CM), whereas the remaining participants ($n = 16$) did not appear to employ this strategy (group N). Because this strategy might influence the utility of the three conditions devoid of visual feedback (RF, RV, RR), separate analyses of the behavior of group CM and group N were performed.
- Finally, the behavior of group CM was compared to the behavior of group N. The visual condition (V) was not included in this analysis.

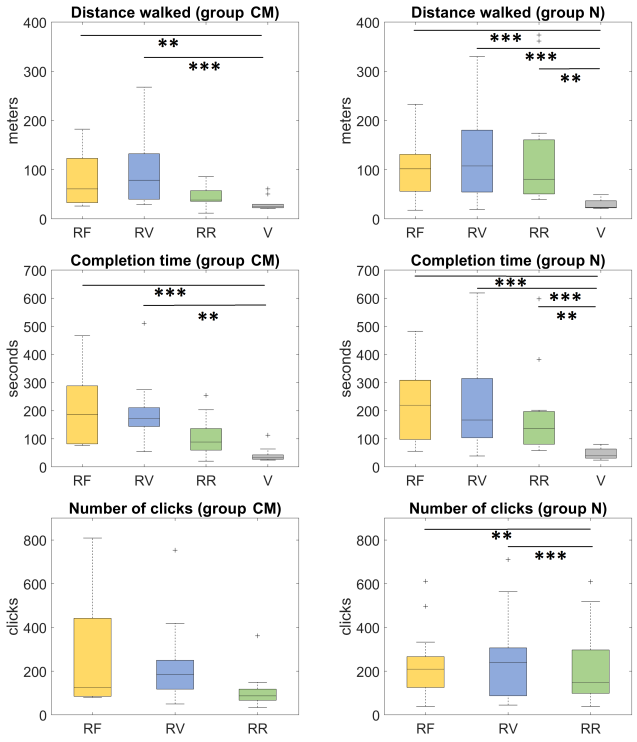


Fig. 9: Boxplots summarizing the results related to walked distance (top), completion time (middle), and number of clicks (bottom) for group CM ($n = 10$, 1st column) and group N ($n = 16$, 2nd column) across conditions. Asterisks and bars indicate, where present, a significant difference (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$ at paired comparison).

The data obtained from the behavioral measures were treated as interval data. However, outliers were identified and the assumption of normality was violated for almost all conditions and groups, as assessed by inspection of boxplots and Shapiro-Wilk tests ($p < .05$), respectively. Thus, the non-parametic Friedman test was used for all statistical within-subjects comparisons, and the Mann-Whitney U test used for between-group comparisons.

Cross-condition comparison: Friedman tests were used to determine if there were statistically significant differences between the three behavioral measures for all participants. Pairwise comparisons of conditions were performed using Dunn-Bonferroni tests. With respect to *distance walked* (Figure 8, left), the test suggested that there was a statistically significant difference ($\chi^2(3) = 42.046, p < .001$), and the pairwise comparisons revealed significant differences between V and RF ($p = .001$), V and RF ($p < .001$), and V and RR ($p < .001$). Similarly, the test performed in relation to *completion time* (Figure 8, middle) indicated a statistically significant difference between the four conditions ($\chi^2(3) = 43.385, p < .001$), and the pairwise comparisons found significant differences between V and RF ($p < .001$), V and RF ($p < .001$), and V and RR ($p < .001$). Finally, the test performed in relation to *number of clicks* (Figure 8, right) across RF, RV, and RR did not indicate that the difference between conditions was statistically significant ($\chi^2(2) = 5.650, p < .059$). Thus, as expected the participants walked significantly shorter and completed the task significantly faster when visual stimuli was presented. Moreover, no differences between the three conditions devoid of visuals feedback were found for any of the three behavioral measures when considering the data obtained from all participants.

Cross-condition comparison for group CM: Friedman tests and pairwise comparisons using Dunn-Bonferroni tests were also used to

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2019.2898787, IEEE Transactions on Visualization and Computer Graphics

determine if there were statistically significant differences between conditions when considering only group CM (Figure 9, 1st column). A significant difference was found with respect to *distance walked* ($\chi^2(3) = 21.720, p = .273$), and the pairwise comparisons suggest that V yielded significantly shorter walking distances compared to RF ($p = .003$) and RV ($p < .001$), but the distance was not significantly different when V was compared to RR ($p = .340$). No significant differences were found between the three conditions devoid of visual feedback. A significant difference was also found in regard to *completion time* ($\chi^2(3) = 22.560, p < .001$), and the pairwise comparisons indicated that V yielded significantly faster completion times compared to RF ($p < .001$) and RV $p = .001$, but the difference between V and RR was not significant ($p = .500$). Again, no significant differences were found between the three conditions devoid of visual feedback. Finally, no significant difference was found with respect to *number of clicks* ($\chi^2(2) = 1.400, p = .497$).

Cross-condition comparison for group N: Again Friedman tests and pairwise comparisons using Dunn-Bonferroni tests were also used to determine if there were statistically significant differences between the behavioral measures when considering only group N (Figure 9, 2nd column). A statistically significant difference was found with respect to *distance walked* ($\chi^2(3) = 22.575, p < .001$), and the post hoc comparisons indicated that the participants walked significantly shorter during V compared to RF ($p < .001$), RV $p < .001$, and RR $p = .004$. A statistically significant difference was also found in regard to *completion time* ($\chi^2(3) = 24.300, p < .001$), and the post hoc comparisons indicated that V yielded significantly faster completion times than RF ($p < .001$), RV ($p < .001$), and RR ($p = .001$). A significant difference between the three conditions was also found with respect to *number of clicks* ($\chi^2(2) = 21.375, p < .001$), and the post hoc comparisons indicated that there were significant differences between RR and RF ($p = 0.001$) and between RR and RV ($p < 0.001$). Specifically, the participants in group N clicked significantly less during RR compared to RF and RV.

Comparison of group CM and group N: To determine if the behavioral measures differed significantly between group CM and group N (see Fig. 9), nine pairwise comparisons were performed using Mann-Whitney U tests. That is, for each of the three measures, we compared the three conditions devoid of visual feedback of group CM with the corresponding conditions of group N. With respect to *walking distance*, RR was found to differ significantly between group CM and group N ($U = 24, z = -2.951, p = .002$), and RR was also found to differ significantly in terms of *number of clicks* between the two groups ($U = 42, z = -2.003, p = .047$). That is, during exposure to RR, the participants in group CM walked significantly shorter and they clicked significantly less compared to the participants in group N.

5.3 Preference ratings

Out of the 26 participants, 3 preferred RF, 3 preferred RV, and 20 preferred RR. A chi-square goodness-of-fit test was performed to determine whether the frequencies were equal. The minimum expected frequency was 8.7. The chi-square goodness-of-fit test indicated that there was a statistically significant difference in the number of participants who preferred each of the three conditions ($\chi^2(2) = 22.231, p < .001$). The majority of the group CM preferred RR condition (7/10), 3 people preferred RF condition (3/10) and none chose RV condition to be their preference. Similarly, the majority of group N preferred RR condition (12/16), 2 people preferred RF condition (2/16) and 2 people chose RV condition (2/16).

6 Discussion

Our previous work [2], indicated that people with musical backgrounds may perform better when relying on echolocation for completing a navigation task. Nonetheless, the present results suggest improvements in our approach.

Generally, the test results provided evidence of statistical significance only between condition V and the rest of the audio conditions (RR, RF and RV). The results indicated that the travelled distances were shorter and completion times faster during the visual condition. This difference is to be expected, as vision generally is the dominant sense among unimpaired users, apart from specific cases where audition dominates [29]. Without constant repeated practice, it is likely to be difficult for most people to navigate blindly when relying on purely auditory information about the environment [47]. The duration of the test (1-1.5 hrs) and completion times suggested that this was not enough for the naïve participants to learn to navigate without difficulties in the VE. Nonetheless, the test did demonstrate that the VE provided a controlled and safe environment for training purposes, as many participants repeatedly collided with the walls without noticing that the walls were in front of them.

In group CM, there was a significant difference between the four conditions among participants. These results support what we observed during the pilot test, where RR was similar to V in terms of completion time and distance walked. This gives us some reason to suspect spatial cognitive maps can be produced based on echolocation in the VE after an adequate multimodal training period [36, 54]. Regarding group N, participants walked shorter and faster in the V condition. Furthermore, a significant difference between the amount of clicks was found between the three audio conditions in this group. Participants were clicking less during exposure to RR compared to RV and RF. This might indicate that the information was more substantial in RR condition [16]. Nonetheless, compared to group CM, participants clicked more, which might indicate that they needed more information from the environment than group CM did, or that they were unable to derive as much information from each click.

Spatial Cognitive Mapping One of the findings of the navigation test was the self-reported acquisition of spatial cognitive maps, which appear to have been created by the participant during navigation task. The CM group may have been better than the N group at identifying obstacles and calculating the route in terms of direction, using localization cues provided by generic HRTFs, and distance, detecting reflections [44]. Since the CM group appears to have performed better, a stronger musical ability could have contributed [5] together with a better knowledge of the VR equipment. However, this is debatable, as there is likely to be highly complex interrelation between auditory perception, audio reproduction, spatial memory, and cognitive processes. Due to many individual differences among participants, it is difficult to identify specific correlations between the employed navigation strategies and the required auditory information needed to aid navigation in VEs. Furthermore, some participants were able to differentiate between first order reflections and the late reverberation tail in signals. This shows an ability to decode complex reflection patterns, which is quite a challenging task for a non-trained population. New tests and evaluation protocols are needed in order to determine which aspects of human hearing allow for better decoding of spatial echos. Non-musicians might also perform relatively well if their spatial orientation in regards to spatial cognitive maps is better developed than musicians. The participants' previous experience with spatial audio technologies, their localization accuracy with non-individual HRTFs [18], their sensitivity to spectral shape [3], or the connection with body/head movements in embodied cognition [41] should be carefully considered in a complete user characterization.

The precedence effect might also have played a role during the test and may have influenced the results for group N, as some test participants from this group may have suppressed the audio signals arriving secondly, especially when the participants were not able to rely on the visual information after the training session [51]. This consideration is based on the observations and self-reported data, indicating that some participants completed the task by chance.

Training and Testing Procedures The design of the training session indicated that it was important for participants to have the opportunity of cross-modal binding between visual information from the environment and its auditory characterization. If this had not been a possibility then it would probably have been more difficult for them to derive meaning from the acoustic information and use it for navigation when visual stimuli was no longer presented. However, due to indi-

vidual differences the session time took between 20 and 30 minutes. Moreover, some participants spent a considerable amount of time on trying to get familiar with Oculus Touch controllers and this may have affected their performance.Thus, the learning process of echolocation while moving is likely to be a time-consuming process, especially in technology-mediated environments. Body-based cues can also be related to vibro-tactile feedback that might compensate for the limitations in the training process or be used in echolocation-inspired navigation aids. The final goal should be to provide a technologically-augmented solution for an effective and improved spatial orientation for different situations and users [30].

Since the training session only involved the RR and V conditions, participants noticed that the sounds differed during the testing session. As the conditions were randomized during the navigation test, the first condition generally performed worst, as was noticed by the instructor. Participants were often asking, if they were heading the right direction. Thus, it seems likely that if participants got a training session through all the three conditions they might have learned how to differentiate between early reflections and reverberations, and thereby may have learned how to avoid the suppression effect. Accordingly, this was the case during the pilot test, described in Sec 4.2.1, as the same test participants were chosen for previous pilot tests where they knew already the differences in audio signals.

Strategies All strategies were identified based on self-reporting. In regards to the presented scores, the dynamic strategy was the best among CM group, while the static strategy was the worst. Previous research indicates that movement in an environment where the sound source is present helps navigation due to Doppler frequency shifts generated by the movement of the sound source and observer [36]. In case of the current study, this would occur when participants generated a clicks continuously (i.e., they followed the dynamic strategy). Auditory feedback helps participants estimate if they are in motion and might activate a vestibular response. On the other hand, in case of the current study reverberation would only provide static auditory distance information. Therefore moving test participants might have produced better results than in the the ones employing the static strategy. The wall strategy appears to be the second best, as some of the participants were also constantly on the move, but some were playing the sound first and then moving along the walls.

In general, individual differences and varying approaches exist when navigating and selecting the best strategy [4]. In particular, the current findings provide a high-level view which will be able to guide an ad-hoc experimental design aimed at characterizing users and identifying basic skill requirements, peculiar strategies, and individual learning protocols.

7 Conclusion and Future Work

In this paper, we presented a study of navigation with echolocation in VR. Based on our previous experiments, we designed a training procedure where the participants got acquainted with the mouth-generated clicking sound and real-time spatial sound propagation using the RIR from a real-world settings. The participants were placed in a virtual cave and performed a navigation task. During the test 10 out of 26 participants (group CM) reported the acquisition of spatial cognitive maps of the VE. Our results showed that not all trained musicians were able to perform the task equally fast, suggesting the need of a further user characterization. Furthermore, the results related to group CM revealed significant differences between RR and the other two audio conditions (RF and RV), and more interestingly no significant differences in navigation performances compared with the ground-truth vision-only condition. A statistically significant difference was also found between group CM and group N (participants who did not report spatial cognitive mapping) in the RR condition in terms of travelled distance and number of clicks with more efficient performances for the CM group. Finally, the test allowed us to identify self-reported navigation strategies with auditory echoes among 26 participants. Three high-level strategies could be labelled as follow: (i) dynamic click-and-move, (ii) static stop-click-and-move accordingly, and (iii) wall-referenced. The dynamic strategy elicited the best performance in terms of completion time and travelled distance among CM group in RR condition.

In general, the work indicates that current VR technologies can provide a safe and controlled settings for this type of experiments and tasks. However, several details of the setup should be improved in future experimental sessions. The pilot test showed that keyboard usage was not optimal due to the restrictions it imposed on the participants' movements. Therefore the bottons of the Oculus Touch controller were used for controlling the virtual movement in the second study. Pressing the buttons of the controller might also have divided the participants' attention, as some of them had never tried VR technology before. Furthermore, a more natural movement algorithm (e.g., using walking-in-place or employing arm gestures), may have been more comfortable from the user interaction point of view. Moreover we have used arm gesture algorithm already in the previous studies thus for flying [1].

The observation that dynamic head and body movements provide a better performance is in line with previous work [44], and it would be relevant for future work to explore how self-generated mouth-clicks would affect participants' sense of agency and performance. Accordingly, real-time microphone input should be implemented in order to investigate the perceived embodiment of the auditory feedback.

Moreover, HRTF personalization procedures based on anthropometry or listener preferences (see [19] for a literature review on this topic) should be considered in order to investigate spectral sensitivities and the impact of customized localization cues compared to generic HRTFs. Finally, an objective characterization of navigation strategies with a detailed analysis on participants' movements will be included in future studies to further develop new complementary screening tests and individual protocols for training sessions. Different strategies might be identified more definitively by introducing specific constraints, such as imposing constraints on the walking velocity or limiting head movements. Finally, future studies might focus on special populations of visually-impaired people, including both expert and non-expert echolocators, and it would also be relevant to explore populations working in immersive audio production or in the performative arts in order to stress the connection between sound and movement.

Acknowledgments

This study was supported by the internationalization grant of the 2016-2021 strategic program "Knowledge for the World" awarded by Aalborg University to Michele Geronazzo.

References

[1] A. Andreasen, N. C. Nilsson, and S. Serafin. Agency enhances body ownership illusion of being a virtual bat. In *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 505–506, 2018.

[2] A. Andreasen, J. Zovnercuka, K. Konovalovs, M. Geronazzo, R. Paisa, and S. Serafin. Navigate as a Bat. Real-Time Echolocation System in Virtual Reality. In *Proc. 15th Int. Conf. Sound and Music Computing (SMC 2018)*, pp. 198–205. Cyprus, July 2018.

[3] G. Andéol, E. A. Macpherson, and A. T. Sabin. Sound localization in noise and sensitivity to spectral shape. *Hearing Research*, 304:20–27, Oct. 2013.

[4] L. J. Anooshian. Diversity within Spatial Cognition: Strategies Underlying Spatial Knowledge. *Environment and Behavior*, 28(4):471–493, July 1996.

[5] J. Beament. *How We Hear Music: The Relationship between Music and the Hearing Mechanism*. Boydell and Brewer, ned - new edition ed., 2001.

[6] M. M. Beigi and A. Zell. A boosting approach for object classification in biosonar based robot navigation. In *Robotics and Automation, 2008. ICRA 2008. IEEE International Conference on*, pp. 3270–3275, 2008.

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

[8] B. B. Boren, M. Geronazzo, P. Majdak, and E. Choueiri. PHOnA: A Public Dataset of Measured Headphone Transfer Functions. In *Proc. 137th Conv. Audio Eng. Society*, Oct. 2014.

[9] K. Bormann. Presence and the Utility of Audio Spatialization. *Presence*, 14(3):278–297, June 2005.

[10] N. Burgess. Spatial memory: how egocentric and allocentric combine. *Trends in Cognitive Sciences*, 10(12):551–557, Dec. 2006.

[11] F. Chersi and N. Burgess. The Cognitive Architecture of Spatial Navigation: Hippocampal and Striatal Contributions. *Neuron*, 88(1):64–77, Oct. 2015.

[12] M. Dieterich and T. Brandt. Global orientation in space and the lateralization of brain functions. *Current opinion in neurology*, 31(1):96–104, 2018.

[13] J. Driver and C. Spence. Multisensory perception: beyond modularity and convergence. *Current biology*, 10(20):R731–R735, 2000.

[14] G. Dubus and R. Bresin. A Systematic Review of Mapping Strategies for the Sonification of Physical Quantities. *PLoS ONE*, 8(12):e82491, Dec. 2013.

[15] F. Dunn, W. Hartmann, D. Campbell, and N. H. Fletcher. *Springer handbook of acoustics*. Springer, 2015.

[16] M. Ekkel, R. van Lier, and B. Steenbergen. Learning to echolocate in sighted people: a correlational study on attention, working memory and spatial abilities. *Experimental brain research*, 235(3):809–818, 2017.

[17] F. Gaunet and C. Thinus-Blanc. Early-Blind Subjects' Spatial Abilities in the Locomotor Space: Exploratory Strategies and Reaction-to-Change Performance. *Perception*, 25(8):967–981, Aug. 1996. doi: 10.1068/p250967

[18] M. Geronazzo, E. Sikström, J. Kleimola, F. Avanzini, A. De Götzen, and S. Serafin. The impact of an accurate vertical localization with HRTFs on short explorations of immersive virtual reality scenarios. In *Proc. 17th IEEE/ACM Int. Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 90–97. Munich, Germany, Oct. 2018.

[19] M. Geronazzo, S. Spagnol, and F. Avanzini. Do We Need Individual Head-Related Transfer Functions for Vertical Localization? The Case Study of a Spectral Notch Distance Metric. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(7):1243–1256, July 2018.

[20] D. R. Griffin. Echolocation by blind men, bats and radar. *Science*, 100(2609):589–590, 1944.

[21] D. Hassabis, C. Chu, G. Rees, N. Weiskopf, P. D. Molyneux, and E. A. Maguire. Decoding neuronal ensembles in the human hippocampus. *Current Biology*, 19(7):546–554, 2009.

[22] E. W. Hill, , and Others. How Persons with Visual Impairments Explore Novel Spaces: Strategies of Good and Poor Performers. *Journal of Visual Impairment and Blindness*, 87(8):295–301, 1993.

[23] M. W. Holderied, C. Korine, M. B. Fenton, S. Parsons, S. Robson, and G. Jones. Echolocation call intensity in the aerial hawking bat eptesicus bottae (vespertilionidae) studied using stereo videogrammetry. *Journal of Experimental Biology*, 208(7):1321–1327, 2005.

[24] P. T. Hong. *Introduction to digital signal processing: Computer musically speaking*. World Scientific, 2009.

[25] B. F. G. Katz, S. Kammoun, G. Parseihian, O. Gutierrez, A. Brilhault, M. Auvray, P. Truillet, M. Denis, S. Thorpe, and C. Jouffrais. NAVIG: augmented reality guidance system for the visually impaired: Combining object localization, GNSS, and spatial audio. *Virtual Reality*, 16(4):253–269, June 2012.

[26] R. L. Klatzky, J. R. Marston, N. A. Giudice, R. G. Golledge, and J. M. Loomis. Cognitive load of navigating without vision when guided by virtual sound versus spatial language. *Journal of Experimental Psychology: Applied*, 12(4):223–232, 2006.

[27] A. J. Kolarik, S. Cirstea, S. Pardhan, and B. C. Moore. A summary of research investigating echolocation abilities of blind and sighted humans. *Hearing research*, 310:60–68, 2014.

[28] B. S. Kolarik, K. Shahlaie, A. Hassan, A. A. Borders, K. C. Kaufman, G. Gurkoff, A. P. Yonelinas, and A. D. Ekstrom. Impairments in precision, rather than spatial strategy, characterize performance on the virtual morris water maze: A case study. *Neuropsychologia*, 80:90–101, 2016.

[29] J. Lewald. Exceptional ability of blind humans to hear sound motion: implications for the emergence of auditory space. *Neuropsychologia*, 51(1):181–186, 2013.

[30] J. R. Marston, J. M. Loomis, R. L. Klatzky, and R. G. Golledge. Nonvisual Route Following with Guidance from a Simple Haptic or Auditory Display. *Journal of Visual Impairment and Blindness*, 101(4):9, 2007.

[31] R. Mehra, A. Rungta, A. Golas, M. Lin, and D. Manocha. WAVE: Interactive Wave-based Sound Propagation for Virtual Environments. *IEEE Transactions on Visualization and Computer Graphics*, 21(4):434–442, Apr. 2015.

[32] T. Nagel. What is it like to be a bat? *The philosophical review*, 83(4):435–450, 1974.

[33] M. Paquier and V. Koehl. Audibility of headphone positioning variability. In *Audio Engineering Society Convention 128*. Audio Engineering Society, 2010.

[34] N. A. Patel and S. J. Patel. Hand gesture recognition system for human computer interaction (hci). 2018.

[35] E. D. Ragan, D. A. Bowman, and K. J. Huber. Supporting cognitive processing with spatial information presentations in virtual environments. *Virtual Reality*, 16(4):301–314, Nov. 2012.

[36] D. Robinson and G. Kearney. Echolocation in virtual reality. In *Proceedings of the 2016 Interactive Audio Systems Symposium*, 2016.

[37] B. N. Schenkman and M. E. Nilsson. Human Echolocation: Blind and Sighted Persons' Ability to Detect Sounds Recorded in the Presence of a Reflecting Object. *Perception*, 39(4):483–501, Apr. 2010.

[38] V. R. Schinazi, T. Thrash, and D. Chebat. Spatial navigation by congenitally blind individuals. *Wiley Interdiscip Rev Cogn Sci*, 7(1):37–58, Jan. 2016.

[39] C. Schissler, A. Nicholls, and R. Mehra. Efficient HRTF-based Spatial Audio for Area and Volumetric Sources. *IEEE Transactions on Visualization and Computer Graphics*, 22(4):1356–1366, Apr. 2016.

[40] D. Schonstein, L. Ferre, and B. F. Katz. Comparison of headphones and equalization for virtual auditory source localization. *The Journal of the Acoustical Society of America*, 123(5):3724–3724, 2008.

[41] T. W. Schubert. A New Conception of Spatial Presence: Once Again, with Feeling. *Communication Theory*, 19(2):161–187, May 2009.

[42] S. Serafin, M. Geronazzo, N. C. Nilsson, C. Erkut, and R. Nordahl. Sonic Interactions in Virtual Reality: State of the Art, Current Challenges and Future Directions. *IEEE Computer Graphics and Applications*, 38(2):31–43, 2018.

[43] S. M. Smith and E. Vela. Environmental context-dependent memory: A review and meta-analysis. *Psychonomic Bulletin & Review*, 8(2):203–220, June 2001.

[44] L. Thaler and M. A. Goodale. Echolocation in humans: an overview. *Wiley Interdisciplinary Reviews: Cognitive Science*, 7(6):382–393, 2016.

[45] L. Thaler, G. M. Reich, X. Zhang, D. Wang, G. E. Smith, Z. Tao, R. S. A. B. R. Abdullah, M. Cherniakov, C. J. Baker, D. Kish, et al. Mouth-clicks used by blind expert human echolocators–signal description and model based signal synthesis. *PLoS computational biology*, 13(8):e1005670, 2017.

[46] E. C. Tolman. Cognitive maps in rats and men. *Image and Environment: Cognitive mapping and spatial behavior*, pp. 27–50, 1973.

[47] A. Tonelli, L. Brayda, and M. Gori. Depth echolocation learnt by novice sighted people. *PloS one*, 11(6):e0156654, 2016.

[48] I. Viaud-Delmon and O. Warusfel. From ear to body: the auditory-motor loop in spatial cognition. *Front. Neurosci*, 8:283, 2014.

[49] T. Virtanen, R. Singh, and B. Raj. *Techniques for noise robustness in automatic speech recognition*. John Wiley & Sons, 2012.

[50] B. N. Walker and J. Lindsay. Navigation Performance With a Virtual Auditory Display: Effects of Beacon Sound, Capture Radius, and Practice. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 48(2):265–278, 2006.

[51] H. Wallach, E. B. Newman, and M. R. Rosenzweig. A precedence effect in sound localization. *The Journal of the Acoustical Society of America*, 21(4):468–468, 1949.

[52] L. Wallmeier, N. Geßele, and L. Wiegrebe. Echolocation versus echo suppression in humans. *Proceedings of the Royal Society of London B: Biological Sciences*, 280(1769):20131428, 2013.

[53] D. Waters and H. Adulula. The virtual bat: Echolocation in virtual reality. Georgia Institute of Technology, 2001.

[54] J. M. Wiener, S. J. Büchner, and C. Hölscher. Taxonomy of human wayfinding tasks: A knowledge-based approach. *Spatial Cognition & Computation*, 9(2):152–165, 2009.

[55] J. M. Wiener, S. J. Büchner, and C. Hölscher. Taxonomy of Human Wayfinding Tasks: A Knowledge-Based Approach. *Spatial Cognition & Computation*, 9(2):152–165, May 2009.

[56] T. Wolbers, J. M. Wiener, H. A. Mallot, and C. Büchel. Differential recruitment of the hippocampus, medial prefrontal cortex, and the human motion complex during path integration in humans. *Journal of Neuroscience*, 27(35):9408–9416, 2007.

[57] K. Woollett and E. A. Maguire. Acquiring "the knowledge" of london's layout drives structural brain changes. *Current biology*, 21(24):2109–2114, 2011.

[58] B. Xie. *Head-Related Transfer Function and Virtual Auditory Display*. J. Ross Publishing, Plantatation, FL, May 2013.