On the improvement of accommodation to non-individual HRTFs via VR active learning and inclusion of a 3D room response

David Poirier-Quinot, Brian F. G. Katz

Acta Acustica 2021, **5**, 25

https://doi.org/10.1051/aacus/2021019

Available online at:
https://acta-acustica.edpsciences.org

SCIENTIFIC ARTICLE OPEN ACCESS

On the improvement of accommodation to non-individual HRTFs via VR active learning and inclusion of a 3D room response

David Poirier-Quinot* and Brian F.G. Katz

Sorbonne Université, CNRS, UMR 7190, Institut Jean Le Rond ∂'Alembert, 75252 Paris, France

Received 7 November 2020, Accepted 14 May 2021

Abstract – This study examines the efficiency of a training protocol using a virtual reality application designed to accelerate individual's selection of, and accommodation to, non-individualized HRTF profiles. This training introduces three elements to hasten audio localization performance improvement: an interactive HRTF selection method, a parametric training program based on active learning, and a relatively dry room acoustic simulation designed to increase the quantity of spatial cues presented. Participants rapidly selected an HRTF (≈5 min) followed by training over three sessions of 12 min distributed over 5 days. To study the impact of the room acoustic component on localization performance evolution, participants were divided into two groups: one acting as control reference, training with only anechoic renderings, the other training in reverberant conditions. The efficiency of the training program was assessed across groups and the entire protocol was assessed through direct comparisons with results reported in previous studies. Results indicate that the proposed training program led to improved learning rates compared to that of previous studies, and that the included room response accelerated the learning process.

Keywords: Binaural, Localization accuracy, HRTF, Learning, Virtual reality

1 Introduction

Binaural synthesis is a signal processing technique used to render spatial auditory scenes over headphones. It relies on the application of direction-dependent audio cues to a monophonic signal, mimicking time and frequency transformations resulting from the propagation of an acoustic wave from a sound source to the listener's ear canals [1, 2]. The technique is used to simulate 3D sounds in practically all wearable augmented and Virtual Reality (VR) systems today.

A recurrent problem in binaural synthesis is that any discrepancy between the simulation and the real phenomenon inevitably impacts the perceived auditory space. A typical discrepancy is the use of direction-dependent cues – referred to as Head Related Transfer Functions (HRTFs) – not measured on the user listening to the rendering. The resulting *non-individualized* synthesis is often the cause of degraded externalization or decreased localization accuracy [3, 4].

Using non-individual HRTFs is common practice as measuring them on a per-user basis is not practical. Methods have been designed to, given a database of existing HRTFs, lead users to select one that minimizes a given criteria, e.g. localization errors [5]. Still, these perceptual *best-match* HRTFs generally result in less precise auditory space perception than individual HRTFs [6]. To further improve one's affinity to a given HRTF, training procedures have been proposed, showing that users could adapt to non-individual HRTFs, exhibiting localization performance approaching that of users relying on individual HRTFs [6, 7].

Paired together, the HRTF selection and training procedures that have been proposed to date generally require users to spend more than an hour to achieve acceptable localization performance (see Sect. 2). The first objective of the present study is to introduce and evaluate a novel HRTF selection and training procedure, designed to accelerate the overall adaptation process. The second objective of the present study is to assess whether the addition of a room acoustic can have a pronounced impact on auditory accommodation to non-individual HRTFs compared to anechoic conditions, given the additional information provided.

A review of previous work is presented in Section 2. The exact scope of the research and the hypothesis under consideration are presented in Section 3. The description of the experiment is presented in Section 4. The results are reported in Section 5, and discussed in Section 6.

*Corresponding author: david.poirier-quinot@sorbonne-universite.fr

2 Previous work

2.1 HRTF selection methods

Various methods have been proposed for HRTF individualization [8]: generating transfer functions using morphological measurements [9, 10], tuning existing transfer functions using subjective criteria [11, 12], selecting HRTF from an existing set based on morphological measurements [13] or subjective ratings [14–16], etc.

The individualization method used in the present study, detailed in Section 4.4, is based on subjective HRTF selection. The concept of these methods is to expose users to stimuli spatialized with various HRTFs and to have them rate which renders best the expected auditory source positions [6]. The principal alternative is objective selection, often taking the form of a localization test [5], where the best HRTF is the one that maximizes participant localization accuracy.

The premise of the present study is that HRTF selection can be reduced to its essential: that given the variance on best-match selection [17, 18], and its limited impact on immediate localization performance [6, 18], it is preferable to capitalize on training to improve localization accuracy. Still, previous studies have shown that using perceptually poorly-rated HRTFs led to a significant increase in adaptation time compared to best-rated ones [6, 19]. The selection method introduced in Section 4.4 has been designed as a tradeoff between these two considerations: providing users with a *"good enough"* match HRTF in a minimum amount of time.

2.2 HRTF learning

2.2.1 Introduction

While localization accuracy is not the only issue resulting from non-individualized rendering [20], the current focus is on this criterion for the following literature review as well as the training program presented in Section 4.6. Readers are referred to Wright and Zhang [21] or Mendonça [22] for more general reviews on the broader topic of HRTF learning.

It has been established that one can adapt to modified HRTFs, e.g. after ear-molds inserted in pinna [7, 23–25], or learn to use non-individual HRTFs [3, 6, 19, 26, 27]. Studies have even shown that one can adapt to distorted HRTFs, e.g. in Majdak et al. [28] where participants suffering from hearing loss learned to use HRTFs whose spectrum had been warped to move audio cues back into frequency bands they could perceive. HRTF learning is not only possible, but lasting in time [19, 26, 29]: users have been shown to retain performance improvements up to 4 months after training [26]. Studies have showed that, given enough time, users using non-individual HRTFs can achieve localization performance on par with participants using their own individual HRTFs [6, 19].

2.2.2 Impact of training protocol parameters

Learning methods explored in previous studies are often based on a localization task. This type of learning is referred to as *explicit* learning [22], as opposed to *implicit* learning where the training task does not immediately focus participant attention on localization cues [6, 19]. Performance-wise, there is no evidence that suggests either type is better than the other. Implicit learning gives more leeway for task design *gamification*. The technique is more and more applied to the design of HRTF learning methods [6, 19, 27, 30], and while its impact on HRTF learning rates remains uncertain [27], its benefit for learning in general is however well established [31]. Explicit learning on the other hand more readily produces training protocols where participants are *consciously* focusing on the learning process [32], potentially helping with the unconscious auditory mental map re-adjustment.

As much as the nature of the task, providing feedback can play an important role during learning. VR technologies are more and more relied upon to increase feedback density in the hope of increasing HRTF learning rates. While Majdak et al. [28] results encourage the use of a visual virtual environment, it has been reported that proprioceptive feedback can equally be used to improve learning rates [6, 33]. There is however a growing consensus on the use of adaptive (i.e. head-tracked) binaural rendering during training to improve learning rates [7, 27], despite the generalized use of head-locked localization tasks to assess performance evolution [22].

Studies on the training stimulus suggest that learning can extend to more than the stimulus used during learning [27]. This result is likely dependent on stimuli relative "*quality*" regarding auditory localization, i.e. whether they present the transient energy and the broad frequency content necessary for auditory space discrimination [34, 35].

There is no clear cut result on optimum training sessions duration and spread. Training session duration reported in previous studies ranges from $\approx$ 8 min [36] to $\approx$ 2 h [28]. Comparative analysis argues in favor of several short training sessions over long ones [22]. Training session spread is also widely distributed in the literature, ranging from all sessions in one day [35] vs. one every week or every other week [19]. Where Kumpik et al. [35] results suggest spreading training over time benefits learning (all in one day vs. spread over 7 days), direct comparison of Stitt et al. [19] and Parseihian and Katz [6] suggests that weekly sessions and daily sessions result in the same overall performance improvement (for equal total training durations). There is some example of latent learning (improvement between sessions) in the literature [36], naturally encouraging the spread of training sessions. Regardless of duration and spread, studies have shown that learning saturation occurs after a while. In Majdak et al. [37], most of the training effect took place within the first 400 trials ($\approx$160 min), a result comparable to that reported by Carlile et al. [38] who reached saturation after 7–9 blocks of 36 trials each.

One of the critical questions not fully answered to date is the role of the HRTF in the training process. It would appear that a certain degree of affinity between a participant and the training HRTF facilitates learning [6, 19]. The question remains whether this affinity also furthers the saturation point.

2.3 Impact of room response on localization accuracy

Previous studies have examined the impact of various degrees of acoustic conditions on binaural cues, such as ground reflections for small animals [39], or subjective perception of an auditory scene [40]. There are surprisingly few studies which have examined the impact of room acoustics on localization in particular, with tested conditions often being quite limited, and results varying between studies. The room acoustic response can be summarized by the term *reverberation,* used here in its broadest sense, pertaining to single or multiple early reflections, late reverberation, or combinations of both.

In an early study, Hartmann [41] examined the effect of Reverberation Time (RT) and ceiling height on localization of distant real speakers (12 m) in the frontal-horizontal plane only using a 500 Hz tone burst. Using the ESPRO variable acoustics facility at IRCAM, RT of 1 s and 5.5 s were compared, as well as a lowered ceiling condition with an RT of 2.8 s, the later changing both the geometry and RT. Results showed no effect between RT conditions, while the changes resulting from the lowered ceiling condition **improved** azimuthal localization performance. In contrast, Rakerd and Hartmann [42] tested specific single reflection directions compared to anechoic conditions, showing that a lateral reflection **decreased** azimuthal localization performance. Subsequently, Guski [43] also focused on single reflections in a dry room (RT of ≈ 0.5 s), and showed that the impact of a lateral reflection on localization performance was not systematic, depending on reflection position relative to the source, while a floor reflection improved localization. Results also suggested that the presence of a floor reflection **benefited** elevation localization performance where a ceiling reflection (1 m above listener) **impeded** overall localization accuracy, contrary to Hartmann [41]. Using a spherical array of speakers for individual reflection directions and a horizontal array for late reverberation (reverberator units with broadband RT of 0.4 s), Bech [44] examined the level and decay thresholds of reflections inducing coloration or image shift of a target sound source. Results were analyzed and discussed relative to previous studies, including those on the precedence effect.

Taking advantage of virtual reality audio binaural synthesis, Begault [45] studied the addition of two spaced floor reflections, 64 early reflections calculated from a 2D ray-tracing simulation of a reversefan shaped room, and a diffuse field late reverberation, using the same HRTF across subjects. Tested speech stimuli on the horizontal plane showed **no effect** of these additions on azimuthal performance, while a general vertical bias was observed across subjects in the reverberant condition. While no explanation was proposed for said bias, one can hypothesize that the unrealistic reflection conditions (both in 2D and the limited room modeling of the time) could induce perceptual errors. Testing localization ability over repetitions, Shinn-Cunningham [46] employed noise bursts at a distance of 1 m in a real room (RT of 0.5 s). Results showed improvement in localization performance with exposure through test repetitions (contrary to results of Hartmann [41]) for seven subjects. While mean elevation performance was judged slightly poorer than other published studies in anechoic conditions, cross-comparisons of such studies with so few subjects are difficult to interpret.

Further exploring binaural synthesis conditions, a pair of studies tested horizontal source positions using a 3D ray-tracing room model (dynamic reaction to head-rotations) combined with a static late reverberation with a RT of 1.5 s for individual and generic HRTFs [47, 48]. Results showed **reduced** azimuthal error in the reverberant condition along with a vertical bias shift in localization for speech stimuli, presented solely in the horizontal plane.

Similar to the above previously mentioned works, Angel et al. [49] examined the effect of adding a single floor reflection in the context of binaural synthesis. Localization of a 1 s amplitude modulated noise burst was evaluated using individual HRTFs and a diffuse late field reverberation (RT of 0.5 and 1 s) for sources either on the horizontal plane or along the 45° cone of confusion.[1] **Improvements** were observed regarding front–back confusions on the horizontal plane for 6-out-of-9 subjects, as well as a reduction in azimuthal errors. **No effect** was observed on elevation errors for the cone of confusion condition, or between late reverberation time conditions. The inclusion of head-tracking, providing dynamic rendering of acoustic cues with changes in head orientation, was determined more beneficial than the inclusion of the floor reflection.

More recently, Nykänen et al. [50] examined the impact of reverberation using recordings made via dummy-head in a real classroom with speech and noise burst stimuli on the horizontal plane only. Results showed that front–back confusion rates were *lower* in the reverberant conditions compared to anechoic conditions.

In traversing these previous studies, the results are rather mixed, if not contradictory. However, it would appear that early arriving reflections with the same azimuth (floor/ceiling) as the direct sound tend to improve lateral localization performance. Results are inconclusive with regards to improvements in elevation localization performance. Tested source distances ranged from 1 m to 12 m, with source positions being almost exclusively limited to the horizontal plane and at a limited number of positions. Test stimuli comprised speech, tone bursts, or noise bursts of various duration. In virtual conditions, HRTFs ranged from a single "generic" HRTF for all subjects to individual measured HRTFs with no observed effects.

3 Aim and scope of the research

The current study investigates the efficiency of a novel (1) HRTF selection and (2) training program, as well as (3) the impact of a virtual room acoustic simulation on

[1] A cone-of-confusion is defined by the locus of positions having the same inter-aural time difference, thereby limiting localization cues to spectral changes only [51]. Typical localization errors are front/back or up/down confusions, where the perceived source position is mirrored across symmetry planes. Left/right confusions are non-existent in listeners without substantial hearing loss in one ear.

HRTF learning rate. All three components have been designed as part of a solution to achieve non-individual HRTF accommodation and subsequent audio localization proficiency in a minimum amount of time and under unsupervised conditions.

The HRTF selection method was designed in the hope of providing participants with a good enough perceptual fit to accelerate the training process [6] while requiring as little of their time as possible. The learning program was designed to be efficient and entertaining, informing participants on the problems of non-individual binaural rendering and providing exercises to overcome them. Finally, the room acoustic response simulation combined with the anechoic rendering was proposed under the hypothesis that multiplying position-dependent spatial cues would further participants' ability to adapt to a new HRTF. The simulated acoustic space was not associated with the visual scene presented, to avoid multimodal interactions. A neutral virtual visual scene was presented during the experiment, as the intent was to examine the impact of reverberation as a **scene-agnostic localization enhancer**. The term *self-coherent* room response is hereafter used to refer to this paradigm.

To characterize the efficiency of the HRTF selection and training program, the experimental protocol has been conceived to enable comparative analysis with previous studies [6, 19]. Learning is assessed based on localization performance for sources on the whole sphere, not constrained to frontal directions or horizontal plane, using an ego-centered response interface in VR. To characterize the impact of the room acoustic simulation on HRTF adaptation, a control group was constituted with participants training under anechoic condition.

The hypotheses to be tested are as follows:

H1 Active involvement of participants during the HRTF selection procedure reduces initial (a) localization errors and (b) confusion rates.

H2 A self-coherent room response reduces (a) localization errors and (b) confusion rates prior to training.

H3 Initial participant performance predicts degree of improvement.

H4 A self-coherent room response improves HRTF learning rate evaluated via (a) localization errors and (b) confusion rates.

H5 The proposed HRTF selection method is at least as good (regarding localization accuracy with the selected HRTF) as those proposed previously in the literature.

H6 The proposed training program improves learning rates compared to previously proposed methods.

4 Experimental design

4.1 General experiment description

A total of 24 adults participated in the experiment (age 19–64 years, mean 29 ± 2 years, 7 women), none self-reported any hearing deficit. Participants first selected their *best-match* HRTF from a pre-existing set, based on the selection process described in Section 4.4. They then proceeded to a first localization task, described in Section 4.5.

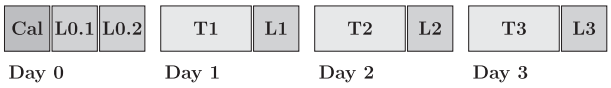

Figure 1. Schematic timeline of the experiment sequence. Li are localization tasks, Ti are training tasks. An average Ti + Li lasted ≈ 20 min (all but Day 0).

The complete experiment sequence is illustrated in Figure 1. All participants performed the first localization task L0.1 in the anechoic condition. They were then evenly distributed into two groups: G-anech or G-reverb, performing the second localization task L0.2 in either anechoic or reverberant conditions. Group assignment was designed so that initial performance of both groups (L0.1) was evenly balanced regarding analysis metrics described in Section 5.1. Participants then underwent three training sessions, each separated from the next by 1–3 nights, each followed by a localization task for performance assessment. Localization tasks lasted ≈ 5 min, training sessions lasted exactly 12 min, as per [6, 19]. To serve as incentive, the experiment was staged as a contest where for each group, participants ranking first and second would receive €250 and €150, respectively. A video illustrating the experiment is available online.[2]

4.2 Test interface and binaural rendering

The experiment was conducted in an acoustically damped and isolated room (ambient noise level < 30 dBA). Participants were equipped with a tracked head-mounted display (Oculus CV1), open circumaural reference headphones (Sennheiser HD 600), and a pair of hand tracked controllers (Oculus Touch). This setup provided tracking information for both head and hands as well as presentation of the test interface to the participant via the visual display. Tracking latency (<6 ms) and precision (≈1 cm) were sufficient for use in this study [52, 53]. The virtual scene and user interface were designed and ran in Unity v2017.3, rendered at a frame-rate of ≈ 90 fps. The entire experiment was conducted on a PC running a 64-bit Windows 10 on a 3.6 GHz Intel Core i7-7700 CPU with 64 Go of RAM coupled to an NVIDIA GeForce GTX 1080Ti graphic card. After setup, the entire experiment was user guided with the test administrator only present in case of difficulties.

Anechoic binaural rendering was performed using the Anaglyph binaural audio plugin v0.9.3b, embedded in a Cycling '74 Max v8.0.5 patch. Anaglyph uses HRTF convolution with a variable delay allowing for customization of Interaural Time Differences (ITD) through a personalizable morphological model. For improved suitability for sources in the near-field, frequency dependant Interaural Level Difference (ILD) correction is applied using a spherical head shadowing model in addition to HRTF parallax correction,

[2] Experiment video: https://youtu.be/j2FFSz-h_Jo.

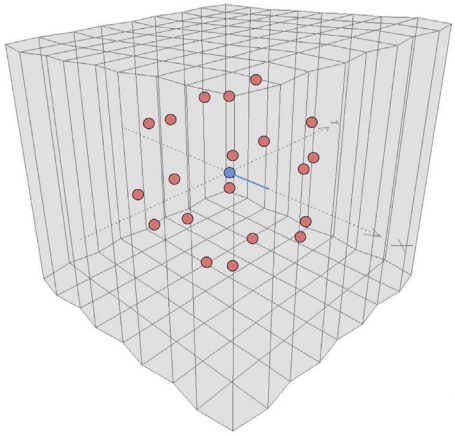

Figure 2. (Color online) Room model used to simulate the 3D room response. ● Source, ● Listener.

Table 1. Reverberation T30 across frequency bands (direct path included in analysis) for the reverberation condition.

Freq. band (Hz)	63	125	250	500	1 k	2 k	4 k	8 k
T30 (s)	0.14	0.19	0.20	0.14	0.15	0.13	0.13	0.16

providing the correct measured HRTF position filter for each ear independently. These corrections are only meaningful for sources at a radius of under 1.95 m (distance for which the HRTFs used in the present study were measured) i.e. only apply to the hand-held probe used during the training task (see Sect. 4.6). Full details of the functionality of the binaural engine are available in Poirier-Quinot and Katz [54]. Audio-visual latency was below the ≈ 15 ms threshold of detectability [55].

4.3 Simulated room reverberation

To provide a self-coherent room reverberation condition, a Geometrical Acoustics (GA) room simulation was used to generate physically realistic Ambisonic Room Impulse Responses (RIRs), using CATT-Acoustic v.9.0. c:3/TUCT v1.1a:4. This GA software has been previously shown to be capable of generating comparable spatial RIRs when subjectively compared to measured data following a variety of perceptual attributes [56].

The reverberation condition employed a convolution with a second-order Ambisonic RIR with the direct sound contribution removed, as **the direct sound was rendered separately as per the anechoic binaural condition**. RIRs were simulated for 20 source positions, uniformly distributed on a 1.95 m radius sphere around the receiver. Second-order Ambisonic as well as the RIR grid density were adopted as a trade-off between spatial precision and processing power requirement, drawing on results from previous studies [57–60]. The room was based on a 5.7 × 5.7 × 5.4 m^3 cube, with shaping to avoid flutter effects, and a slight incline of the ceiling angle, illustrated in Figure 2. Ambisonic RIRs were simulated using the following settings

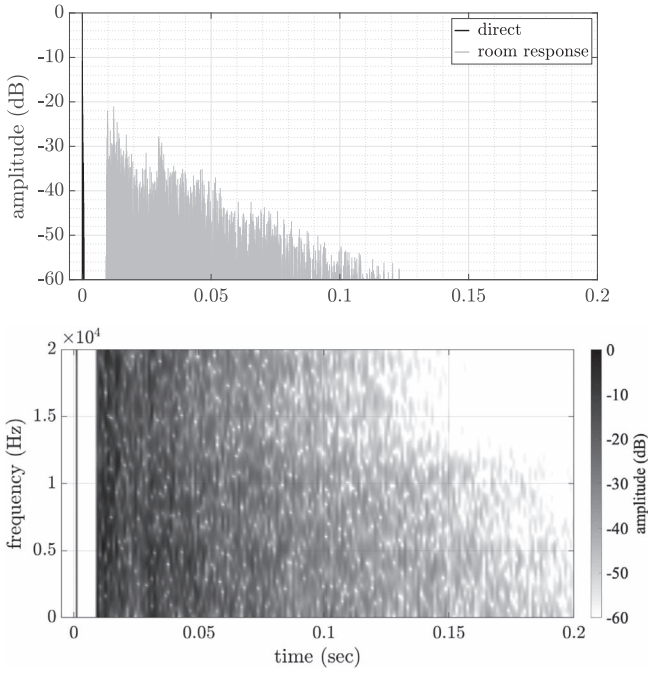

Figure 3. Example temporal (top) and frequency (bottom) response of the numerically simulated room acoustic for a source in front of the listener. The "direct" component has been added for reference, not present in the RIR used for the reverberation.

in the GA software: Algorithm 1: Short calculation, basic auralization and max split order 1 with 150 000 rays. Resulting RIRs had an RT (T30) of ≈ 0.15 s across all frequency bands (see Tab. 1). An example of the rendered RIR is shown in Figure 3, highlighting the density and decay rate of the room response, further emphasizing the continuity of reflections from the early to the latter part of the response when a realistic condition is considered. A spectrum highlighting the temporal-frequency characteristics of the RIR is also provided, highlighting the general uniformity and lack of a pronounced coloration effect. The use of a relatively short reverberation time was deemed an obvious requirement for inclusion in a generalized VR environment. If the reverberation was too pronounced, then it would contribute a noticeable reverberant effect to whatever VR scenario would be created by VR designers, and would be accumulated with any reverberator audio production effects associated to the actual VR scene. As such, this study examines the use of what could be termed a **subtle** room effect, so as not to overshadow, interact, or more importantly detract from the acoustic precepts of a VR scenario environment.

For a given position to spatialize during the localization task, an RIR was constructed on the fly in the rendering engine from a linear interpolation between the three nearest simulated RIR positions before convolution with the monophonic input stimulus signal. The resulting Ambisonic stream was then decoded using the virtual speaker approach [61] with 12 virtual speakers uniformly distributed on a sphere. The binaural encoding of each virtual speaker signal was also performed within the Anaglyph plugin, using the same ITD-corrected HRTF as for encoding

the direct/anechoic path. The final signal presented during the reverberation condition was the sum of both direct and reverb streams.

4.4 HRTF selection method

Participants were situated in a virtual scene, facing a 7-element menu of available HRTFs. The presented HRTFs were a "subjective orthogonal subset" [62, 63] of the LISTEN database [64]. In each hand, they held a small taser (virtual representation of hand-tracked VR controllers) that they could activate to create an audio-visual spark. After a general introduction to the shortcomings of nonindividualized binaural rendering, participants were left to explore various sound source positions around their head with the taser. Their instructions were to select the HRTF that minimized overall spatial incoherence (confusions, misalignment, etc.) based on the auditory, proprioceptive, and visual perception of the taser position in the hand. No other stimuli were present during this task.

HRTFs of the LISTEN database are composed of 187 pairs of impulse responses, measured on a 1.95 radius sphere at 15° azimuth/elevation intervals, with a gap on the southern hemisphere below −45°. HRTF ITDs were adjusted based on participant head circumference to focus the selection on fine direction-dependent cues. The taser audio stimulus was a sequence of three bursts of white noise, 40 ms each with a 4 ms cosine-squared onset/offset ramp and a 70 ms inter-onset interval for a total duration of 180 ms. Full spectrum burst sounds were chosen in order to favor an adaptation to the complete spectral cues of the HRTF [47].

4.5 Localization task

Participants were situated in a virtual scene, facing a visual anchor, surrounded by a 1.95 m radius semi-transparent sphere. The sphere was lightly textured to facilitate spatial memory and provide an adequate frame of reference during the localization task [37]. Fixing one's head position and looking at the visual anchor triggered a circular progress bar, loading for ≈ 1 s before triggering an audio stimuli. The stimulus was the same noise burst as in Section 4.4. Paired with the visual anchor mechanism of fixated gaze/-head-orientation, the short audio stimulus prevented participants initiating head movement during presentation of the target stimulus [65]. The stimulus was randomly spatialized at one of 20 potential positions, evenly distributed on a sphere. Each position was repeated three times for a total of 60 localization trials. Participants used a pair of handheld *blasters* to indicate the perceived direction of origin of the audio source on the surrounding sphere, inspired from the "manual pointing" reporting method assessed in Bahu et al. [66]. When activated (trigger button) a blaster would shoot a small red ball onto the semi-transparent sphere that would correspond to participant perceived direction of origin. To facilitate aiming, each blaster was equipped with a laser sight. When evaluated on visible targets during debug sessions, the overall setup resulted in pointing errors below 2°.

4.6 Training task

The gamified training task resembled a hide-and-seek game, where participants had to identify a hidden audio source among visual potential targets, designate it, and repeat, with different degrees of complexity during the course of training. Upon startup, participants were presented with a selection menu from which they could enter any of the already *unlocked* training scenarios. A training scenario was composed of a predefined number of trials, each trial starting with the creation of at least two visual potential targets (white spheres in Fig. 4), and ending with participants designating one of them as the hidden audio source or *active* target. After facing all the trials of a given scenario, participants were returned to the scenario selection menu. Training automatically ended after the allotted 12 min. The training VR scene situated the participant on top of a 2 m radius platform, surrounded by a 360° sky dome, providing a frame of reference as for the sphere texture of the localization task.

During a trial, participants were free to look around to identify all the visual potential target positions of that trial. The active target would remain silent except when they looked directly at the visual anchor (green sphere in Fig. 4) and stood at the center of the platform, thus preventing the use of head-movements for localization. In their hands, participants held the same pair of tasers of Section 4.4. The primary use of the tasers was to serve as spatial audio probes: when hesitating between two or more visual target positions, participants could move a taser (i.e. a hand) towards each of them and trigger it at will to compare its sound to that of the hidden audio source. Both taser and hidden audio source used the same noise burst as in Section 4.4. This comparison mechanism was deemed an **essential** component of the training, leading participants to carefully listen to audio spectral cues, reinforcing sound-to-position relationships via proprioception and visual feedback. The taser also served as a selection pointer, with which participants could designate which visual target they thought was the active one. Upon visual target selection, the true active target emitted a sound, indicating if the choice was correct as well as revealing its position.

A total of 14 training scenarios were designed, divided into four "difficulty levels", representing increasingly more complex scenarios. Scenarios from the first difficulty level each focused on introducing one of the specific known issues of non-individualized binaural rendering (front–back confusions, cone of confusion, angular resolution, etc.). Subsequent scenarios re-exposed these problems, further increasing in difficulty. A complete list of all training scenarios as well as specifics on level design mechanics are provided in Appendix.

5 Results

Section 5.1 presents the analysis tools and metrics employed in quantifying the results. Section 5.2 examines the degree of participant active involvement in the HRTF

Figure 4. (Color online) Training setup: (top-left) participant in the experiment room, (bottom-left) third person view of the training platform, (right) participant viewpoint during the training.

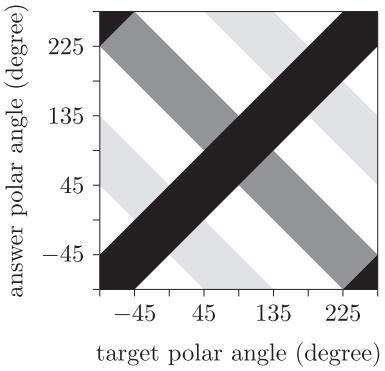

Figure 5. Definition of the 4 different cone-of-confusion response classification zones, from (5). ■ *Precision*, ■ *Front-Back*, ■ *Up-Down*, □ *Combined*.

selection task. Section 5.3 asserts initial group performance equivalence during the first anechoic localization task L0.1. The impact of reverberation on localization accuracy prior to learning (L0.1 vs. L0.2) is assessed in Section 5.4. Participants initial performances are correlated to their relative improvement in Section 5.5. The impact of reverberation on learning rate (L1 to L3 progression) is assessed in Section 5.6. The impact of the novel proposed learning paradigm, comparing the results of G-anech with that of previous studies, is assessed in Section 5.7.

5.1 Performance metrics

Analysis of localization performance was conducted using the interaural polar coordinate system [67], allowing for a rough separation of the role of different interaural cues throughout the analysis [19]. Two types of metrics were used during the analysis: *angular* and *confusion* metrics, all computed by comparing participant responses against target true position.

The four angular metrics considered were the interaural lateral and polar errors, the great-circle error, and the *folded* interaural polar error [19]. Lateral (resp. polar) error is the absolute difference between response and true target lateral (resp. polar) angle, wrapped to [−180: 180]. Great-circle error is the minimum arc, expressed in degrees, between response and true target position:

$$\text{great circle} = \arctan\left(\frac{\left\|xyz_{\text{true}} \times xyz_{\text{response}}\right\|}{xyz_{\text{true}} \cdot xyz_{\text{response}}}\right).$$

Folded polar error is computed after compensation for any observed front–back or up–down confusions, applying a front–back (resp. up–down) axis-symmetry to the response position before calculating the polar error. Lateral and polar errors analyze specific aspects of the responses, highlighting what types of errors are occurring, while the basic global great-circle error indicates the overall magnitude of errors. It is understood that some metrics are derivatives of others, or not totally independent.

Polar angle confusions were classified using a traditional segmentation of the cone of confusion (see [6, 19]), (revised in [5]). The classification results in three potential confusion types: front–back, up–down, and combined, with a fourth type corresponding to precision errors, represented schematically in Figure 5. The *precision* category designates any response close enough to the real target so as not to be associated to the other confusion types.

The impact of independent variables (reverberation condition, session, etc.) on performance metrics was assessed using a Friedman test, as most distributions proved to follow non-normal (skewed) distributions.

Post-hoc paired-sample comparison was based on a Wilcoxon signed rank test for angle error distributions, and a chi-squared test for confusion rate distributions. Statistical significance was determined for p-values below the 0.05 threshold, the notation $p < \epsilon_1$ is adopted to indicate p-values below 10^{-3}. Reported p-values are those of post-hoc tests. Effect size d is reported for pairs of distributions with p-values below 0.05, using Cohen and Phi statistics for angle error distributions and confusion rate distributions respectively. The notation $d < \epsilon_2$ is adopted to indicate effect sizes below 0.1 (i.e. when two group means differ by < 0.1 standard deviation). Significant differences between fits pertaining to the analysis of learning rates in Section 5.7.2 were assessed based on comparisons of their coefficients. Significant difference is discussed when at least one of the coefficients of two fits differ beyond 50% of their estimate's 95% Confidence Interval (CI) [68]. The symbol "±" represents standard error throughout the paper.

5.2 Effect of HRTF selection task behavior on initial performance (H1)

Participant initial performances were compared to their behavior during the HRTF selection task, to assess whether the degree of active involvement in the task translates to improved initial localization performances. Participant involvement was quantified by examining several "*behavioral*" metrics: number of times the acoustic probe was activated, portion of the sphere explored, number of times the active HRTF was changed, and the duration of the selection task. The metric "sphere coverage" is a

Table 2. Correlation coefficients between participant behavior during the HRTF selection task and initial localization performance. Coefficients whose magnitudes are <0.15 are not reported.

	Num. probings	% sphere coverage	Num HRTF switches	Duration
Great-circle	−0.19	0.15	•	•
Polar	−0.18	•	•	•
Lateral	•	−0.22	•	•
Front–back	−0.30	•	−0.18	•
Up–down	•	•	0.16	•
Combined	0.19	•	•	0.18

Table 3. Groups initial performance (L0.1, anechoic condition). Bold text indicates significant difference between groups.

	Absolute angular error mean values (degrees)			
	Great-circ.	Lateral	Polar	Polar fold.
G-anech	46.4 ± 1.3	15.7 ± 0.5	65.6 ± 1.9	47.5 ± 1.7
G-reverb	45.1 ± 1.3	12.9 ± 0.4	63.3 ± 2.0	45.0 ± 1.7
p-value	0.131	$p < \epsilon_1$	0.158	0.061
	Interaural confusions error classifications (%)			
	Front–back	Up–down	Combined	Precision
G-anech	15.3	8.8	25.8	50.1
G-reverb	17.4	5.8	21.5	55.3
p-value	0.285	0.033	0.055	0.051

percentage, computed based on how many *"regions"* of the sphere were probed, where regions are defined as Voronoi cells around the 20 potential target positions of the localization task. Correlation coefficients between these metrics and initial L0.1 performances are reported in Table 2. Results show the magnitude of correlations never exceeding 0.30, indicating weak to no correlation. These results do not support **H1**.

5.3 Group baseline: initial performance comparison

Following the initial localization test results in L0.1, participants were divided into 2 groups, evenly matching great-circle and front–back confusion errors. There remained however a slight mismatch between group means at the start for the metrics lateral and up–down confusion errors. This small yet significant offset, was in favor of G-reverb (difference of $\approx 3°$ and 3% resp.). Detailed group baseline performance during L0.1 are reported in Table 3 and shown in Figure 6.

5.4 Instantaneous impact of room response on performance (H2)

Figure 6 illustrates the evolution of group performance from L0.1 to L0.2, where G-reverb started using the reverberation while G-anech remained in anechoic conditions, prior to any training. A slight overall improvement trend can be observed, likely due to procedural training. However, neither group showed any significant improvement on angular metrics between the two sessions. Baseline group mean differences observed in L0.1 on lateral errors and up–down confusions were also present in L0.2.

G-anech confusion rates did not significantly improve between the two sessions. In contrast, G-reverb combined confusion rate improved from 21.5% to 17.4% ($p = 0.046$). Consequently, G-reverb combined confusion rate was significantly below the 22.6% rate of G-anech in L0.2 ($p = 0.012$). No other differences were observed for G-reverb confusion rates between sessions L0.1 and L0.2.

Further L0.2 comparisons revealed that G-reverb front-back confusion rates were significantly higher than that of G-anech (21.0% vs. 14.2% resp., $p < \epsilon_1$). In the absence of intra-group performance evolution between these two sessions, this result only hints at reverberation having a negative impact on immediate front-back confusions. This observation could also simply be the result of the confusion classification method "*redistributing*" G-reverb errors from combined and up–down confusions (decreasing) to frontback confusions (increasing) between the two sessions.

5.5 Predicting relative improvement based on initial performance (H3)

Participant relative improvement between L0.2 and L3 as a function of initial L0.2 performance is reported in Figure 7. Both data sets are strongly correlated for lateral angle, front back, and up–down confusion errors ($|r| > 0.5$), arguing in favor of H3. Additionally, scattered values show the overall even distribution of participants improvement during training. Negative improvements on front–back and up–down confusions, i.e. points with y-axis values above zero, are a result of the confusion classification method redistributing errors between categories, as the sum of the 4 categories always tabulates to 100%. Participants did

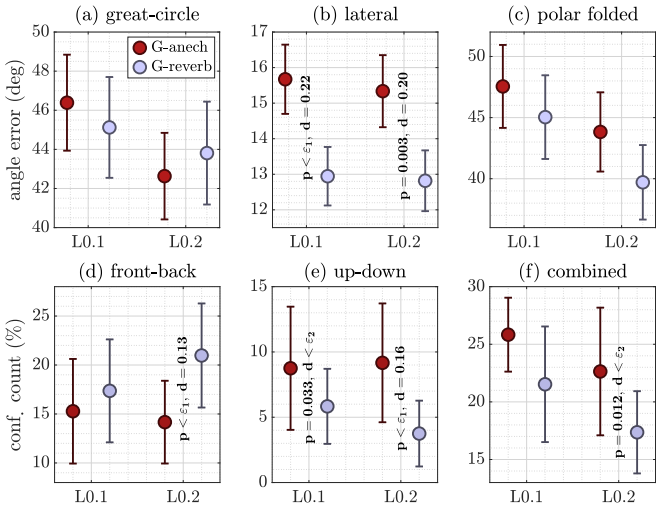

Figure 6. (Color online) Comparison of mean and 95% Confidence Interval (CI) of angular and confusion errors between sessions and groups for the first two sessions. 720 pts per error bar for angular errors. 12 pts – one percentage value per participant – per error bar for confusion errors. Reported p-values indicate significant differences between sessions. For confusion errors, significance assessment is performed on a per-trial basis (i.e. not per-participant, thus providing 720 Boolean values per distribution). For reference, precision confusion scores in *L0.2* increased (non-significantly) compared to those of *L0.1* reported in Table 3: to 54.0% and 57.9% for groups *G-anech* and *G-reverb* resp.

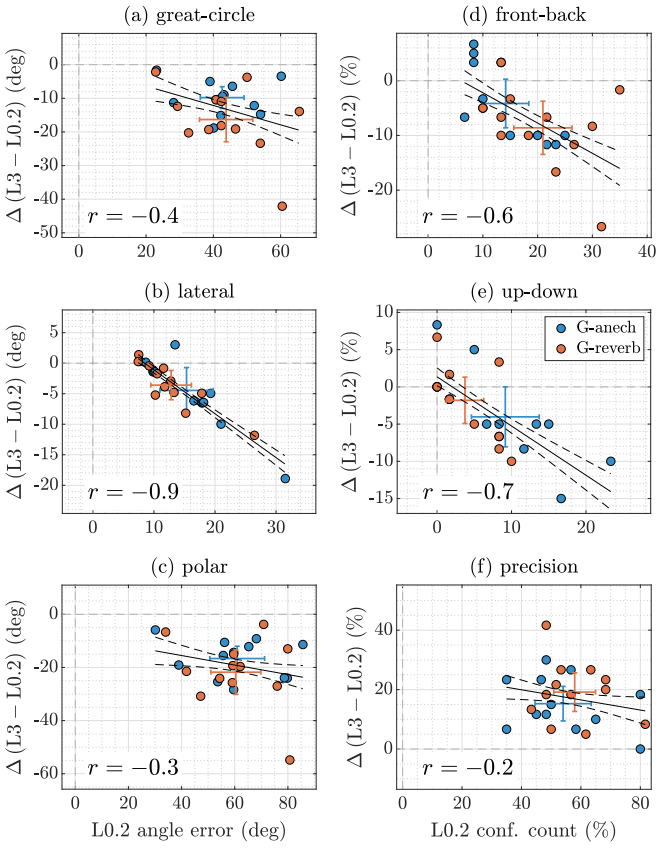

Figure 7. (Color online) Participants relative Δ (L3 – L0.2) performance improvement as a function of their initial L0.2 performance on (a–c) angular and (d-f) confusion errors. The overall linear regression across both groups is shown, with correlation coefficient, r. Error bar lengths and positions indicate 95% CI and means (12 pts per value) for each group.

improve overall with regards to confusions after training, as illustrated by the increase in precision confusion scores in Figure 7f.

5.6 Impact of room response on performance evolution (H4)

Groups performance evolution from Day 1 to Day 3 on both types of metrics is reported in Figures 8 and 9. For both groups, a steady and significant decrease can be observed on all angle errors throughout training. Less clear-cut, the overall improvement on confusion errors is still readily apparent.

For all but front–back confusions, G-reverb consistently outperformed G-anech from L1 onwards. Setting aside those two metrics for which initial group performance were not evenly balanced (lateral errors and up–down confusions), results strongly suggest that the inclusion of the room response had a positive impact on performance evolution during training.

5.7 Impact of HRTF selection method and training program: comparison to previous studies (H5 & H6)

Results of the current study are compared to those of three previous experiments concerned with HRTF selection and HRTF training in anechoic conditions. The following brief descriptions for each of those experiments highlights those elements relevant to the current comparison (while not covering the entirety of those studies).

- Exp-Parseihian [6]. Study on HRTF accommodation in which test group G3, composed of 5 participants, trained with their best-match HRTF, selected from the same subset as the one used in the present study. The selection was based on participant ratings of the precision of audio trajectories created with each HRTF, referred to as the trajectory method. Training consisted of a 12 min game, inducing HRTF adaptation by coupling proprioceptive feedback and passive listening. Designed to be compatible with studies with visually impaired individuals, no visual interface or feedback was provided. Training schedule supposed one training session per day for three consecutive days. Each training session was followed by a performance evaluation based on a localization task. G3 participant performance evolution was compared to that of G1, a control group consisting of 5 participants, also using their best-match HRTF but undertaking only the first training session.
- Exp-stitt [19]. Study on HRTF accommodation in which test groups W4 and W10, composed of a total of 16 participants, trained with their worst-match

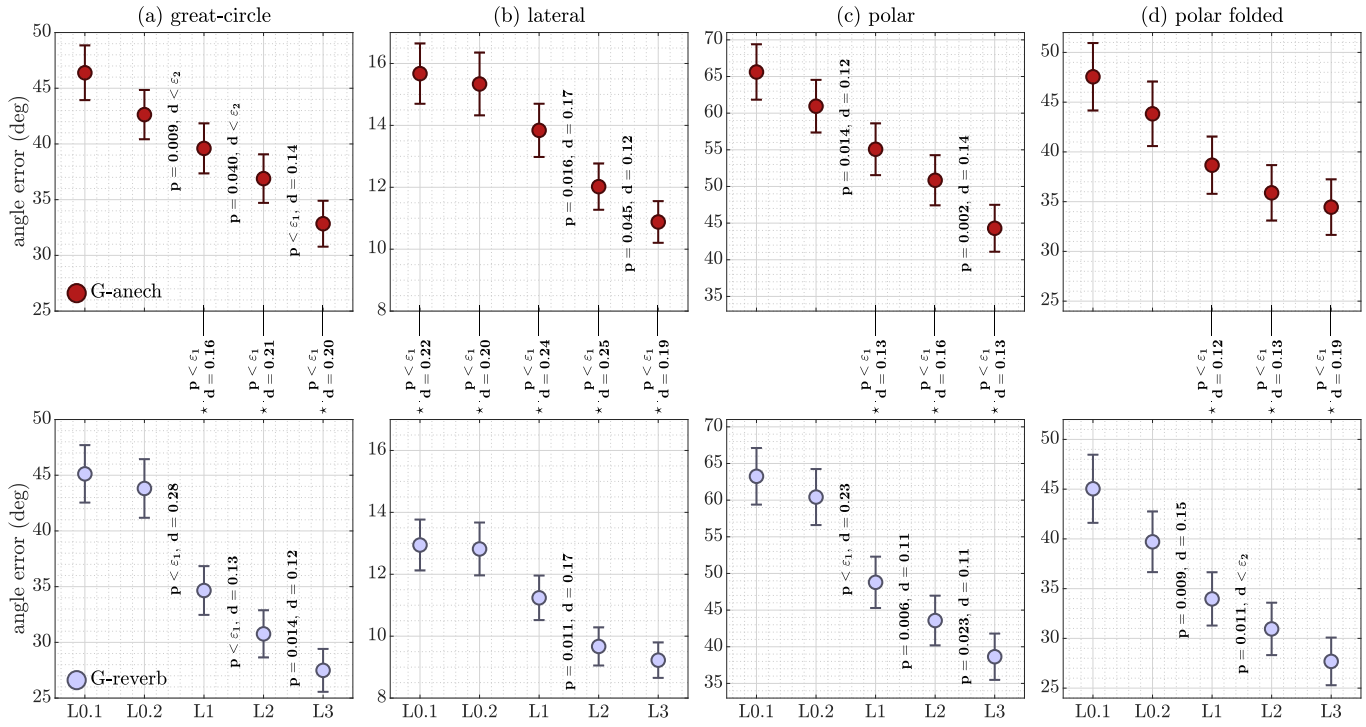

Figure 8. (Color online) Comparison of mean and 95% CI of angular errors between sessions and groups throughout the whole training (720 pts per error bar). The * symbol indicates the distribution with the lowest mean.

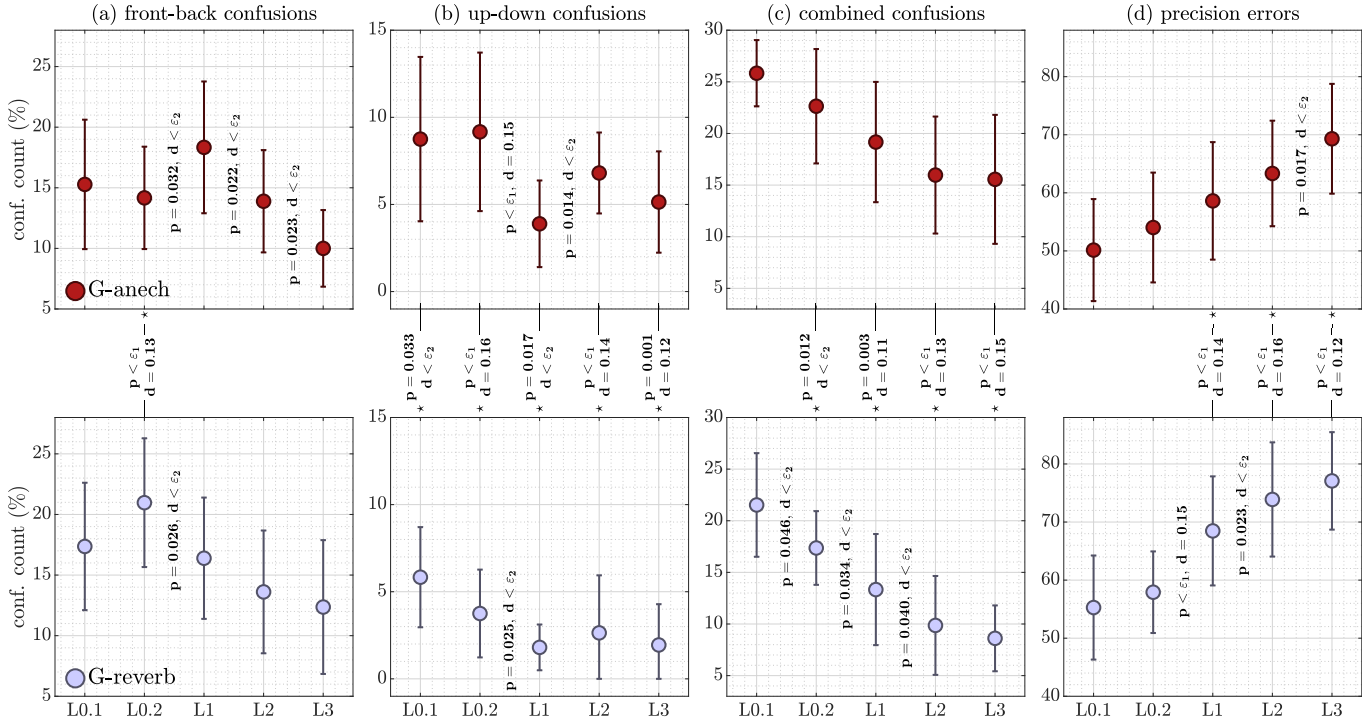

Figure 9. (Color online) Comparison of mean and 95% CI of confusion error classification percentages between sessions and groups throughout the whole training (12 pts – one percentage value per participant – per error bar, 720 pts per distribution for significant difference assessment). The * symbol indicates the distribution with the lowest mean.

Table 4. Mean duration of the HRTF selection sequence reported in each experiment. HRTF selection duration as reported in similar studies have been added for reference. Methods marked $\frac{1}{2}_{3} \equiv$ provide a full ranking of an HRTF set (e.g. trajectory method), the remainder are designed for the selection of a unique best-match HRTF (e.g. active method). Indicated durations have been divided by the number of task repetitions compared to total reported duration in each experiment (e.g. 3 in exp-zagala) for accurate comparison.

Experiment	Method	HRTFs	Duration
Exp-Parseihian	Trajectory method $\frac{1}{2}_{3} \equiv$	7	Not reported
Exp-stitt	Trajectory method$\frac{1}{2}_{3} \equiv$	7	Not reported
Exp-zagala	Trajectory method$\frac{1}{2}_{3} \equiv$	8	9.0 min
Exp-zagala	Localization method$\frac{1}{2}_{3} \equiv$	8	8.2 min
Exp-current	Active method	7	4.7 min
Andreopoulou and Katz [63]	Trajectory method$\frac{1}{2}_{3} \equiv$	24	26.0 min
Andreopoulou and Katz [17]	Trajectory method$\frac{1}{2}_{3} \equiv$	7	6.5 min
Iwaya [16]	Tournament method	32	15.0 min

HRTF once a week for 3 weeks. But for the 1 week inter training session gap, this experiment used the same protocol as exp-Parseihian (HRTF selection method, training game, HRTF subset, etc.).

- Exp-zagala [5]. Study on HRTF quality evaluation in which 28 participants' best-match HRTFs were determined based on two different methods. The first method was the trajectory method, described above. The second relied on the results of a basic localization task, similar to that described in Section 4.5, referred to as localization method. Seven of the 8 HRTFs evaluated were from the same subset as that used in the present study. Each task was repeated three times to examine repeatability.

5.7.1 Impact of the HRTF selection method on initial performances (H5)

The average duration of the HRTF selection stage for each experiment is reported in Table 4. The selection method used in the present study, referred to as the active method, required on the order of half the time required by the other three (based on the duration reported in exp-zagala). It must be noted that this method was designed to select only a unique bestmatch HRTF, while the other methods yield a full ranked list of the presented HRTFs.

Figure 10 illustrates the localization performance of participants immediately after the HRTF selection in all three experiments. Participants using the proposed active method significantly outperformed those using the trajectory method, but for the front-back error rate of group zagala traj. As detailed in Figure 10, results related to the group zagala traj are presented on a per-metric best-case outcome of the two tested methods. Participants using the proposed active method nevertheless significantly outperformed previous cited studies on initial great-circle, up–down confusion, and combined confusion errors.

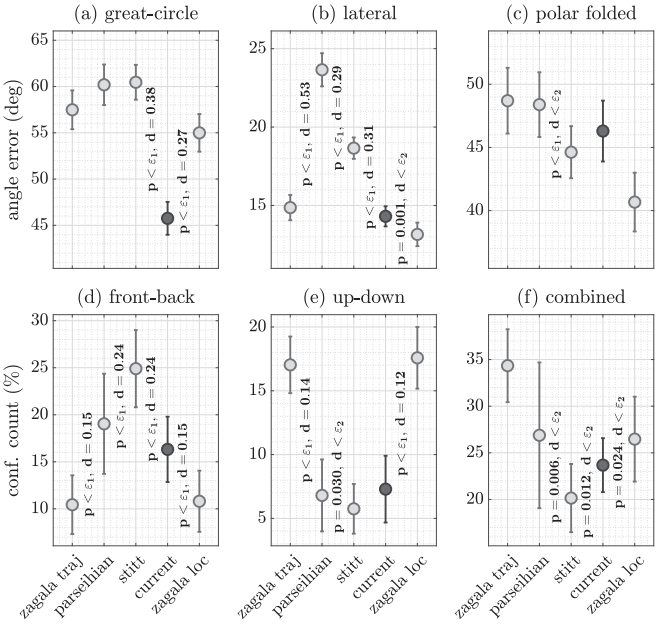

Figure 10. Comparison of mean and 95% CI of angular and confusion errors between experiments (HRTF selection methods). Presented results correspond to the first session of each experiment. Parseihian presents the results of groups G1 and G3 in exp-Parseihian. Zagala traj presents the results of participants during the localization task using their best-match HRTF in exp-zagala. zagala loc represents performance with post-hoc determined best-match HRTF calculated for each metric in exp-zagala. current present the combined results of groups G-anech and G-reverb in the current study. stitt presents the results of groups W4 and W10 in exp-stitt. This last group is added to the comparison despite having already been through one training session, as no localization task was performed before the training in exp-stitt. From left to right, angle metrics error bars (top) respectively represent 1092 pts, 1250 pts, 2000 pts, 1440 pts, and 1092 pts, while confusion metrics error bars (bottom) respectively represent 28 pts, 10 pts, 16 pts, 24 pts, and 28 pts.

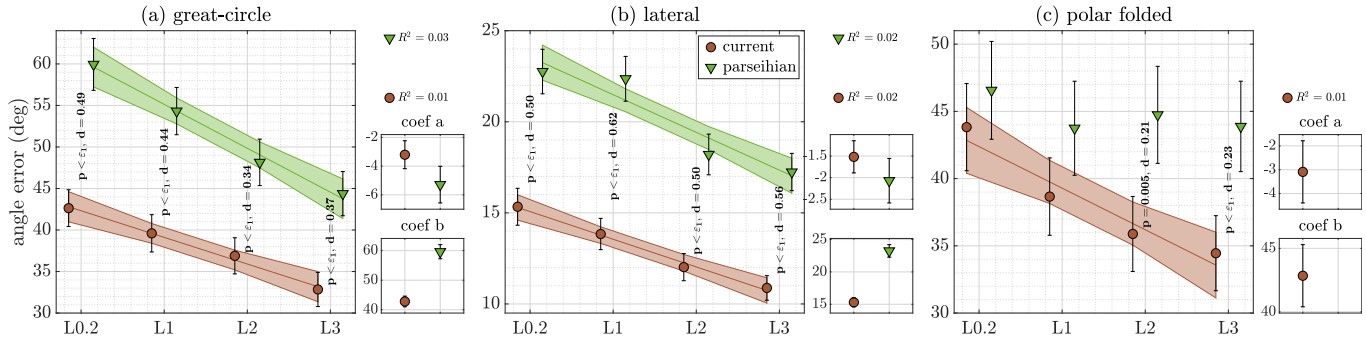

Figure 11. (Color online) Comparison of mean and 95% CI of angular errors between sessions and experiments. Current present the results of group G-anech in this experiment. Parseihian presents the results of group G3 in exp-Parseihian. Colored lines and bands correspond to the value and 95% CI of an $ax + b$ linear regression. p-values are reported when pair-wise comparisons between contiguous distributions proved significant. Right hand column of each subfigure report the R^2 goodness of fit and the regression coefficients values and 95% CI. The error bars associated to each experiment session respectively represent the 720, and 625 pts of 12 and 5 participants.

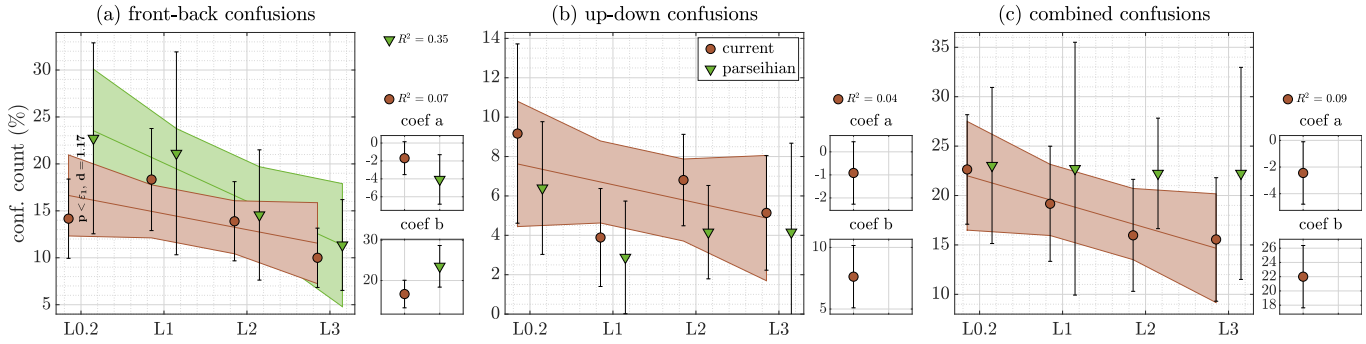

Figure 12. (Color online) Comparison of mean and 95% CI of confusion error percentages between sessions and experiments. The error bars associated to current, and parseihian sessions respectively represent 12 and 5 pts.

5.7.2 Impact of the training protocol on performance evolution (H6)

The evolution of exp-Parseihian and exp-current participant performance across training sessions is illustrated in Figures 11 and 12. The linear regression fit applied on participant results is of the form $ax + b$, as per Majdak et al. [28]. The low goodness of fit values reported on both figures are a direct result of large inter and intra participant variance. Regression coefficients are still judged meaningful as long as session ID proved to have a significant impact on the metric under consideration. Regression-related elements are omitted from Figures 11 and 12 when this condition is not met.

Results on *initial performance* coefficient b reflect those reported in Section 5.7.1: exp-current participants started training with lower great-circle and lateral angle errors than those of exp-Parseihian. A direct comparison of the *improvement rates* coefficient, a, suggests that there is no significant difference between training methods, except for lateral error evolution. Comparison of 95% CI overlap of coefficient a with zero, however, reveals that expcurrent participants alone improved on polar folded error and combined confusion rates. A similar comparison suggests that exp-Parseihian participants significantly improved on front–back confusions where exp-current did not, a result mitigated by unbalanced participants initial performance as discussed in Section 6.

Pairwise post-hoc comparison p-values are shown on Figures 11 and 12 for those distributions showing significant session × experiment interactions. Having a slightly better initial performance and improving at the same rate, exp-current participants outperformed those of exp-Parseihian on great-circle and lateral error for all sessions. Having comparable initial performance and improving faster, exp-current participants outperformed those of exp-Parseihian on polar folded confusion rates from session L2 onwards.

6 Discussion

The comparison of participants active involvement during the HRTF selection with their initial localization performance in Section 5.2 did not argue in favor of H1.

The comparison of G-reverb and G-anech initial performance in Section 5.4 does not argue in favor of H2. The only

significant, if somewhat small impact of reverberation on performance observed prior to training was on combined confusions, reduced to ≈ 17% compared to ≈ 21% in the initial anechoic condition.

Results presented in Section 5.5 argue in favor of H3, which can be interpreted as participants who improved the most are those who started the worst, having more room for improvement. This observation was previously made in Stitt et al. [19] and Majdak et al. [28], underlying the logarithmic progression model proposed in the latter. It is important to note that no extreme cases were observed where a participant failed to improve, as reported by Stitt et al. This is most likely due the context of that study, focusing on worst-case scenario training with a worst-match HRTF, while the current study employed a rapid best-match HRTF training protocol. These results again emphasize the importance of suitable HRTF matching even when training.

The comparison of G-reverb and G-anech performance evolution in Section 5.6 clearly supports H4. Despite the short duration of the training, some of the metrics monitored tend to plateau from L2 onwards (lateral, polar folded, and combined confusions in Figs. 8 and 9). Comparison of the plateaued values suggests that training under reverberant condition could, beyond accelerating learning rates, lead to better long-term overall performance. Two or more additional training sessions would be required to assert this extension of H4.

The comparison of performance prior to training in Section 5.7.1 argues in favor of H5. The proposed HRTF selection method performed surprisingly well, even compared to the post-hoc per-metric best-case zagala loc reference. The authors believe that beyond providing a good-match HRTF, the main advantage of the proposed method is that it intuitively confronts users to the impact of HRTF matching on spatial perception. Coupled with a selection task suited to novices and experts alike (not fastidious for the former, allowing the later precise testing), this method somewhat led participants to actively work on their learning strategy, framing the whole training as an opportunity to improve on a valuable game skill.

Finally, the inter-experiment comparison of performance evolution in Section 5.7.2 supports H6. When exp-current participants started with performance on par to those of the other experiments, they either learned at faster (polar folded errors, combined confusions) or similar rates (up–down confusions). When exp-current participants started with better performance, they learned at similar (great-circle and lateral errors) or slower rates (front–back confusion). Pending further testing, this last result, but for front-back confusions, argues in favor of H6, as it has been shown that long-term evolution followed a logarithmic progression [28], i.e. that the more room for improvement participants have, the faster they tend to learn.

According to informal interviews and observation of participants during the experiment, one of the advantages of this training program is that it coupled small and focused training sequences (scenarios) with an explicit scoring feedback. The authors believe that, with the dynamic adjustment of generated configurations targeting localization conditions where participants showed difficulty, these elements were key to the observed training efficiency.

As discussed in Section 4.3, the second-order Ambisonic reverberation and sparse 20-point RIR grid were adopted as a trade-off between spatial precision and processing power requirement. Coupled to the artifacts of linear interpolation [69], these design choices limit the scope of the results. Additional work is required to characterize the impact of these choices, including the prescribed reverberation time and room shape, to further investigate the impact the room acoustic response had on learning efficiency.

7 Conclusion

This paper presented the results of a perceptual study designed to assess a novel HRTF selection method and training program, conceived to reduce the time required to obtain acceptable binaural localization performance. The 24 participants of the experiment started by selecting a non-individual HRTF from an existing database, with which they trained during three 12 min sessions. The rapid selection method employed an active exploration by participants where they judged audio-visual-proprioceptive coherence as a function of HRTF. Participants were divided into two groups, training under either anechoic (control) or reverberant conditions, to assess whether additional audio spatial cues provided via a 3D room response improved learning rates.

Analysis of initial localization performance (prior to training) indicated that the 5 min active HRTF selection method led to localization performance as good as, if not better than, previously suggested methods [5]. Analysis of participant evolution under anechoic conditions indicated that the training program led to improved learning rates compared with that of previous studies [6]. Finally, comparisons between group performance showed that the proposed self-coherent, subtle scene-agnostic room acoustic response accelerated non-individual HRTF learning compared to anechoic conditions.

Conflict of interest

The Authors declared no conflict of interests.

Acknowledgments

This work was funded in part through a fundamental research collaboration partnership between Sorbonne Université, CNRS, Institut ∂'Alembert and Facebook Reality Labs.

References

1. J. Blauert: Spatial hearing: The psychophysics of human sound localization. MIT Press, 1997.

2. D.R. Begault, 3-D Sound for Virtual Reality and Multimedia, Academic Press, Cambridge, 1994.
3. E.M. Wenzel, M. Arruda, D.J. Kistler, F.L. Wightman: Localization using nonindividualized head-related transfer functions. The Journal of the Acoustical Society of America 94, 1 (1993) 111–123. https://doi.org/10.1121/1.407089.
4. A.W. Bronkhorst: Localization of real and virtual sound sources. The Journal of the Acoustical Society of America 98, 5 (1995) 2542–2553. https://doi.org/10.1121/1.413219.
5. F. Zagala, M. Noisternig, B.F.G. Katz: Comparison of direct and indirect perceptual head-related transfer function selection methods. The Journal of the Acoustical Society of America 147, 5 (2020) 3376–3389. https://doi.org/10.1121/10.0001183.
6. G. Parseihian, B.F.G. Katz: Rapid head-related transfer function adaptation using a virtual auditory environment, The Journal of the Acoustical Society of America 131, 4 (2012) 2948–2957. https://doi.org/10.1121/1.3687448.
7. S. Carlile, K. Balachandar, H. Kelly: Accommodating to new ears: the effects of sensory and sensory-motor feedback. The Journal of the Acoustical Society of America 135, 4 (2014) 2002–2011. https://doi.org/10.1121/1.4868369.
8. S. Xu, Z. Li, G. Salvendy: Individualization of head-related transfer function for three dimensional virtual auditory display: a review. Intl Conf on Virtual Reality, Springer, 2007, pp. 397–407. https://doi.org/10.1007/978-3-540-73335-5_4
9. B.F.G. Katz: Boundary element method calculation of individual head-related transfer function. II. Impedance effects and comparisons to real measurements. The Journal of the Acoustical Society of America 110, 5 (2001) 2449–2455. https://doi.org/10.1121/1.1412441.
10. R.O. Duda, V.R. Algazi, D.M. Thompson: The use of head-and-torso models for improved spatial sound synthesis. Audio Engineering Society Convention 113 (2002) 1–18.
11. J.C. Middlebrooks, E.A. Macpherson, Z.A. Onsan: Psychophysical customization of directional transfer functions for virtual sound localization. The Journal of the Acoustical Society of America 108, 6 (2000) 3088–3091. https://doi.org/10.1121/1.1322026.
12. A. Silzle: Selection and tuning of HRTFs. Audio Engineering Society Convention 112 (2002) 1–14.
13. D. Schönstein, B.F.G. Katz: HRTF selection for binaural synthesis from a database using morphological parameters. Intl Congress on Acoustics (2010) 1–6.
14. B.U. Seeber, H. Fastl: Subjective selection of non-individual head-related transfer functions. Intl Conf on Auditory Display (2003) 259–262.
15. D. Zotkin, J. Hwang, R. Duraiswaini, L.S. Davis: HRTF personalization using anthropometric measurements, in Workshop on Applications of Sig Proc to Audio and Acoustics, IEEE, 2003, pp. 157–160. https://doi.org/10.1109/ASPAA.2003.1285855.
16. Y. Iwaya: Individualization of head-related transfer functions with tournament-style listening test: Listening with other's ears. Acoustical Science & Technology 27, 6 (2006) 340–343. https://doi.org/10.1250/ast.27.340.
17. A. Andreopoulou, B.F.G. Katz: Investigation on subjective HRTF rating repeatability. Audio Engineering Society Convention 140 (2016a) 9597, 1–10.
18. D. Poirier-Quinot, B.F.G. Katz: Assessing the impact of head-related transfer function individualization on performance: Case of a virtual reality shooter game. Journal of the Audio Engineering Society 68, 4 (2020). https://doi.org/10.17743/jaes.2020.0004.
19. P. Stitt, L. Picinali, B.F.G. Katz: Auditory accommodation to poorly matched nonindividual spectral localization cues through active learning. Scientific Reports 9, 1 (2019) 1063, 1–14. https://doi.org/10.1038/s41598-018-37873-0.
20. L. Simon, N. Zacharov, B.F.G. Katz: Perceptual attributes for the comparison of Head- Related Transfer Functions. The Journal of the Acoustical Society of America 140 (2016) 3623–3632. https://doi.org/10.1121/1.4966115.
21. B.A. Wright, Y. Zhang: A review of learning with normal and altered sound- localization cues in human adults. International Journal of Audiology 45, sup1 (2006) 92–98. https://doi.org/10.1080/14992020600783004.
22. C. Mendonça: A review on auditory space adaptations to altered head-related cues. Frontiers in Neuroscience 8, 219 (2014) 1–14. https://doi.org/10.3389/fnins.2014.00219.
23. P.M. Hofman, J.G. Van Riswick, A.J. Van Opstal: Relearning sound localization with new ears. Nature Neuroscience 1, 5 (1998) 417–421. https://doi.org/10.1038/1633.
24. M.M. Van Wanrooij, A.J. Van Opstal: Relearning sound localization with a new ear. Journal of Neuroscience 25, 22 (2005) 5413–5424. https://doi.org/10.1523/JNEUR0SCI.0850-05.2005.
25. R. Trapeau, V. Aubrais, M. Schönwiesner: Fast and persistent adaptation to new spectral cues for sound localization suggests a many-to-one mapping mechanism, The Journal of the Acoustical Society of America 140, 2 (2016) 879–890. https://doi.org/10.1121/1.4960568.
26. P. Zahorik, P. Bangayan, V. Sundareswaran, K. Wang, C. Tam: Perceptual recalibration in human sound localization: Learning to remediate front-back reversals. The Journal of the Acoustical Society of America 120, 1 (2006) 343–359. https://doi.org/10.1121/1.2208429.
27. M.A. Steadman, C. Kim, J.-H. Lestang, D.F. Goodman, L. Picinali: Short-term effects of sound localization training in virtual reality. Scientific Reports 9, 1 (2019) 1–17. https://doi.org/10.1038/s41598-019-54811-w.
28. P. Majdak, T. Walder, B. Laback: Effect of long-term training on sound localization performance with spectrally warped and band- limited head-related transfer functions. The Journal of the Acoustical Society of America 134, 3 (2013) 2148–2159. https://doi.org/10.1121/1.4816543.
29. C. Mendonça, G. Campos, P. Dias, J.A. Santos: Learning auditory space: Generalization and long-term effects. PloS One 8, 10 (2013) 1–14. https://doi.org/10.1371/journal.pone.0077900.
30. A. Honda, H. Shibata, J. Gyoba, K. Saitou, Y. Iwaya, Y. Suzuki: Transfer effects on sound localization performances from playing a virtual three-dimensional auditory game. Applied Acoustics 68, 8 (2007) 885–896. https://doi.org/10.1016/j.apacoust.2006.08.007.
31. J. Hamari, J. Koivisto, H. Sarsa: Does gamification work? A literature review of empirical studies on gamification, in: Intl Conf on System Sciences, IEEE, 2014, pp. 3025–3034. https://doi.org/10.1109/HICSS.2014.377.
32. C. Mendonça, G. Campos, P. Dias, J. Vieira, J.P. Ferreira, J. A. Santos: On the improvement of localization accuracy with nonindividualized HRTF-based sounds. Journal of the Audio Engineering Society 60, 10 (2012) 821–830.
33. T. Bouchara, T.-G. Bara, P.-L. Weiss, A. Guilbert: Influence of vision on short-term sound localization training with non-individualized HRTF. EAA Spatial Audio Signal Processing Symp (2019) 55–60. https://doi.org/10.25836/sasp.2019.04.
34. F. Dramas, B.F.G. Katz, C. Jouffrais: Auditory-guided reaching movements in the peripersonal frontal space. The Journal of the Acoustical Society of America 123, 5 (2008) 3723–3723. https://doi.org/10.1121/1.2935195.
35. D.P. Kumpik, O. Kacelnik, A.J. King: Adaptive reweighting of auditory localization cues in response to chronic unilateral earplugging in humans. Journal of Neuroscience 30, 14 (2010) 4883–4894. https://doi.org/10.1523/JNEUR0SCI.5488-09.2010.

36. K. Molloy, D.R. Moore, E. Sohoglu, S. Amitay: Less is more: latent learning is maximized by shorter training sessions in auditory perceptual learning. PloS One 7, 5 (2012) 1–13. https://doi.org/10.1371/journal.pone.0036929.
37. P. Majdak, M.J. Goupell, B. Laback: 3-D localization of virtual sound sources: Effects of visual environment, pointing method, and training. Attention, Perception, & Psychophysics 72, 2 (2010) 454–469. https://doi.org/10.3758/APP.72.2.454.
38. S. Carlile, P. Leong, S. Hyams: The nature and distribution of errors in sound localization by human listeners. Hearing Research 114 (1997) 179–196. https://doi.org/10.1016/S0378-5955(97)00161-5.
39. B. Gourévitch, R. Brette: The impact of early reflections on binaural cues. The Journal of the Acoustical Society of America 132, 1 (2012) 9–27. https://doi.org/10.1121/1.4726052.
40. N. Kaplanis, S. Bech, S.H. Jensen, T. van Waterschoot: Perception of reverberation in small rooms: a literature study. Audio Eng Soc Conf on Spatial Audio 55 (2014) 1–14.
41. W.M. Hartmann: Localization of sound in rooms. The Journal of the Acoustical Society of America 74, 5 (1983) 1380–1391. https://doi.org/10.1121/1.390163.
42. B. Rakerd, W. Hartmann: Localization of sound in rooms, II: The effects of a single reflecting surface. The Journal of the Acoustical Society of America 78, 2 (1985) 524–533. https://doi.org/10.1121/1.392474.
43. R. Guski: Auditory localization: Effects of reflecting surfaces. Perception 19, 6 (1990) 819–830. https://doi.org/10.1068/p190819.
44. S. Bech: Spatial aspects of reproduced sound in small rooms. The Journal of the Acoustical Society of America 103, 1 (1998) 434–435. https://doi.org/10.1121/1.421098.
45. D.R. Begault: Perceptual effects of synthetic reverberation on three-dimensional audio systems. Journal of the Audio Engineering Society 40, 11 (1992) 895–904.
46. B.G. Shinn-Cunningham: "Learning reverberation: Considerations for spatial auditory displays. Proc Intl Conf on Auditory Display (2000) 126–134.
47. D.R. Begault, E.M. Wenzel, M.R. Anderson: Direct comparison of the impact of head tracking, reverberation, and individualized head- related transfer functions on the spatial perception of a virtual speech source. Journal of the Audio Engineering Society 49, 10 (2001) 904–916.
48. D.R. Begault, E.M. Wenzel, A.S. Lee, M.R. Anderson: Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source. Audio Engineering Society Convention 108, 5134 (2000) 1–19.
49. E.J. Angel, R. Algazi, R.O. Duda: On the design of canonical sound localization environments. Audio Engineering Society Convention 113 (2002) 5714, 1–12.
50. A. Nykänen, A. Zedigh, P. Mohlin: Effects on localization performance from moving the sources in binaural reproductions, in Intl Cong and Exposition on Noise, Control Engineering 4 (2013) 3193–3201.
51. B.F.G. Katz, R. Nicol: Binaural spatial reproduction, in Sensory Evaluation of Sound, Zacharov N., Ed., CRC Press, Boca Raton, 2019, pp. 349–388.
52. A. Borrego, J. Latorre, M. Alcañiz, R. Llorens: Comparison of Oculus Rift and HTC Vive: Feasibility for virtual reality-based exploration, navigation, exergaming, and rehabilitation. Games for Health Journal 7, 3 (2018) 151–156. https://doi.org/10.1089/g4h.2017.0114.
53. A. Becher, J. Angerer, T. Grauschopf: Novel approach to measure motion-to-photon and mouth-to-ear latency in distributed virtual reality systems, in: GIVR/AR Workshop (2018) 1–14, arxiv.org/abs/1809.06320.
54. D. Poirier-Quinot, B.F.G. Katz: The Anaglyph binaural audio engine. Audio Engineering Society Convention 144 (2018) 1–4.
55. D. Brungart, A.J. Kordik, B.D. Simpson: Effects of head-tracker latency in virtual audio displays, Journal of the Audio Engineering Society 54 (2006) 32–44.
56. B.N. Postma, B.F.G. Katz: Perceptive and objective evaluation of calibrated room acoustic simulation auralizations. The Journal of the Acoustical Society of America 140, 6 (2016) 4326–4337. https://doi.org/10.1121/1.4971422.
57. S. Bertet, J. Daniel, E. Parizet, O. Warusfel: Investigation on localisation accuracy for first and higher order ambisonics reproduced sound sources. Acta Acustica 99, 4 (2013) 642–657. https://doi.org/10.3813/AAA.918643.
58. L. Picinali, A. Wallin, Y. Levtov, D. Poirier- Quinot: Comparative perceptual evaluation between different methods for implementing reverberation in a binaural context, in AES Convention 142, Berlin, Germany, 2017, p. 9742, 1–7. https://hal.archives-ouvertes.fr/hal-01790217.
59. I. Engel, C. Henry, S.V. Amengual Gari, P.W. Robinson, L. Picinali: Perceptual implications of different Ambisonics-based methods for binaural reverberation. The Journal of the Acoustical Society of America 149, 2 (2021) 895–910. https://doi.org/10.1121/10.0003437.
60. I. Engel, C. Henry, S.V.A. Gari, P.W. Robinson, D. Poirier-Quinot, L. Picinali: Perceptual comparison of ambisonics-based reverberation methods in binaural listening, in: EAA Spatial Audio Signal Processing Symposium, Paris, France, 2019, pp. 121–126. https://doi.org/10.25836/sasp.2019.11.
61. M. Noisternig, T. Musil, A. Sontacchi, R. Holdrich: 3D binaural sound reproduction using a virtual Ambisonic approach, Intl Symp on Virtual Env, HCI and Meas Systems, IEEE, 2003, pp. 174–178. https://doi.org/10.1109/VECIMS.2003.1227050
62. B.F.G. Katz, G. Parseihian: Perceptually based head-related transfer function database optimization, The Journal of the Acoustical Society of America 131, 2 (2012) 99105. https://doi.org/10.1121/1.3672641.
63. A. Andreopoulou, B.F.G. Katz: Subjective HRTF evaluations for obtaining global similarity metrics of assessors and assessees. Journal of Multimodal User Interfaces 10, 3 (2016b) 259–271. https://doi.org/10.1007/s12193-016-0214-y.
64. O. Warusfel: IRCAM Listen HRTF database, 2003. http://recherche.ircam.fr/equipes/salles/listen, last checked 2018-09-29
65. R.S. Woodworth, H. Schlosberg: Experimental psychology, Rev ed., Holt, Oxford, England, 1954.
66. H. Bahu, T. Carpentier, M. Noisternig, O. Warusfel: Comparison of different egocentric pointing methods for 3D sound localization experiments. Acta Acustica 102, 1 (2016) 107–118. https://doi.org/10.3813/AAA.918928.
67. M. Morimoto, H. Aokata: Localization cues of sound sources in the upper hemisphere. Journal of the Acoustical Society of Japan 5, 3 (1984) 165–173. https://doi.org/10.1250/ast.5.165.
68. G. Cumming: The new statistics: Why and how. Psychological Science 25, 1 (2014) 7–29. https://doi.org/10.1177/0956797613504966.
69. M. Zaunschirm, F. Zotter, M. Frank: Perceptual evaluation of variable-orientation binaural room impulse response rendering. Audio Engineering Society, 2019.

Appendix

Training scenarios

This Appendix presents design details of the experimental protocol and gamified training scenario evolution during training.

Participants started T1 with only the first and easiest scenario unlocked. The nth scenario could be unlocked by correctly identifying more than a certain percentage (75% by default) of the active targets in the $(n-1)$th scenario. The best attempt percentage was displayed beside each scenario, as a progress indicator for participants to know which issue required training. From one training session to the next, participant scores and unlocked scenarios were preserved in their game profile file. Once unlocked, a scenario could be accessed at will during the remainder of the training as well as during the subsequent training sessions.

Scenario definition was implemented in a `.json` configuration file, using a simple syntax so that nondevelopers could easily define new ones. A typical scenario entry consisted of a title, a short description to help participants understand its focus, an integer indicating how many trials this scenario should spawn, as well as a number of *sets*. Each set defined two or more surface segments of the unit sphere or *spawning zones*. At the beginning of each trial, the training program selected a set and spawned one visual target per zone in that set. One of the zones would then be selected to become the active zone, i.e. the visual target it contained would become the active target. The set and the active zone for a given trial were selected based on probability weights, adjusted according to participants previous achievements. The zone-based spawning system was adopted so that a scenario could be designed to work on a given problem while presenting participants with ever changing potential visual target positions. The weighting system was designed so that participants would progressively be confronted with only those [set + active zone] configurations they struggled to master, and did not waste time on those that no longer presented any difficulties.[3] Two example training scenarios are illustrated in Figure A.1.

An exhaustive list of the training scenarios is provided here. The proposed classification, grouping scenarios under "*difficulty level*" categories (beginner, novice, etc.) was defined to give participants a sense of progression throughout the training.

Beginner level scenarios

- Single: Unique target per trial (tutorial scenario).
- Left–Right: Two visual targets per trial, left–right symmetry (tutorial + audio check scenario).
- Front–Back: Two visual targets, front–back symmetry.
- Up–Down: Two visual targets, up–down symmetry (see Fig. A.1 (left)).
- Accuracy: Two visual targets, located in a given quadrant of the sphere (no front–back, up–down, or left–right ambiguities). Targets spaced by 40 degrees or more.

Novice Level Scenarios

- Left–Right: Two visual targets, left–right symmetry. Based on a finer zone segmentation than that of Figure A.1 (left): 8 × 3 zones with 40 degree margins, resulting in more ambiguous localization scenarios.
- Front–Back: Two visual targets, front–back symmetry, using the 8 × 3 zone segmentation.
- Up–Down: Two visual targets, up–down symmetry, using the 8 × 3 zone segmentation.
- Cone of confusion: Four visual targets located on a given cone of confusion (constant interaural azimuth, see Figure A.1 (right)).

Advanced level scenarios

- Cone of confusion: Three to five visual targets located on a given cone of confusion, using a finer zone segmentation than that of the novice scenario.
- Front–Back 1: Two visual targets, front–back symmetry. Same zones segmentation than that of the novice scenario but with a more rigorous achievement threshold (lower number of mistakes accepted).
- Up–Down: Two visual targets, up–down symmetry. Same zones segmentation than that of the novice scenario but with a more rigorous achievement threshold.

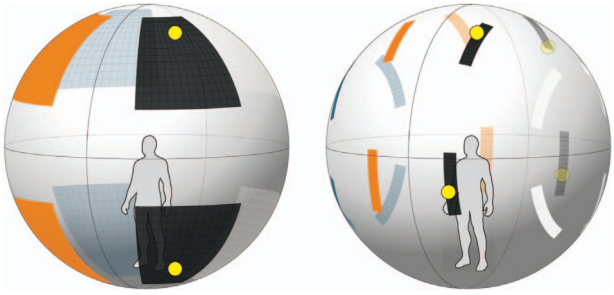

Figure A.1. (Color online) Example training scenarios used in the experiment: (left) "*Beginner up-down confusions*" and (right) "*Novice cone of confusion*". Each colored patch corresponds to a zone; zones of identical colors correspond to a set. Yellow dots illustrate a generated set of visual targets for a trial using the black set. A scenario specific flag could be set to force symmetric target sets, e.g. along the horizontal plane as illustrated for up–down confusions training (left).

[3] The [set + active zone] configuration was randomly picked from those available for a given scenario based on a warped probability distribution, that gave the (n + 1)th configuration (ranked based on participant score to each configuration in descending order) w time more chances to be picked up than the nth configuration. The warping factor w was defined in [2, 10] on a per-scenario basis.

- Front–Back 2: Two visual targets, front–back symmetry. Based on a 16 × 4 zone segmentation with a more rigorous achievement threshold than that of the "Front–Back 1" scenario.

Master level scenario

- Tutti: Twenty visual targets homogeneously distributed on the sphere (above −45° elevation).

Cite this article as: Poirier-Quinot D. & Katz B.F.G. 2021. On the improvement of accommodation to non-individual HRTFs via VR active learning and inclusion of a 3D room response. Acta Acustica, **5**, 25.