DOI: 10.61782/fa.2023.0904

WHEN HRTF TRAINING DOESN'T WORK

David Poirier-Quinot [1*] **Martin S. Lawless** [2] **Brian F. G. Katz** [1]

[1] Sorbonne University, CNRS, UMR 7190, Institut Jean Le Rond ∂'Alembert, Paris, France
[2] State University of New York Maritime College, US

ABSTRACT

Previous studies have established the efficiency of training programs to help users improve on binaural auditory localization when using a non-individual head-related transfer function. This paper reports the results of an experiment where participants trained with such a program, though they failed to improve. Interestingly, the same training program had successfully been used in a previous study. After a brief description of the program and the experiment protocol, a comparative analysis of the results of both studies is provided. The discussion then focuses on the differences between these studies in an attempt to better understand what caused the training to fail.

Keywords: *binaural, HRTF, training, localization*

1. INTRODUCTION

Most wearable Augmented Reality (AR) and Virtual Reality (VR) systems today use binaural synthesis to simulate 3-dimensional environments over headphones. The objective of this technique is to render spatial auditory scenes by applying direction-dependent audio cues to monophonic signals to alter the timing and frequency content and makes it appear to the listener that the sound originates externally from a location in the virtual space [1, 2]. The set of direction-dependent audio cues, including timing and level differences between the left and right ears caused by the location of the source as well as distortions in the signal due to reflections from the head, torso, and pinna, are all contained within the head-related transfer function (HRTF), which represents the propagation of an acoustic wave from a set of specific source positions to the listener's ears. Therefore, an individual HRTF [1] tends to be unique to each listener and is not as effective for rendering spatial audio scenes for other listeners, often causing a degradation in the externalization or localization accuracy of the virtual sources [3, 4].

AR and VR systems commonly use non-individual HRTFs because measuring each consumer's HRTF is currently impractical. Past studies have shown that users can adapt to non-individual HRTFs, exhibiting improved localization performance approaching that of listeners using individual HRTFs [5–7]. However, modern literature on this "rapid" HRTF adaptation via training has been limited to studies conducted in the lab (in contrast to studies such as [8–10] using long-term passive adaptation) in controlled environments with high-end computing and pro-audio hardware. These limitations would be prohibitive for everyday consumers of AR and VR devices.

The objective of the present study was initially to determine if an HRTF training program could be conducted effectively at home using an off-the-shelf head mounted display (HMD). Participants completed a three-day HRTF training using an established learning program [11] at home with an HMD, and their localization accuracy performance was evaluated over the course of three training sessions to assess how their performance changed with time. The results of this experiment were compared

**Corresponding author*: *david.poirier-quinot@dalembert.upmc.fr.*

Copyright: *©2023 David Poirier-Quinot et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 Unported License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.*

[1] We use the term *individual* to identify the HRTF of the user, *individualized* or *personalized* to indicate an HRTF modified or selected to accommodate the user best, and *non-individual* or *non-individualized* to indicate an HRTF that has not been tailored to the user. A so-called *generic* or *dummy-head* HRTF is a specific instance of a non-individual HRTF.

to a previous study [11], which used the same training program in the lab, to ascertain whether the participants exhibited similar levels of localization performance improvement over the course of the training.

2. MATERIALS AND METHODS

Fourteen participants trained on HRTF adaptation for auditory localization, all using the same non-individual HRTF. They trained at home, for 12 minutes per day over three consecutive days. They were equipped with a Quest 2 HMD, using the built-in headphones, running a training program designed to adapt auditory cue interpretation using feedback and proprioception. The training timeline is depicted in Figure 1.

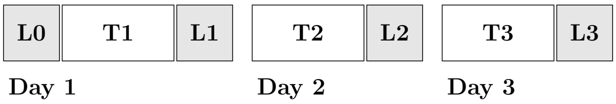

Figure 1: Schematic timeline of the experiment sequence. **L***i* are localization tasks used to evaluate participant performance, **T***i* are training tasks.

The HRTF, or more precisely the binaural room impulse response (BRIR), participants trained with was that of the KU100 dummy-head, recorded with a 5° resolution on a 0.8 m radius sphere in a dry room ($T_{30,1000\,\mathrm{Hz}}$ of 0.12 s). This $4.2 \times 3.8 \times 3.2$ m^3 room was selected as it is very similar to the one simulated in [11] that helped accelerate training compared to anechoic conditions.

The training program, detailed and evaluated in [11], was developed in Unity. It is divided into 14 difficulty levels where participants faced the various challenges of auditory localization with non-individual binaural rendering: front-back confusions, localization blur, etc.. Each level is composed of trials; for each trial, participants indicate which of the visual targets surrounding them is the one emitting a sound. The number of visual targets (*i.e.* of potential decoys) displayed depends on the current level's focus and difficulty grading. To further help with the task, participants are equipped with a virtual auditory source, a "probe" attached to each of their hands, to allow them to listen to how the HRTF sounds at any given position.

Before the first training, to establish a baseline reference, and after each training session, participants performed an auditory localization task in the same virtual environment as the training. This evaluation task, identical to that used in [6, 7, 11, 12] consisted in a series of trials during which participants pointed towards the perceived location of the sound source. Participants could no longer use the probe during the localization task. Participants were tested on twenty different source locations distributed around them on the sphere, each repeated three times for each localization session.

The audio stimulus used during both the training and localization task consisted of a sequence of three white noise bursts, lasting a total of 180 ms to limit the use of head movement and dynamic localization cues. Played as a loop during the training, so that participants had time to estimate its position using their probes, it was only played once during the localization task. During the training, participants would only hear the stimulus when their head was at rest position [11]), preventing them from relying on, and thus learning, dynamic cues via head movement despite the looping stimulus. The binaural rendering was generated off-line, prior to the experiment, with a different audio file created for each potential position on the sphere by convolving the input bursts with each of the KU100 BRIRs. As in [11], eight different versions of the stimulus were created for each position, using different noise seeds and applying a ±3 dB level roving to ensure participants did not rely on synthesis artifacts to identify auditory source positions.

Along with the HMD, participants were given instructions to set up the experiment at home. They were instructed to use the HMD built-in headphones, to train while standing up in a clutter-free 2 m wide zone to encourage full auditory sphere exploration, and to disable WiFi during training to avoid distracting notifications.

2.1 Data analysis

Localization accuracy was assessed based on the methodology presented in [13]. The global extent of the localization error is first assessed using the great-circle error. Critical localization confusions are then evaluated using the response classification scheme proposed in [13, *c.f.* Figure 4] with a 45° threshold, illustrated in Figure 2. Finally, the local extent of localization error is evaluated using the local great-circle, local lateral, and local polar weighted error. These "local" metrics are computed by considering only responses classified as "precision" responses, *i.e.* that fell within a 45° cone around the source location, to avoid localization reversals inflating angular metric values [12].

Analyses of variances (ANOVAs) [14] were conducted for each of the dependent variables of mean global

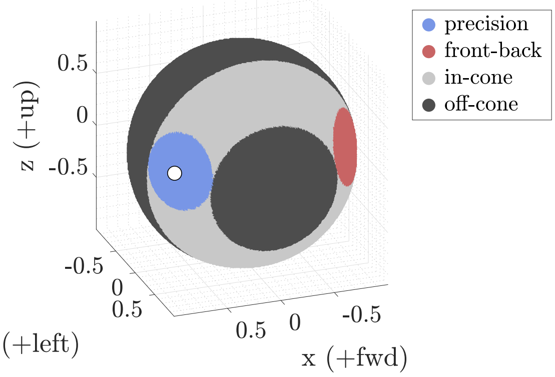

Figure 2: Response classification scheme proposed in [13], defining the response type as a function of the response position on the sphere. The white circle indicates the considered target position, at spherical coordinates (35°, 10°). The listener is facing X with his left ear pointing towards Y. The angle threshold used in the figure is of 20° compared to the 45° used in the analysis to better illustrate the various response type regions.

and local great circle angle, lateral, and polar errors (using R [15]), to assess the evolution of participant performance within and across sessions. The factors included in the analysis were source location repetition (1, 2, or 3) and session number (**L0**–**L3**), as well as the first-order interaction between these two factors. Likewise, generalized linear mixed models (GLMMs), constructed as repeated measures logistic regressions [14], were used to evaluate the evolution of the percentage of reversal errors based on the same factors of the ANOVAs. For the GLMMs, the significance of each factor was determined by performing goodness of fit comparisons between the full models and models with single-term deletions.

Similar statistical models were used to compare the localization performance of the participants in the present study to those in [11] in Section 3.2. The models were adjusted to include the factor of group: participants in the present study are referred to as the $\mathbf{G_{current}}$ group, those in the previous study [11] as the $\mathbf{G_{previous}}$ group. The first-order interaction terms between groups and the other factors were also included in the models.

For all tests, statistical significance was determined for p-values below a 0.05 threshold. The notation $p < \varepsilon$ is adopted to indicate p-values below 10^{-3}. Post-hoc pairwise comparisons for significant factors were made with Tukey-Kramer adjusted p-values.

3. RESULTS AND DISCUSSION

3.1 Training assessment in the current study

The evolution of the four angular metrics for $\mathbf{G_{current}}$ is illustrated in Figure 3. ANOVA indicated that training had a significant impact on the global great-circle error ($p < \varepsilon$). Post-hoc comparison showed that it significantly decreased only between **L0** and **L1** ($50.4 \pm 2.1°$ vs. $44.4 \pm 2.1°$, $p < \varepsilon$), not evolving after **L1** ($p = 1.00$ for each **L1** through **L3** comparison). Further analysis showed that training had no significant impact on local great-circle error, lateral error, and weighted-polar ($p > 0.05$).

The evolution of precision responses rate and confusion rates for $\mathbf{G_{current}}$ is illustrated in Figure 4. ANOVA indicated that training had a significant impact on the percentage of in-cone and off-cone confusions ($p < \varepsilon$ for both). Post-hoc comparisons showed that between **L0** and **L1**, the in-cone confusion rate dropped from 21.0% to 16.9%, and the off-cone confusion rate from 1.8% to 0.7%. Neither confusion rate evolved after **L1**. No significant effect was observed for training on the front-back confusion rate.

3.2 Training comparison with the previous study

The training protocol used in the current study was nearly identical to that used in [11]. As such, the results of $\mathbf{G_{current}}$ participants are compared to those reported in [11] in this section. Only the results of participants of the “G-reverb” group in [11], training with a room acoustic condition comparable to that of the present study, will be considered in this analysis. This group is referred to here as $\mathbf{G_{previous}}$. $\mathbf{G_{current}}$ and $\mathbf{G_{previous}}$ trained during the same amount of time, with what appears to be the same dedication, as they completed a comparable average and standard deviation of 103 ± 46 and 109 ± 47 “training trials”, respectively, over the 3 days of their training. To better appreciate the comparison, an exhaustive list of the differences between the current training protocol and that used in [11] is provided in Table 1.

The factors of session and group as well as the interaction between them significantly affected the global great-circle angle error ($p < \varepsilon$, $p < 0.002$, and $p < \varepsilon$, respectively). Post-hoc means comparisons showed that

Table 1: Differences between the protocols of the present study vs. that used in [11]. ITD stands for Interaural Time Difference and HOA for High Order Ambisonic.

	Previous study [11] ($\mathbf{G_{previous}}$)	Current study ($\mathbf{G_{current}}$)
HRTF used	Best-match HRTF issued from a 2 min subjective selection from LISTEN [16] subset [17]	KU100 dummy-head
Virtual acoustic environment	Simulated test room HOA [18]	Actual test room BRIR
ITD individualization	ITD based on head circumference	No ITD adjustment
Headphones	Sennheiser HD 600	Quest 2 integrated headphones
Location	In the lab	At participant's home

both $\mathbf{G_{previous}}$ and $\mathbf{G_{current}}$ started with statistically similar performance levels in **L0** ($p = 1.00$), with errors of $47.2 \pm 2.4°$ and $48.7 \pm 2.1°$ respectively. Both groups exhibited significantly lower global great-circle angle errors in **L1** ($p < \varepsilon$), statistically similar to one another ($p = 0.46$) at $37.4 \pm 2.4°$ and $42.7 \pm 2.1°$, respectively. After **L1** however, $\mathbf{G_{current}}$ did not improve in their global great-circle angle performance in **L2** or **L3** ($p = 1.00$ for all three comparisons), while $\mathbf{G_{previous}}$ demonstrated clear localization improvement ($p < \varepsilon$ for all three comparisons). This result is illustrated in Figure 3(a).

Once again, the global great-circle error was decomposed into local angular metrics and confusion error rates. The factors of session and the interaction term between session and group were significant in the ANOVA model for local great-circle angle error ($p < \varepsilon$). Both groups of participants had the same initial local great-circle errors in **L0** ($p = 0.97$), illustrated in Figure 3(b). $\mathbf{G_{current}}$ did not demonstrate any performance improvement across the four sessions ($p > 0.96$). On the other hand, $\mathbf{G_{previous}}$ local great-circle error significantly improved from $25.1 \pm 0.8°$ in **L0** to $20.4 \pm 0.8°$ in **L3** ($p < \varepsilon$).

ANOVA of the absolute local lateral error indicated that both group and session factors were significant ($p < \varepsilon$ for both treatments), as well as the interaction between them ($p < 0.02$). As for local great-circle error, post-hoc analysis did not show any difference between both groups in **L0** ($p = 0.54$). $\mathbf{G_{previous}}$ showed a marked improvement in absolute local lateral error between **L0** and **L2** ($p < \varepsilon$) and between **L0** and **L3** ($p < \varepsilon$), but there were no significant differences between **L1**, **L2**, and **L3** ($p > 0.05$ for all comparisons). On the other hand, $\mathbf{G_{current}}$ showed no improvement at all during training, having similar absolute local lateral errors across all four sessions ($p > 0.05$ for all comparisons). This result is illustrated in Figure 3(c).

For absolute local weighted-polar error, only the main factors of session ($p < 0.003$) and group ($p < \varepsilon$) were significant, the interaction term was not ($p = 0.08$). On average between the two groups, the participants showed improvement between **L0** and **L2** ($p < 0.03$), and between **L0** and **L3** ($p < 0.003$), for a total improvement of $2.2 \pm 0.6°$, showed in Figure 3(d). There was no difference between **L1–L3** ($p > 0.05$ for all comparisons). Across the four sessions, the participants in $\mathbf{G_{previous}}$ had a better absolute local weighted-polar error than $\mathbf{G_{current}}$ by $3.4 \pm 0.9°$.

The interaction effect between the factors of session and group was also significant for front-back, in-cone, and off-cone confusion rates ($p < \varepsilon$, $p < 0.003$, and $p < 0.04$, respectively). Both $\mathbf{G_{previous}}$ and $\mathbf{G_{current}}$ had similar amounts of front-back confusion rates in **L0** (17.4% and 16.6%, respectively; $p = 1.00$). $\mathbf{G_{previous}}$ participants front-back confusion decreased to 11.8% in **L1** ($p < 0.04$), to then stagnate from **L1** to **L3** ($p > 0.05$ for all comparisons between sessions). By comparison, $\mathbf{G_{current}}$ stagnated during the whole training, and still had a front-back confusion rate of 16.3% in their last session **L3** ($p > 0.05$ for all comparisons), illustrated Figure 4(b). $\mathbf{G_{previous}}$ had fewer in-cone confusions overall, starting at 11.8% in **L0** and significantly improving to 4.1% in **L2** ($p < \varepsilon$). While the in-cone confusion rate in **L3** was lower for $\mathbf{G_{previous}}$, it was not significant different from **L2** ($p = 0.87$). In comparison, $\mathbf{G_{current}}$ had 21.0% in-cone confusions in **L0**, significantly more than $\mathbf{G_{previous}}$ ($p < 0.05$), and only improved to 15.2% in the final session ($p < 0.02$), illustrated in Figure 4(c). Lastly, while both groups initially had similar off-cone confusions rates in **L0** (1.3% for $\mathbf{G_{previous}}$ and 1.8% for $\mathbf{G_{current}}$, $p = 1.00$) and demonstrated improvement over time, $\mathbf{G_{previous}}$ ended the sessions with a slightly better rate of 0.1% compared

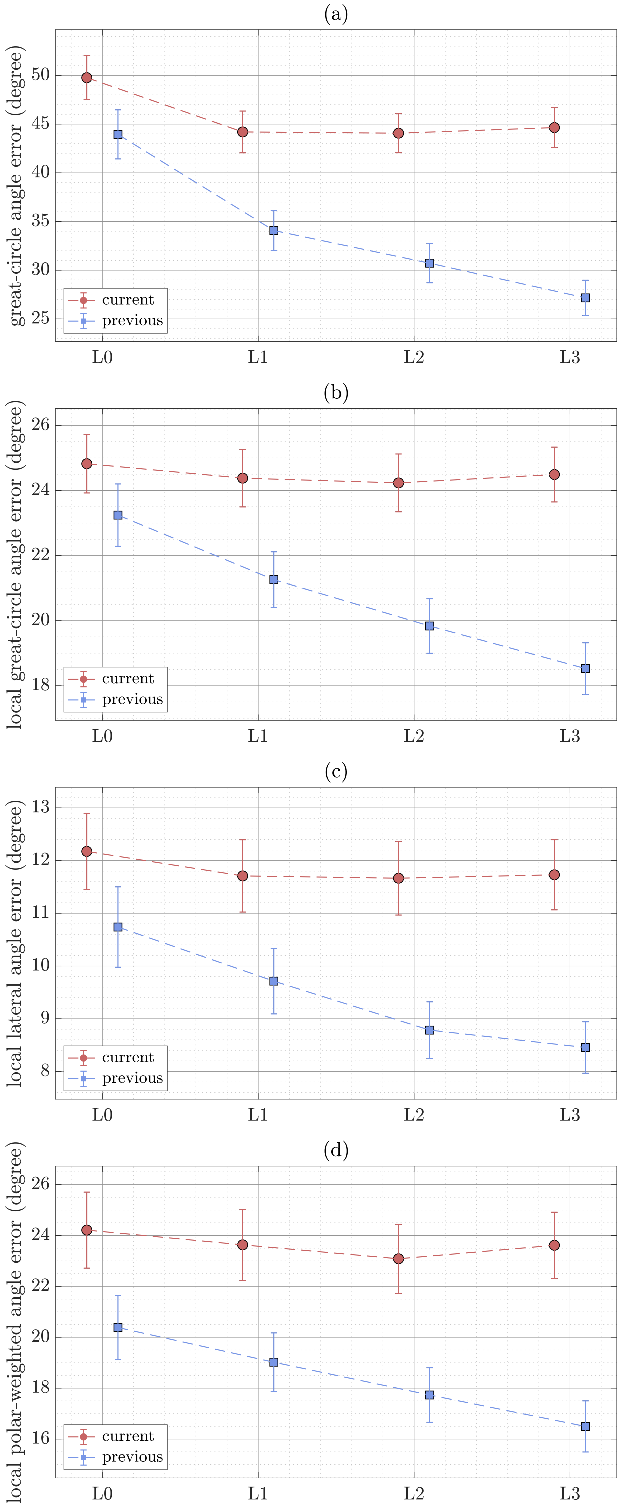

Figure 3: Evolution of mean and 95% Confidence Interval (CI) of angular errors across sessions for $\mathbf{G_{current}}$ and $\mathbf{G_{previous}}$.

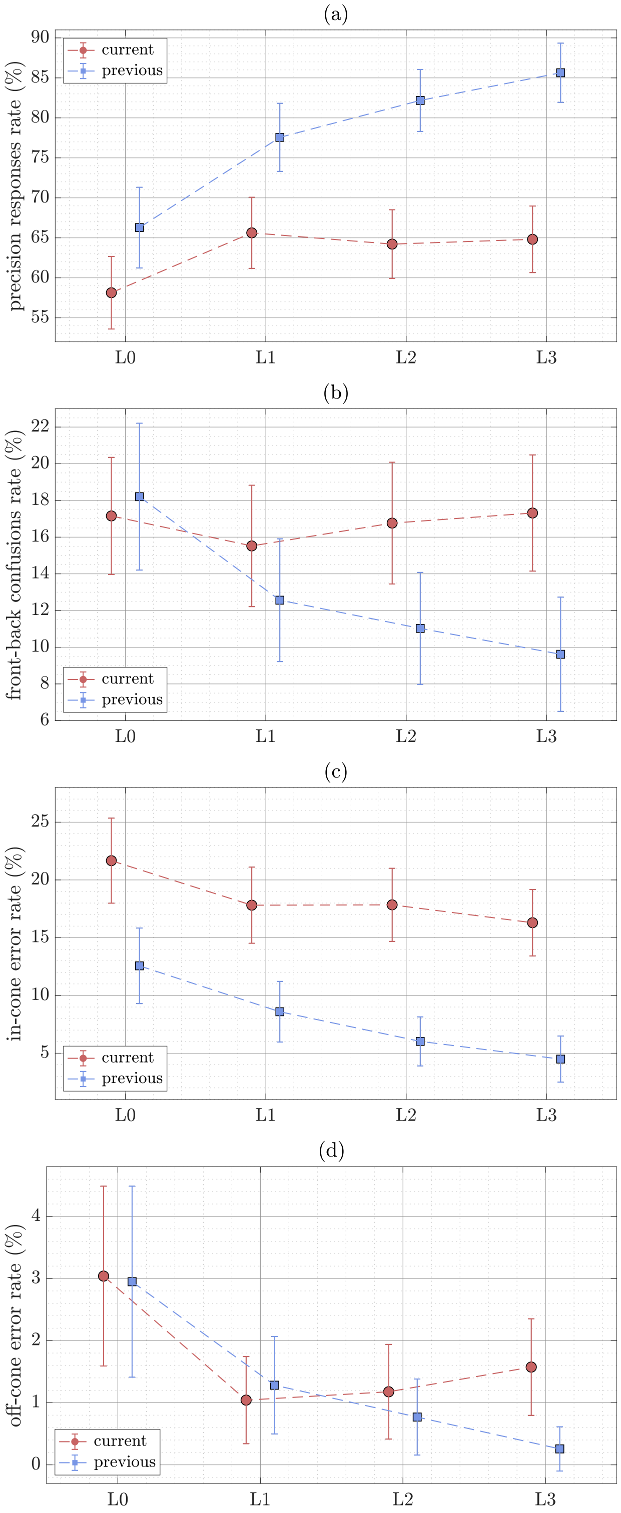

Figure 4: Evolution of mean and 95% Confidence Interval (CI) of response category rates across sessions for $\mathbf{G_{current}}$ and $\mathbf{G_{previous}}$. The four category rates sum to 100%.

to 0.9% for $\mathbf{G}_{\mathbf{current}}$, though these are not significantly different from each other ($p = 0.23$). Off-cone confusion rate evolution across sessions for both groups is illustrated Figure 4(d).

The impact of training on individual performance of $\mathbf{G}_{\mathbf{current}}$ and $\mathbf{G}_{\mathbf{previous}}$ is illustrated in Figure 5. This figure depicts an individual's change in "precision" response rate (*i.e.* 1−confusion rate) on the Y-axis and change in local accuracy on the X-axis between **L1** and **L3**. Improvement in precision rate is indicated by a greater precision rate in **L3** (higher on the Y-axis), while improvement in local accuracy is shown by lower great-circle angle error (left on the X-axis). The majority of the participants in $\mathbf{G}_{\mathbf{previous}}$ demonstrate improvement in one or both categories. $\mathbf{G}_{\mathbf{current}}$, however, is more centered about the middle of the graph with only some individuals exhibiting improvement, primarily in local accuracy. The figure also suggests that there is no obvious cluster of "good" vs. "bad" learners in $\mathbf{G}_{\mathbf{current}}$, as has been seen in previous studies [6, 7].

3.3 Discussion: why so little improvement for $\mathbf{G}_{\mathbf{current}}$ compared to $\mathbf{G}_{\mathbf{previous}}$?

The first interesting result reported in Section 3.2 is the difference between the performance of both groups before the training in **L0**. It might be that the lack of HRTF selection in $\mathbf{G}_{\mathbf{current}}$ compared to $\mathbf{G}_{\mathbf{previous}}$ is responsible for this difference. For comparison, results of reported in [6, 7] indicate that participants using a worst-match HRTF with individualized ITD had an initial great-circle error above 60°, and a precision response rate below 50%. Additionally, the results of $\mathbf{G}_{\mathbf{current}}$ are very close to those obtained by participants using a random HRTF without ITD individualization in [19], reporting an initial great-circle error of 51° and a precision response rate of 55%. This initial difference could also be attributed in part to the absence of ITD adjustment for $\mathbf{G}_{\mathbf{current}}$ compared to $\mathbf{G}_{\mathbf{previous}}$. The similitude between their initial lateral local angle errors however argues against this hypothesis.

The second striking result is the difference in learning from **L1** to **L3** between the two groups, illustrated in Figure 5. We examine this evolution from **L1** to reduce the contribution of procedural learning which appears most prevalent between **L0** and **L1**. Previous studies results suggest that a certain percentage of non-proficient learners should be expected in every HRTF learning experiment [6, 7, 13]. This percentage is well above zero however, even in non-favorable conditions as when participants train with a worst-match HRTFs (*e.g.* 5 non-learners out of 8 participants in the W10 group in [7]). As such, it is unlikely that the observed difference is due to an unlucky participant selection in the current experiment. The next obvious factor might be the HRTF. Looking at the results of participants training on their worst-match HRTF in [6, 7], some participants still show a significant improvement between evaluations **L1** and **L3** in these studies. Still, it might be that the KU100 HRTF used in the present study was very different from every $\mathbf{G}_{\mathbf{current}}$ participants' individual HRTF, thereby presenting a significantly lower potential for improvement as argued in [7]. As a reminder, $\mathbf{G}_{\mathbf{previous}}$ participants used their best-match HRTF, selected before **L0** based on subjec-

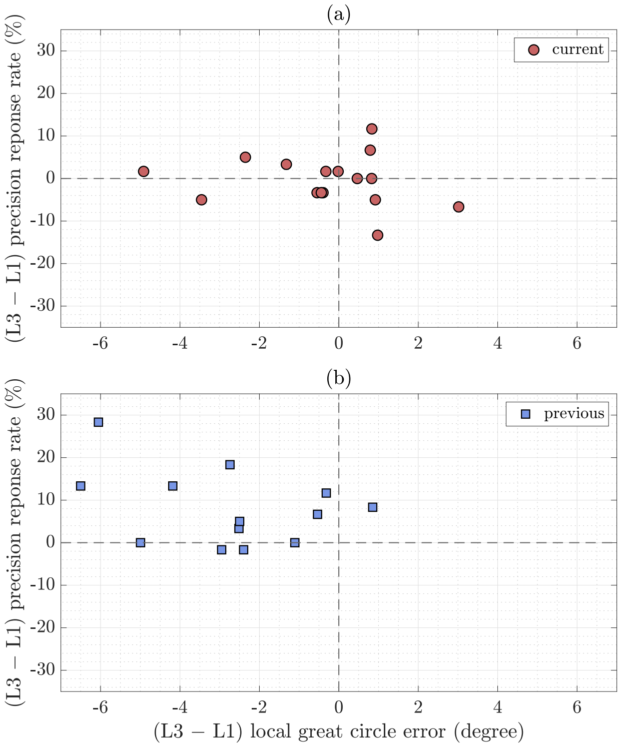

Figure 5: (**L3** − **L1**) difference in individual participant performance for (a) $\mathbf{G}_{\mathbf{current}}$ and (b) $\mathbf{G}_{\mathbf{previous}}$. The Y-axis represents the evolution of precision type response rate, while the X-axis shows the evolution of participants local great-circle angle error. Points in the top-left corner represent participants with the best relative performance improvement between **L1** and **L3**.

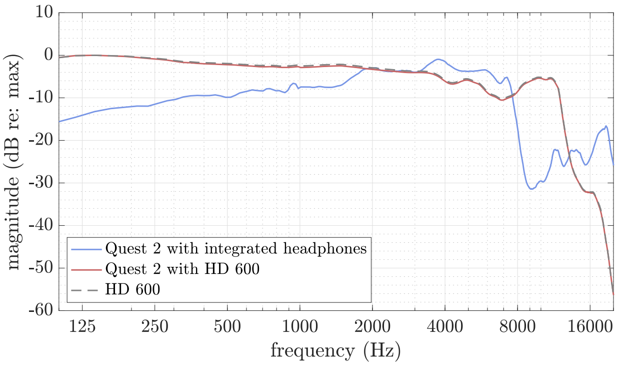

Figure 6: Frequency response (1 channel) of the Quest 2 integrated headphones, measured with a KU100 dummy-head. The logarithmic sine-sweep used as a stimulus was uploaded to the Quest Unity using the PCM (lossless) decompression algorithm. For comparison, the response of the Quest 2 connected (3.5 mm jack output) to a HD 600 headset has been added ("Quest 2 with HD 600"), as well as that of the HD 600 alone, using the same unity application, not running on the Quest 2 but on a computer.

tive affinity among a 7-elements subset from the LISTEN database [17]. This is a serious claim, however, and requires further tests to be evaluated.

Another potential factor might be that $\mathbf{G}_{\mathbf{current}}$ trained with the Quest 2 integrated headphones, whereas $\mathbf{G}_{\mathbf{previous}}$ trained with a Sennheiser HD 600. The frequency responses of both devices were measured on a KU100 dummy head, reported in Figure 6. The usable (± 5 dB) frequency range for the Quest 2 appears to be 300 to 8000 Hz, with a steep drop off after 7500 Hz. Notably, the reduction in energy in the a critical localization spectral cue high-frequency region also could have significantly affected participants' poor training performance in the current study. Comparatively, the measured frequency range for the HD 600 headphones was 40 to 12 000 Hz.

There remain two principal additional factors that could be the reason for $\mathbf{G}_{\mathbf{current}}$ participant's poor performance improvement during the training compared to $\mathbf{G}_{\mathbf{previous}}$: the use of different room acoustics, and the fact that participants trained at home in an uncontrolled environment. Additional tests are required and are being conducted to investigate the role of each element.

If there is a silver lining to the absence of perceptual HRTF adaptation reported after **L1**, it is that the evolution of localization metrics observed between **L0** and **L1** might then be used as a measure of expected procedural learning improvement in those metrics. Based on the existing literature [13], the **L0–L1** evolution leading to the **L1–L3** performance plateau seems indeed both too fast and not large enough to be attributed to perceptual learning. If so, typical procedural learning evolution of those metrics are notably 5° improvement in overall great-circle angle error, 7 % improvement in precision response rate, 4 % in in-cone error rate, and 2 % in off-cone error rates. The remaining metrics showed no marked improvements.

4. SUMMARY AND CONCLUSION

This paper reported the performance evolution of participants training on auditory localization with a non-individual HRTF-BRIR. Surprisingly, said evolution was minimal: performance metrics show a slight improvement after the first training session and none after the second and third (12 min each), suggesting only procedural learning but not perceptual HRTF adaptation. These results were compared to those of a previous study where participants trained with the same program, this time very obviously improving after each session. The comparison suggests that the limited learning could be a result of the headphones/audio-hardware used (Quest 2 integrated headphones), of a lack of affinity between participants and the HRTF used (KU100), or of the uncontrolled nature of the training environment. A complementary study is underway to understand better what happened here. The results of this study, and the follow-up work, should help to design more robust HRTF training programs.

5. ACKNOWLEDGMENTS

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

6. 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*. Cambridge: Academic Press, 1994.

[3] E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman, "Localization using nonindividualized head-related transfer functions," *J Acous*

Soc America, vol. 94, no. 1, pp. 111–123, 1993, doi:10.1121/1.407089.

[4] A. W. Bronkhorst, "Localization of real and virtual sound sources," *J Acous Soc America*, vol. 98, no. 5, pp. 2542–2553, 1995, doi:10.1121/1.413219.

[5] S. Carlile, K. Balachandar, and H. Kelly, "Accommodating to new ears: the effects of sensory and sensory-motor feedback," *J Acous Soc America*, vol. 135, no. 4, pp. 2002–2011, 2014, doi:10.1121/1.4868369.

[6] G. Parseihian and B. F. G. Katz, "Rapid head-related transfer function adaptation using a virtual auditory environment," *J Acous Soc America*, vol. 131, no. 4, pp. 2948–2957, 2012, doi:10.1121/1.3687448.

[7] P. Stitt, L. Picinali, and B. F. G. Katz, "Auditory accommodation to poorly matched non-individual spectral localization cues through active learning," *Scientific Reports*, vol. 9, no. 1, pp. 1063:1–14, 2019, doi:10.1038/s41598-018-37873-0.

[8] P. T. Young, "Auditory localization with acoustical transposition of the ears," *J Experimental Psychology*, vol. 11, pp. 399–429, Dec 1928.

[9] P. M. Hofman, J. G. A. Van Riswick, and A. J. Van Opstal, "Relearning sound localization with new ears," *Nature Neuroscience*, vol. 1, no. 5, pp. 417–421, 1998.

[10] R. Trapeau and M. Schönwiesner, "Adaptation to shifted interaural time differences changes encoding of sound location in human auditory cortex," *Neuroimage*, vol. 118, pp. 26–38, June 2015, doi:10.1016/j.neuroimage.2015.06.006.

[11] D. Poirier-Quinot and B. F. G. Katz, "On the improvement of accommodation to non-individual HRTFs via VR active learning and inclusion of a 3D room response," *Acta Acustica*, vol. 5, no. 25, pp. 1–17, 2021, doi:10.1051/aacus/2021019.

[12] D. Poirier-Quinot and M. S. Lawless, "Impact of wearing a head-mounted display on localization accuracy of real sound sources," *Acta Acustica*, vol. 7, p. 3, 2023, doi:10.1051/aacus/2022055.

[13] D. Poirier-Quinot, M. S. Lawless, P. Stitt, and B. F. G. Katz, "HRTF Performance Evaluation: Methodology and Metrics for Localisation Accuracy and Learning Assessment," in *Advances in Fundamental and Applied Research on Spatial Audio* (B. F. G. Katz and P. Majdak, eds.), ch. 2, Rijeka: IntechOpen, 2022, doi:10.5772/intechopen.104931.

[14] J. J. Faraway, *Extending Linear Models with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. Chapman & HallCRC, 2006.

[15] R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022.

[16] O. Warusfel, "IRCAM Listen HRTF database," 2003, (url). last checked 2018-09-29.

[17] B. F. G. Katz and G. Parseihian, "Perceptually based head-related transfer function database optimization," *J Acous Soc America*, vol. 131, no. 2, pp. 99–105, 2012, doi:10.1121/1.3672641.

[18] D. Poirier-Quinot and B. F. G. Katz, "The Anaglyph binaural audio engine," in *Audio Eng Soc Conv 144*, (Milan), pp. EB431:1–4, May 2018, (url).

[19] M. A. Steadman, C. Kim, J.-H. Lestang, D. F. Goodman, and L. Picinali, "Short-term effects of sound localization training in virtual reality," *Scientific Reports*, vol. 9, no. 1, pp. 1–17, 2019, doi:10.1038/s41598-019-54811-w.