

We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

7,000

Open access books available

186,000

International authors and editors

200M

Downloads

Our authors are among the

154

Countries delivered to

TOP 1%

most cited scientists

12.2%

Contributors from top 500 universities



WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com



HRTF Performance Evaluation: Methodology and Metrics for Localisation Accuracy and Learning Assessment

*David Poirier-Quinot, Martin S. Lawless, Peter Stitt
and Brian F.G. Katz*

Abstract

Through a review of the current literature, this chapter defines a methodology for the analysis of HRTF localisation performance, as applied to assess the quality of an HRTF selection or learning program. A case study is subsequently proposed, applying this methodology to a cross-comparison on the results of five contemporary experiments on HRTF learning. The objective is to propose a set of steps and metrics to allow for a systematic assessment of participant performance (baseline, learning rates, foreseeable performance plateau limits, etc.) to ease future inter-study comparisons.

Keywords: spatial hearing, binaural, localisation accuracy, evaluation, HRTF selection, HRTF training

1. Introduction

If you reached this point, you are probably familiar with the concept of binaural rendering. You likely also know that it is used for producing spatial sound over headphones in most of today's personal mixed reality experiences. While conceptually sound, binaural rendering is subject to several limitations in practice, some of them leading users to perceive distorted versions of the encoded 3D scene. Those distortions range from slight localisation blur to critical scenarios where auditory events are perceived on the opposite hemisphere from their actual position. Researchers have been working on techniques to address this problem of binaural localisation accuracy for some time now. To establish the benefit of these techniques, they predominantly, and quite naturally, rely on localisation performance evaluations.

The problem that concerns us here is that there is no standard for said evaluation. As a consequence, fully appreciating the value of a technique often requires careful reading and interpretation of both protocols and associated results. This becomes truly problematic when comparing the results of several studies, where differences in protocol and evaluation metrics make for complicated analysis at best, simply impossible in some cases. Without inter-study comparison, it becomes

hard to reach any conclusion on the overall and added value of an HRTF selection, synthesis, or learning method. The objective of this chapter is to lay the foundations of such a standard.

1.1 Context

One of the most frequent causes of auditory space distortion in binaural rendering is related to the use of *non-individual* Head Related Transfer Functions (HRTF)¹. An HRTF is a collection of filter pairs that, applied to a mono signal, modify it so that it has the same characteristics as if it had physically been travelling from a specific point in space to our ears. The term HRTF refers to the set of filter pairs, each corresponding to a different source position, typically forming a sphere of fixed radius around the listener. When sound travels to our ears, the acoustic wave interactions with our morphology causes deformations in the perceived signal. From childhood, our brain learned to interpret these acoustic cues as different source positions. Since there exist many variations of ear, head, and torso shapes that each deform the sound differently, so too are there variations in HRTFs. While we are quite adept at sound localisation with our own ears and our own HRTF, the problem arises when we start using someone else's.

In practice, most users will end up experiencing binaural rendering using an HRTF that is not their own, as in the case of a non-individual HRTF, generally taken from an existing database. Presently, measuring an individual's HRTF most often requires specific equipment and access to an anechoic room. Methods exist to simulate an HRTF from geometrical head scans or morphological data, but they suffer the same drawbacks: the techniques are either too costly or burdensome to implement in practical scenarios, or they produce HRTFs that do not exactly match the individual users. As mentioned, using a non-individual HRTF, which the brain has not trained with, often results in distortions of the perceived auditory space. Researchers have been working on this issue, proposing new simulation methods, HRTF selection processes, and even HRTF training programs focused on the reduction of these distortions.

Naturally, all these lines of research end up using a localisation evaluation task to assess the benefit of new techniques. As mentioned above, there exists no standard method for this evaluation, hindering results appraisal and inter-study comparisons.

1.2 Chapter scope and organisation

The objective of this chapter is to outline a set of metrics and propose a methodology to assess localisation performance in the context of HRTF selection and training programs. While the tools proposed can be applied to other contexts, they were designed with HRTF training in mind as not only do they assess instantaneous performance but also performance *evolution*, adding another dimension to the analysis workflow.

Section 2 presents a state of the art of evaluation metrics used to assess localisation accuracy in previous studies. Section 3 introduces the proposed methodology and the set of metrics on which it is built. Section 4 is a case-study, using

¹ We use the term *individual* to identify the HRTF of the user, *individualised* or *personalised* to indicate an HRTF modified or selected to best accommodate the user, and *non-individual* or *non-individualised* to indicate an HRTF that has not been tailored to the user. A so-called *generic* or *dummy-head* HRTF are specific instances of non-individual HRTFs.

the methodology to re-analyse and compare the results of five contemporary experiments on HRTF learning. Section 4 concludes this chapter.

2. State of the art

This section presents and discusses a variety of metrics and methods of analysis introduced in previous studies for the evaluation of auditory localisation performance, in the context of HRTF selection and learning. Further, it discusses what aspect of the data or human behaviour is highlighted by each metric.

2.1 Analysis based on angular distances

The majority of the metrics used in the literature to assess localisation performance are derived from the angular distance from the source position to the participant's response. This section discusses the most common of these metrics, their interpretation, and limitations. It builds upon the work presented in Letowski and Letowski [1].

2.1.1 Egocentric coordinate systems

Many auditory localisation tasks have participants indicating perceived target locations *around them*. As such, egocentric coordinate systems are a logical choice for the assessment of pointing errors. The *spherical* coordinate system, illustrated in **Figure 1a**, uses axes of azimuth and elevation angles. As most researchers are familiar with this coordinate system, it provides an intuitive framework to view and present results.

Alternatively, the *interaural* coordinate system has been proposed to evaluate localisation results as a more natural representation of how sound is perceived. The lateral angle, referred to as the “binaural disparity cue” by Morimoto and Aokata [2], defines *cones-of-confusion* along which the binaural cues of Interaural Level Difference (ILD) and Interaural Time Difference (ITD) are approximately constant. A cone-of-confusion is a set of positions presenting binaural cue/localisation ambiguities, that listeners may not be able to differentiate unless provided with further spectral cues or head movement information [3]. While not truly ‘cones’,

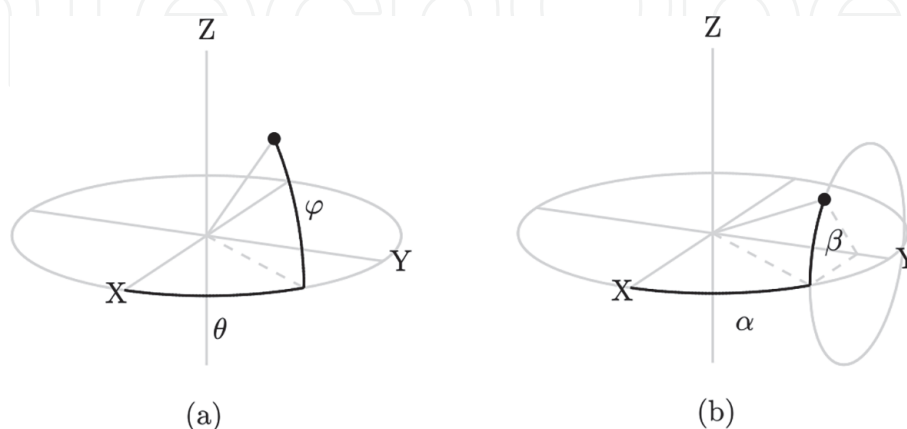


Figure 1. (a) Spherical, and (b) interaural coordinate systems used in the methodology, for a source positioned at angles (55° , 46°) as defined in each coordinate system. Spherical azimuth angle θ is defined in $[-180^\circ:180^\circ]$, elevation angle φ in $[-90^\circ:90^\circ]$. Interaural lateral angle α is defined in $[-90^\circ:90^\circ]$, polar angle β in $[-180^\circ:180^\circ]$. The lateral angle used here is shifted by 90° compared to that originally defined by Morimoto and Aokata [2]. In both systems, listeners are facing X with their left ear pointing towards Y.

these constant ILD or ITD surfaces generally define a circle when the radius is fixed (see [4] for more discussion on the variation with radius of these constant-value surfaces). To maintain accepted terminology in the field, each of these circles is termed a “cone-of-confusion”. The polar angle, or “spectral cue”, is primarily linked with the monaural spectral cues in the HRTF. This independence of binaural and monaural cues makes the interaural coordinate system a compelling choice when assessing localisation performance, particularly when monaural cues are of special interest as in HRTF selection and learning tasks.

Other conventions have been proposed, such as the *double-pole* [5] or *three-pole* [6] coordinate systems. These systems have been designed to circumvent compression issues impacting single-pole (spherical and interaural) coordinate systems, further discussed in Section 2.1.3. They can prove very helpful for some types of data presentation [5], yet can confuse the analysis as more than one coordinate vector can be assigned to any given point in space.

2.1.2 Azimuth, elevation, lateral, and polar errors

Regardless of the coordinate system used, angular errors can be calculated using either the *signed* or *absolute difference* between target and response coordinates. The signed error will give an indication on the “localisation bias” [5] where the absolute error, more often used in the literature [7–10], provides a measure of how close a response is to the target, regardless of error direction. Computing summary statistics from these values can be a first and straightforward step to characterise both the central tendency and dispersion, or “localisation blur” [11], of participant responses [1].

Care must be taken in calculating signed and absolute errors because of the discontinuities in the azimuth and polar angles of the spherical and interaural coordinate systems. If a source is close to the discontinuity and the response crosses it (*e.g.* 179° to -179°), the calculated error will be artificially large. Likewise, summary statistics such as mean or standard deviation should also be computed away from those discontinuities. Another problem that results from working with ego-centric systems is that data distributions will be warped by the sphere curvature, requiring in theory to use circular statistics when comparing statistical distributions. As discussed in [1], linear statistics can however be used in practice if the directional judgements are relatively well concentrated around a central direction.

2.1.3 Compensating for spatial compression

Both the spherical and interaural coordinate systems introduce spatial compression at their poles. In the interaural coordinate system for example, the circumference of the cone-of-confusion at 80° lateral angle is much smaller than that of a cone at 0° lateral angle. Therefore, polar angle errors at the poles (near $\pm 90^\circ$ lateral angle) are more exaggerated than near the median plane. The same problem impacts azimuth errors near the poles (near $\pm 90^\circ$ elevation angle) for the spherical coordinate system.

Previous studies have sought to avoid the spatial compression problem altogether by limiting the analysis to targets away from the poles [12]. The downside of this method is that it limits the scope of the study’s conclusions because a large region of space cannot be studied. Still others have proposed compensation schemes, using for example the lateral angle to weight the response contribution to the average polar error [13–15]. Carlile et al. [13] for example weighted polar response errors using the cosine of the target lateral angle, decreasing response contributions as targets moved towards the interaural axis. This method more

accurately reflects the arc length between the target and response locations on the circle, keeping in mind that this weighting does not take the lateral angle of the response into account.

2.1.4 Using directional statistics to analyse sound localisation accuracy

Due to the discontinuities and spatial compression in the angular metrics of the typical coordinate systems, some work has simply examined the distance between the participant responses and the true target positions to assess the extent of localisation error. The most basic method, the *great-circle error* used in several studies [9, 15, 16], is measured as the distance along the unit sphere between the response and target locations. The great-circle error is independent of the selected coordinate system, not affected by the issues related to discontinuity in the axes or spatial compression.

Great-circle error on its own does not provide information about the direction of the response. Paired with the *angular direction*, it becomes a vector that fully describes the difference between the response and target positions [1]. Similar to *bearing* used to navigate on the globe, angular direction is the angle between the vector of the target towards the positive pole and that of the target towards the response. This vector can be used to compute the mean position of the responses, or *centroid*, and perform directional or spherical statistics. Alternatively, the centroid of the response locations may be calculated by separately summing the x, y, and z coordinates of the responses and dividing by the resultant length [17, 18], though this method may experience some undesirable results for edge cases with widely-scattered locations on the sphere.

To perform statistical analyses of the localisation accuracy, the variance in the response locations must be quantified [19, 20]. Given the two-dimensionality of the data, previous work has used Kent distributions on a sphere [17, 21] to determine ellipses that portray the variance of the data along major and minor axes of the spread of the responses. With Kent distributions, circular statistical tests may be conducted to evaluate the significance of the distance between the centroid of the responses and the target location (such as the Rayleigh z test) or the differences between mean response locations for different conditions (such as the Watson two-sample U^2 test) [22]. Alternatively, Wightman and Kistler [18] suggest the use of the “concentration parameter” κ to characterise the variance, or “dispersion”, of the response locations on the sphere.

2.1.5 Further high level metrics based on angular distances

The *spherical correlation coefficient* has been used to provide an overall measure of the correlation between target and response positions [13, 17, 18]. As with standard correlation, the spherical correlation coefficient ranges from -1 to 1 , where a value of 1 is obtained for two identical data sets, and a value of -1 is obtained for two sets that are reflections of one another. By construction, the spherical correlation coefficient is invariant for global rotations between the two sets.

Rather than looking at single or mean error values to assess localisation accuracy, Hofman et al. [23] and Trapeau et al. [24] studied the linear regression between targets and responses elevation angles. Termed “elevation gain”, the slope of this regression provides a higher level metric that can be used to detect compression or dilation effects in participant responses. Van Wanrooij and Van Opstal [25] extended this technique, applying the regression on target versus response azimuth

as well as elevation angles. To account for azimuthal dependence of the elevation gain, they also introduced the notion of “local elevation gain”, averaging elevation gain values based on a sliding azimuthal window. This metric allows the assessment of how elevation compression and dilation effects impact different regions of the sphere.

2.2 Analysis based on confusions classification

2.2.1 Confusions classification

An analysis based on angular distances alone would fail to distinguish local accuracy misinterpretations from critical space confusions, where responses are often on the opposite hemisphere from target positions. These kinds of errors are very common in studies using non-individualised HRTFs [8, 10, 26, 27], though they also occur when listening with one’s own ears or HRTF [5].

One of the simplest techniques is that used by Honda et al. [28], which defines a hit-miss criterion based on a threshold great-circle error value. Though intuitive, the method does not provide much information on the nature or potential origin of the confusions.

A slightly more elaborate form of confusion classification was used by Middlebrooks [12], which flags responses as confusions when they are in a different hemisphere than that of the target. To avoid reporting small local accuracy errors as confusions for targets near the hemispheres limits, only those responses with polar angle errors greater than 90° were considered when searching for confusions. The classification thus resulted in three types of “quadrant confusions”: front-back, up-down, and left-right. Majdak et al. [14] further improved the definition, introducing a weighting factor to compensate for polar angle compression near the interaural axis. A comparable strategy was adopted by Carlile et al. [13], excluding from confusion checks those targets too close to the interaural axis.

A parallel classification was proposed by Martin et al. [29], determining confusion types based on cone-of-confusion angle values rather than sphere quadrants. The classification was further refined by Yamagishi and Ozawa [30], Parseihian and Katz [8] and Zagala et al. [16], adding “precision” and “combined” confusions to the already existing confusion types. This classification is discussed in more detail in Section 3.1.4.

2.2.2 Separating angular and confusions errors contributions

Given the relatively high incidence of front-back confusions in non-individual HRTF localisation tasks, results often exhibit a bi-modal distribution [10]. Analyses applied to data that contain a large portion of front-back confusions will have large variance and potentially inaccurate averages. The other confusion types also have a similar, if somewhat less characteristic, impact on the data, artificially inflating localisation errors. As such, it is common practice to split the data to analyse confusions separately from *local* performance [1, 12, 14, 31]. A potential problem with this approach is that excluding data from an analysis may result in an unbalanced data set, which limits the use of classical repeated-measures statistics.

Another approach that preserves the sample size of the data consists of ‘folding’ the responses into the same subspace as that of the target prior to the analysis. This technique has only ever been applied to mirror front-back confusions [18], as it may only apply to very specific circumstances and tends to inflate the power of the resulting conclusions [1].

2.3 Additional analysis methods

2.3.1 Decomposing the analysis across sphere regions

Several studies have shown variations in localisation accuracy as a function of region on the sphere due to, amongst other things, cue interpretation [3] or reporting method [32]. In these cases, decomposition schemes were used to better characterise those variations and understand their origins. As mentioned in Section 2.1.5, Van Wanrooij and Van Opstal [25] for example decomposed the analysis of elevation gain across azimuthal regions. Later, Majdak et al. [14] proposed an analysis split into hemi-fields to detect higher accuracy variations for targets in the rear region. Middlebrooks [12] applied a similar spatial decomposition to detect high variability for responses in the upper-rear quadrant, temporarily excluding them from the analysis to better assess variations in remaining regions. The principal drawback of decomposition is that it reduces the statistical power of the analysis, and can result in unbalanced data sets if responses are not evenly spread across the regions under consideration.

2.3.2 Performance evolution modelling and analysis

For the evaluation of HRTF learning, it is essential to assess the progression of participant performance over multiple sessions. On the assumption that any adaptation to an HRTF is a process with diminishing returns with repeated training sessions, localisation performances may be modelled as an exponential decay $y = y_0 \exp(-t/\tau) + c$ [15, 31]. Here y_0 is the initial performance, t is the time (training day, session, etc.), τ is the improvement time constant, and c is the long term performance. This model of performance over time allows for comparisons between studies, such as determining if different protocols lead to faster learning rates or if better long term performance can be achieved. If the training duration proves insufficient to reach a performance plateau/asymptote, like that seen in Stitt et al., [10], the improvement data may be better modelled using the linear form $ax + b$ [9, 31]. In addition to performance modelling, the correlation between training duration and performance metrics has been used to determine if factors other than training duration, like participant attention, should be considered to explain performance evolution [33].

Analysis of performance evolution can be performed per condition (grouping participants) [8, 10] or per participant [23]. Participant performance evaluation makes it harder to draw general conclusions, but potentially provides deeper insight into performance as not all participants exhibit the same ability to adapt to a new HRTF [24]. This adaptation capacity appears to be a function of initial HRTF affinity or “perceptual quality” [10]. For inter-study comparisons, some form of performance scaling or normalisation may first be required to compensate for such affinities, highlighting performance improvement rather than absolute value [10].

3. Methodology for assessing localisation performance

From the literature review in the previous section, a methodology is derived for assessing binaural localisation accuracy. Though it was designed with a focus on HRTF training programs, it should be applicable to any HRTF-related study interested in localisation performance assessment. Section 3.1 introduces the conventions and metrics used in the methodology, itself detailed in Section 3.2. The metrics

Name	Notion examined
Space coverage statistic	Density and homogeneity of the evaluation grid
Confusion rates	Percentage of errors resulting from cone-of-confusion or quadrant ambiguities
Great-circle error	Overall localisation accuracy
Local great-circle error	overall localisation accuracy, excluding confusions
Local lateral error	Localisation accuracy in the horizontal plane, excluding confusions
Local polar error	Localisation accuracy in the vertical plane, excluding confusions
Local azimuth error	Localisation accuracy in the horizontal plane, excluding confusions
Local elevation error	Localisation accuracy in the vertical plane, excluding confusions
Local lateral compression	Whether localisation errors are distorted systematically towards the median plane ZX , excluding confusions
Local elevation compression	Whether localisation errors are distorted systematically towards the horizontal plane XY , excluding confusions
Local lateral bias	Whether there is a systematic rotational offset on responses around the Z axis, excluding confusions
Local elevation bias	Whether there is a systematic upward offset on responses, towards positive Z , excluding confusions
Per-region metrics	Decomposition of the analysis across target regions
Local responses distribution	Whether two sets of responses, excluding confusions, belong to different spherical distributions (using Kent distribution and circular statistics)

Table 1. Summary of the evaluation metrics used in the methodology, grouped by concept similarity.

proposed along with the notions they examine are summarised in **Table 1** at the end of this section. A MATLAB toolbox for the evaluation of all the metrics discussed here is available online².

3.1 Conventions and evaluation metrics

3.1.1 Coordinate systems

The methodology makes use of both spherical and interaural coordinate systems, illustrated in **Figure 1**. While the spherical coordinate system provides an intuitive perspective on the results, the interaural system has been especially designed to separate the analysis of binaural and monaural cues, as discussed in Section 2.1.1, making it a natural choice for the analysis of HRTF-related localisation performance.

3.1.2 Protocol space coverage

Space coverage is a set of metrics, sc_{angle} and sc_{shape} , designed to provide insight on the density of points tested during the localisation task, as well as on the homogeneity of their distribution on the sphere. sc_{angle} represents the density of the

² MATLAB auditory localisation evaluation toolbox: <https://hal.archives-ouvertes.fr/hal-03265190>.

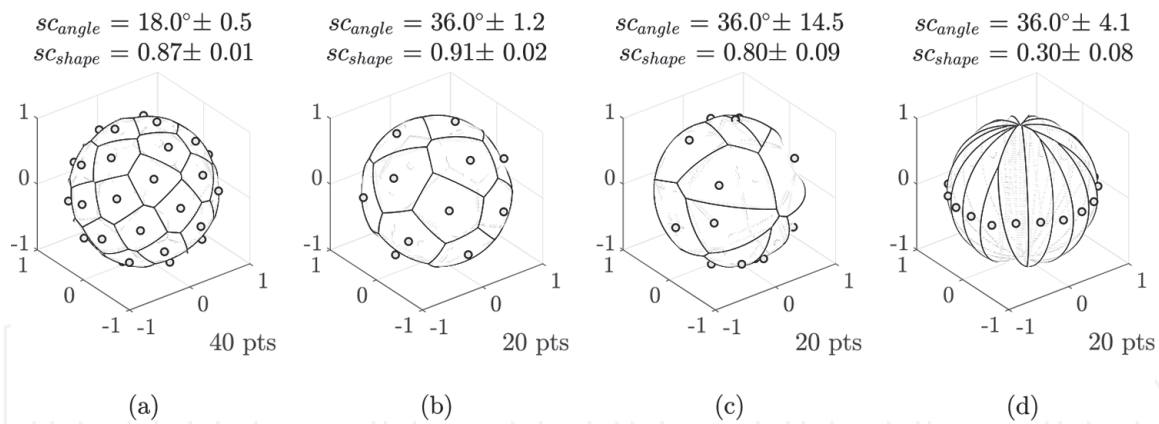


Figure 2. Various test grids and associated space coverage statistics. (a) Homogeneous grid with large number of points, (b) homogeneous grid with small number of points, (c) non-homogeneous grid with small number of points, and (d) horizontal grid with small number of points.

evaluated positions for a given test protocol. It is computed based on the spherical Voronoi diagram built from the evaluated positions, as the average over the solid angles of its cells [34], accompanied, \pm , by its standard deviation. As illustrated in **Figure 2**, denser grids result in smaller sc_{angle} , with standard deviation decreasing for increasingly homogeneous distributions.

sc_{shape} is computed as the average over the shape indices of the cells of the Voronoi diagram, defined as:

$$\text{shape_index} = 4\pi \frac{\text{cell_area}}{(\text{cell_perimeter})^2} \quad (1)$$

where the perimeter is computed as the sum of the great-circle values between the cell vertices, expressed in radians. The squared value of the perimeter, as well as a 4π normalisation factor, are used so that the final shape index value is defined in $[0, 1]$. Cells shaped as circles will have an index close to 1, whereas the index will decrease towards 0 as the cell grows into an elongated polygon. As illustrated in **Figure 2**, sc_{shape} is used in addition to sc_{angle} standard deviation to detect uneven evaluation grid distributions. Note that grid density has a negative impact on sc_{shape} : dropping from 0.91 to 0.84 for uniform grids of 20 and 80 points respectively [35].

3.1.3 Great circle error and angular direction

The great-circle error is defined as the minimum arc between the response and the true target position. This metric provides an intuitive way to assess the local localisation accuracy as the spherical distance between the responses and the target. Given xyz_{target} and $xyz_{response}$ as the vectors in Cartesian coordinates of the target and response positions respectively, the great-circle error is defined in $[0^\circ:180^\circ]$ as:

$$\text{great_circle_error} = \arctan \left(\frac{\|xyz_{target} \times xyz_{response}\|}{xyz_{target} \cdot xyz_{response}} \right) \quad (2)$$

where smaller values correspond to better localisation performances.

The angular direction is coupled to the great circle to enable vector summation of target to response arcs on the sphere. The direction towards the right ear constitutes the positive pole in the interaural coordinate system. The angular direction may then be calculated from the interaural coordinates as:

$$\text{angular}_{\text{dir.}} = \arctan \left(\frac{\cos(\alpha_{\text{resp}}) \sin(\beta_{\text{resp}} - \beta_{\text{target}})}{\cos(\alpha_{\text{target}}) \sin(\alpha_{\text{resp}}) - \sin(\alpha_{\text{target}}) \cos(\alpha_{\text{resp}}) \cos(\beta_{\text{resp}} - \beta_{\text{target}})} \right) \quad (3)$$

where α is the lateral angle and β is the polar angle.

3.1.4 Confusion classification

As discussed in Section 2.2, confusion classification schemes are primarily designed to separate small localisation errors from larger errors caused by erroneous localisation behaviours typically observed in binaural localisation tasks. The scheme used in the methodology is designed around notions borrowed from both cone-of-confusion [8, 10, 16, 29] and sphere quadrant [12, 14] classifications. It separates responses into 4 categories: those near the target (*precision* errors), those opposite the target compared to the YZ plane (*front-back* errors), those within the target cone-of-confusion (*in-cone* errors), and the remainder (*off-cone* errors).

The classification is illustrated in **Figure 3a**. Responses within a 45° radius cone around the target are defined as precision errors. Responses within a 45° cone around the symmetrical of the target position regarding the YZ plane, not already classified as precision errors, are defined as front-back errors. Responses with a lateral angle within 45° of that of the target, not already classified as either precision or front-back confusions, are defined as in-cone errors. Remaining responses are defined as off-cone errors. **Figure 3b** and **c** schematically show several alternate approaches, evaluated before choosing the current method (discussed in more detail below).

The proposed 45° threshold value is somewhat arbitrary, based on a segmentation of localisation error distributions of responses from previous studies [8–10]. This value can be adapted depending on the context of the study and the nominal localisation accuracy expected. To improve understanding, the

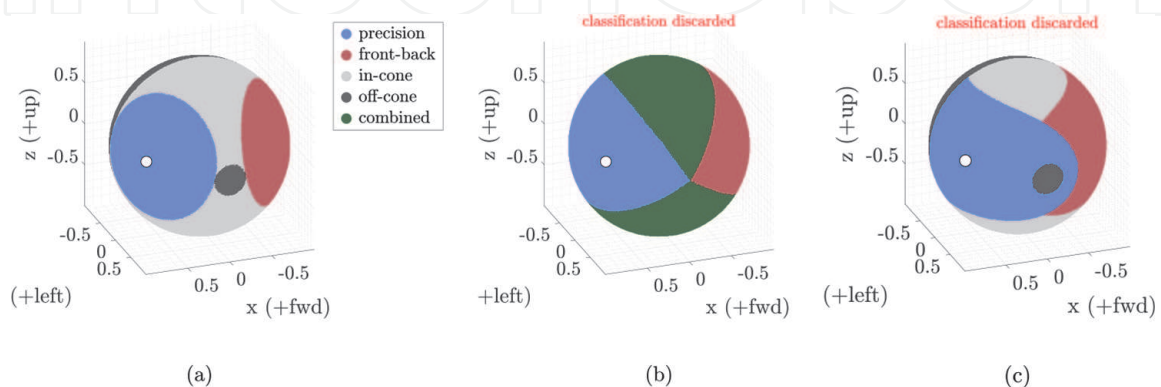


Figure 3. Confusion type as a function of response position on the sphere, for a target at spherical coordinates (35°, 10°) and a listener facing X with his left ear pointing towards Y. (a) Proposed classification scheme, (b) classification used in Stitt et al. [10] based on polar angle only, and (c) attempt at solving pole compression issues of (b).

evolution of confusion zones for a 20° threshold and various target position is illustrated in **Figure 4**. The sum of the four confusion category rates always sums to 100%.

The distinction between in- and off-cone confusions is inspired from the duplex theory [36, 37], separating responses based on whether they are caused by misinterpreting monaural cues (in-cone confusions) or binaural cues (off-cone confusions). The commonly cited front-back confusion category has been maintained, despite not having a clearly identified origin in signal symmetry, as it represents a behaviour frequently observed in localisation studies [38]. Other confusion categories have been considered for this scheme, such as up-down or combined up-down-front-back confusions. They have been discarded however, as their representative patterns were not prevalent in the ≈ 10000 participant responses analysed in Section 4 or the meta analysis on ≈ 80000 responses in free field by Best et al. [38].

Compared to traditional cone-of-confusion classifications defined using only polar angle [8, 10, 16, 29], the main drawback of the proposed scheme is that it is susceptible to ITD mismatch. By only looking at the difference in *polar* angle between target and response, these classifications are not impacted by participants misinterpreting the ITD of the target, focusing on monaural cues interpretation characterisation. As illustrated in **Figure 3b**, the problem of these classifications is that they have high rate of false error detection at the poles of the interaural coordinate system, were a small shift in response can be interpreted as *e.g.* a front-back confusion instead of a precision error.

An attempt was made to propose a new scheme, inspired by the one used in Stitt et al. [10], alleviating the pole issue by increasing the (polar) spread of the precision zone as targets near the poles, constraining said spread to always span 45° of great-circle angle when projected on the sphere. As illustrated in **Figure 3c**, this constraint results in a undesirable warping of the precision error zone for targets within a certain lateral distance from the poles.

The solution proposed for studies needing a classification based on monaural cues interpretation alone is to extend the proposed scheme, artificially adjusting the lateral position of targets prior to the classification to discard errors related to ITD mismatch. This adjustment can be made on a per-participant/target basis, replacing the lateral angle of targets by the mean lateral angle of their associated responses prior to the classification. It can also be performed on a per-response basis by simply assuming that targets and responses always have the same lateral position. The case study of Section 4 uses the second, simple, non-adaptive form of the classification scheme.

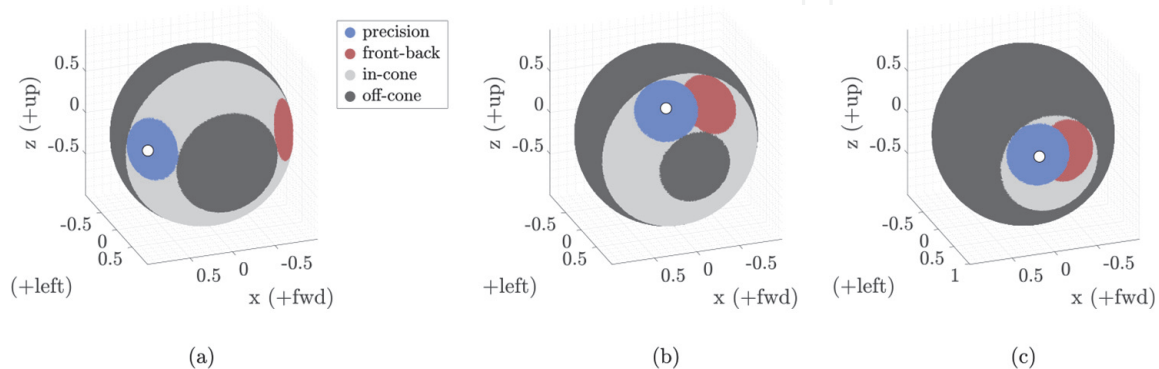


Figure 4. Confusion type as a function of response position on the sphere for the proposed classification scheme with an angle threshold of 20 and a listener facing X with his left ear pointing towards Y. Target at spherical coordinates (a) (35°, 10°), (b) (70°, 40°), and (c) (80°, 10°).

3.1.5 Azimuth, elevation, lateral, and polar errors and biases

Lateral and polar errors are defined as the absolute difference between target and response positions in interaural coordinates. They are used to project localisation errors onto spatial dimensions associated with separate cues in the HRTF, allowing for an analysis of their independent contribution to the overall performance. Both are defined in $[0^\circ:180^\circ]$, where smaller values correspond to better localisation performances. In the methodology, lateral and polar errors will be evaluated only on responses classified as *precision* confusions, hence referred to as *local* lateral and polar errors. This limitation allows to avoid the discontinuities discussed in Section 2.1.2 as well as the hazardous interpretation of values compounding local errors and spatial confusions.

As mentioned in Section 2.1.3, compression at the poles will lead to artificially inflated polar errors for targets near the interaural axis. A weight, proportional to the target lateral position, can be applied to the polar error to compensate for the compression, defining the *polar error weighted* as:

$$\text{polar_error_weighted} = \text{polar_error} * \cos(\alpha_{\text{target}}) \quad (4)$$

This weight is designed so that, for a target and a response that share the same lateral angle, the polar error weighted is equal to the arc length (great-circle) that separates them, regardless of said lateral angle. Note that while lateral error is not impacted by pole compression, it ‘folds’ near the interaural axis: random responses will overall have a lower local lateral error for targets in this region. This is a valuable feature of the interaural system when assessing the symmetric contribution of binaural cues (ITD/ILD) to localisation error. It can nonetheless lead to artificially deflated lateral errors when used in a different context.

Azimuth and elevation errors are defined as the absolute difference between target and response positions in spherical coordinates. They correspond to a more traditional projection of spherical coordinates, more intuitive yet no longer guided by auditory cue separation. Like interaural errors, azimuth and elevation errors are defined $[0^\circ:180^\circ]$ and will be used only for local precision evaluation. As for polar error, azimuth error compression near the poles can be compensated for, defining the *azimuth error weighted* as:

$$\text{azimuth_error_weighted} = \text{azimuth_error} * \cos(\varphi_{\text{target}}) \quad (5)$$

In addition to absolute errors, *signed* lateral and elevation errors are used in the methodology. Mean signed errors, referred to as *biases*, are typically used to examine systematic rotational biases, induced for example by an offset between the tracking system used for measuring the HRTF and that used during the evaluation task, or reporting bias. As for absolute errors, usage of both metrics will be restricted to responses classified as precision confusions.

Finally, lateral and elevation *compression* errors are used to highlight space compression and dilation effects. *Lateral compression*, is defined as $\|\alpha_{\text{target}}\| - \|\alpha_{\text{response}}\|$, so that a positive error corresponds to a compression towards the median plane ZX. Respectively, a negative error corresponds to a dilation away from the median plane. Similarly, the *elevation compression* is defined as $\|\varphi_{\text{target}}\| - \|\varphi_{\text{response}}\|$, so that a positive error corresponds to a compression towards the horizontal plane XY. Respectively, a negative error corresponds to a dilation away from the horizontal plane. Compression errors are for example used to characterise a pointing bias caused by the reporting interface, or to detect lateral compressions

resulting from an ITD mismatch between the presented HRTF and that of the participants.

3.1.6 Sphere regions

The decomposition of the analysis in sphere regions depends on the context. As such, there exists no one ideal decomposition scheme. To support the case study presented in the next section, the sphere will be split into 6 regions: *front-up* ($x > 0$ and $z > 0$), *front-down* ($x > 0$ and $z < 0$), *back-up* ($x < 0$ and $z > 0$), *back-down* ($x < 0$ and $z < 0$), *left* ($y > 0$), and *right* ($y < 0$). This scheme has been chosen to best highlight region specific behaviours while remaining manageable, based on a preliminary analysis of the experiments studied in Section 4. The redundant *left* and *right* regions have been added for systematic checks on lateralisation discrepancies in participant responses.

3.2 Methodology

The methodology is proposed as a set of analysis steps, each building on the previous one to provide a comprehensive assessment of participants localisation performance.

3.2.1 Evaluation task characterisation

The first step of the analysis is to assess how much of the space, *i.e.* sphere, has been tested during the localisation task. In addition to depicting the grid of tested positions, this step reports its space coverage statistics as defined in Section 3.1.2. This provides readers with a simple set of metrics that reflect the spatial thoroughness of the evaluation, a value they can use to qualify the study's conclusions as well as for inter-study comparisons.

Atypical evaluation grids and their potential impact on participant results should also be discussed here. An evaluation on frontal field positions alone is likely to result in better overall performance compared to one encompassing the whole sphere, due to known variations of perceptual accuracy across sphere regions [5]. When using such grids, reporting metrics chance rates, *i.e.* their values for responses randomly distributed on the sphere, as proposed by Majdak et al. [14] can greatly help readers appreciate the presented results. Another problematic example is the use of evaluation grids sparse enough for participants to identify and recall the tested positions, likely impacting participants performance and associated conclusions.

Finally, the stimulus characteristics (type, duration, *etc.*) as well as the reporting method should be described and discussed here, so that any systematic bias they may have on participant responses can be detected during the analysis.

3.2.2 Assess global extent of localisation error

The objective here is to get a rough overview of participant performance during the localisation task, simply answering the question "how far were responses from the true target position?". The assessment is based on the great-circle error as defined in Section 3.1.3.

3.2.3 Assess critical localisation confusions

The next step consists in separating small precision errors from critical confusions. The nature and types of confusions is characterised early on as they can have

a critical impact on localisation performance, often far more detrimental than local localisation accuracy issues. This characterisation is performed using one of the classification methods defined in Section 3.1.3.

3.2.4 Assess local extent of localisation error

This next step takes a closer look at responses classified as precision errors, *i.e.* the non-confused responses, to examine the local localisation performance. The mean great-circle error and angular direction of responses classified as precision confusions is computed to analyse the extent of local errors. Note that this metric does not depend on the confusion classification method used, as precision errors are defined using the same criterion in both methods. Conclusions drawn from this local analysis should naturally be leveraged by the percentage of responses it encompasses.

3.2.5 Horizontal and vertical decomposition of the localisation error

Whether or not this step should be included in the analysis, and which metrics it should make use of, depends on the context of the study. An experiment focusing on perceptual ITD adjustment for example would likely make use of both local lateral error as well as lateral compression. A training program attempting to fine tune participant interpretation of monaural cues would on the other hand base its evaluation on the local polar error. For some studies, this decomposition will not make sense and should be avoided to limit Type I error inflation.

3.2.6 Decompose the analysis across sphere regions

This final step consists in repeating all of the above, decomposing the analysis based on target positions to assess how participants fared in specific regions of the sphere. Given the loss of statistical power and the additional clutter that this analysis represents, it only needs to apply to those studies interested in characterising spatial imbalances in performance. The decomposition can then be performed using either a sphere splitting scheme as the one described in Section 3.1.6, or on a per-target position basis. For example, this approach can be used to support the design of HRTF learning programs that would focus dynamically on those regions/confusions that are the most problematic [9].

To further characterise local localisation behaviours, the analysis can be completed by evaluating average response positions and spherical response distributions. The former, computed by summing local great-circle error *vectors*, as discussed in Section 3.1.3, will help characterise variations of localisation accuracy across sphere regions [21]. The latter, characterised using Kent distributions (see Section 3.1.3), will provide the statistical framework to assess the significance of those variations.

4. Case study

The methodology defined in the previous section is applied here to build a comparative analysis on a selection of studies, focusing on the use of, and adaptation to, binaural cues for auditory localisation. The objective of this case study collection is not so much to present a thorough comparison of these studies as to illustrate how the methodology can be applied to a practical use case, and how its constituting metrics react to concrete scenarios. To further focus the case-study on

these points, significance assessment is based on the overlapping of estimated distributions Confidence Intervals (CIs) rather than on null-hypothesis tests [39].

4.1 Study selection overview

Several studies of the impact of HRTF training on localisation accuracy have been selected from existing literature, for which authors graciously provided raw participant data used in the comparative analysis. A short description of each study is provided in the next section, reporting only those elements that concern the present analysis.

Common to most of the presented studies is the notion of HRTF *perceptual quality*. This term refers to the perceptual matching, localisation wise, between a participant and an HRTF. A low quality HRTF is one that results in bad localisation accuracy. Inversely, the higher the quality, the better the localisation accuracy, the highest quality match corresponding in theory to one's own HRTF. Replicating the potential outcomes of selecting an HRTF from an existing database, three degrees of perceptual matching are considered in these studies in addition to individual HRTF: *worst-match*, *random-match*, and *best-match* HRTF. Best and worst-match HRTFs represent respectively a best and worst case outcome, typically obtained by asking participants to perform a localisation task with, or a perceptual ranking of, an existing set of HRTFs.

4.1.1 Study description: *exp-majdak*

Majdak et al. [14], a 2010 study on the impact of various reporting methods during training with their individual HRTF. 10 participants trained on auditory localisation: 5 reporting perceived localisation positions with their hand, 5 with their head. Each participant completed 600–2200 localisation trials over a span of 2–32 d. Training and evaluation were performed within each trial: a session was composed of 50 trials, completed in 20–30 min. Each trial consisted of a localisation task with feedback, testing participants on 1380 positions overall, distributed on a sphere, using a 500 ms burst of white noise as stimulus. As the reporting method proved to have only a small impact on training efficiency, the 10 participants have hereafter been aggregated in a single group (**grp-majdak-indiv**), focusing the analysis on the impact of HRTF quality on performance evolution.

4.1.2 Study description: *exp-parseihian*

Parseihian and Katz [8], a 2012 study on accommodation to non-individual HRTF. 12 participants trained on auditory localisation, each completing 3 sessions of 12 min each on 3 consecutive days. Each session consisted of an interactive audio localisation game followed by a localisation task evaluation testing participants on 25 positions distributed on a sphere, using a 180 ms sequence of white noise bursts as stimulus. Before training, each participant ranked a set of 7 *perceptually orthogonal* HRTFs [40, 41] from the LISTEN database [42] based on localisation accuracy as perceived during predefined audio trajectories. The best and worst-match HRTF for each participant was extracted from this ranking. Participants were then divided into 3 groups: 2 that trained with their individual HRTF (**grp-parse-indiv**), 5 with the best-match HRTF (**grp-parse-best**), and 5 with the worst-match HRTF (**grp-parse-worst**). An additional 2 groups that performed only 1 training session are not considered in the current analysis. The ITDs of all HRTFs were adjusted based on individual participant head circumference, using a model derived from a regression between measured ITDs and morphological parameters. This technique is used as a

practical method, easily carried out by end-users, to maximise initial localisation performance accuracy.

*4.1.3 Study description: **exp-stitt***

Stitt et al. [10], a 2019 study on accommodation to non-individual HRTF. 16 participants trained on auditory localisation, each completing 10 sessions of 12 min each over a span of 10–20 weeks. The worst-match HRTF selection, training game, stimulus, and tested audio source positions during the localisation task evaluation at the end of each training session were the same as those of **exp-parseihian**. Participants were divided into 2 groups: 4 training with individual HRTFs (**grp-stitt-indiv**) and 8 with worst-match HRTFs (**grp-stitt-worst**). An additional 8 participants trained for only 4 sessions with their worst-match HRTFs are not considered in the current analysis.

*4.1.4 Study description: **exp-steadman***

Steadman et al. [15], a 2019 study on accommodation to non-individual HRTF. 27 participants trained on auditory localisation, each completing 9 sessions of 12 min each over a span of 3 d. A localisation task evaluation was conducted at the beginning and end of each day as well as between each training session the first day, testing participants on 12 positions distributed on a sphere using a 1.6 s stimulus merging bursts of white noise and speech signal. All participants trained with the same randomly-matched HRTF selected from the 7 LISTEN database of **exp-parseihian**. Participants were distributed in 3 groups, training on various gamified and interactive versions of an audio localisation game, aggregated as one group in the current analysis (**grp-steadman-random**). An additional 9 participants, acting as a control group not undertaking training, are also not considered in the current analysis, as well as the results of a parallel evaluation task performed on another HRTF than that used during training.

*4.1.5 Study description: **exp-poirier***

Poirier-Quinot and Katz [9], a 2021 study on accommodation to non-individual HRTF. 12 participants trained on auditory localisation (**grp-poirier-best**), each completing 3 sessions of 12 min each over a span of 3–5 d. Participants trained using a best-match HRTF selected from the 7 LISTEN database of **exp-parseihian**, though the simplified subjective selection method was only concerned with identifying the best-match HRTF. An additional 12 participants trained with their best-match HRTF in a reverberant condition are not considered in the current analysis. Each session consisted of an interactive audio localisation game followed by a localisation task evaluation testing 20 positions distributed on a sphere using the same stimulus as in **exp-parseihian**.

4.2 Application of the methodology

4.2.1 Time alignment of evaluation sessions

In all these experiments, the training sessions lasted for 12 min, except for **exp-majdak** where both training and evaluation were performed in a single block of 20–30 min. According to **exp-majdak**, the evaluation itself took half that time, leaving a per-session training duration equivalent to that of the other studies. A time realignment across experiments was executed such that the evaluation sessions

compared are separated by equivalent training durations. Thus, the sessions have been renumbered to account for changes in protocol.

In the analysis, evaluation sessions are numbered from 1 to 11, each separated by a 12 min training. **Exp-poirier** and **exp-parseihian** only performed 3 training sessions, hence the missing data-points in subsequent figures. Likewise, **exp-stitt** and **exp-majdak** did not report pre-training performances, missing session 1 data-points. Finally, the number of evaluations in **exp-steadman** spreads out from session 4 onward, switching from an evaluation session after each training to an evaluation at the beginning and end of each 3-sessions training day.

4.2.2 Evaluation task characterisation

The space coverage of target positions evaluated during the localisation task of each study are reported in **Figure 5**. The high density of the grid of **exp-majdak** results in a very low average sc_{angle} compared to those of the other experiments. Its comparatively high standard deviation is due to the absence of test positions in the bottom part of the sphere (polar gap). For comparison, a homogeneous grid with the same number of points would have yielded $sc_{angle} = 0.5^\circ \pm 0.003$. Distribution homogeneity is also responsible for the lower sc_{angle} standard deviation value observed in **exp-poirier** compared to that of **exp-parseihian** and **exp-stitt**. Finally, **exp-steadman**, with fewer test points and a polar gap in the bottom hemisphere, has the highest sc_{angle} value and standard deviation.

As could be expected, all the grids present high sc_{shape} values, being overall evenly distributed on the sphere. Grid density around polar gaps impacts the metric, explaining why **exp-poirier** value is higher than that of **exp-majdak** while both grids are evenly distributed: removing polar gap contributions in these grids would yield sc_{shape} values of 0.91 and 0.84 respectively.

Two different reporting methods were used in the five studies: head pointing (**exp-majdak** and **exp-steadman**) and hand pointing (**exp-majdak**, **exp-parseihian**, **exp-steadman**, **exp-poirier**). This should have little to no impact on the comparative analysis however, as both methods lead to similar reporting biases [32]. **exp-parseihian**, **exp-stitt**, and **exp-poirier** used the same stimulus: a 180 sequence of three white noise bursts. **Exp-majdak** used a slightly longer, unique burst of 500 ms, and **exp-steadman** used a 1.6 s stimulus composed of both white noise bursts and speech signal. All these stimuli are likely to present the transient energy and the broad frequency content necessary for auditory space discrimination [43, 44]. The difference in stimulus duration may have

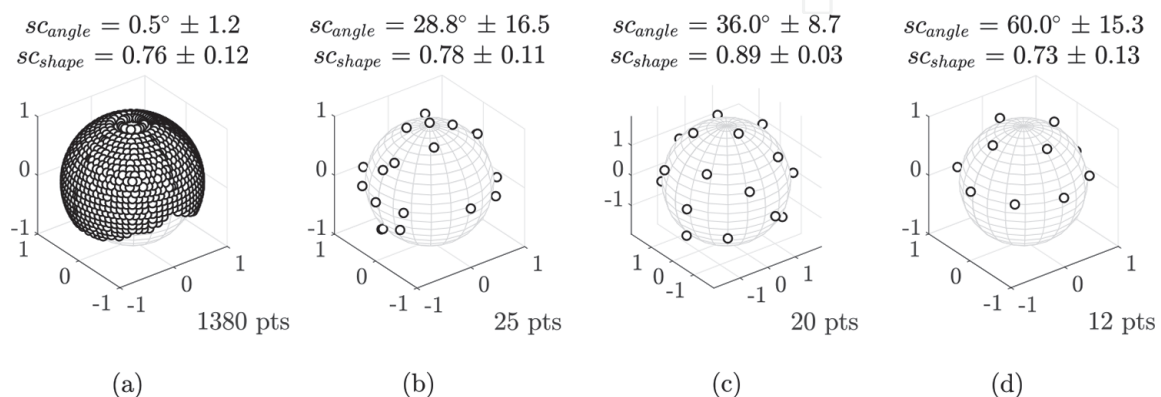


Figure 5. Space coverage statistics of the evaluation task in the selected studies (a) majdak, (b) parseihian/stitt, (c) poirier and (d) steadman.

repercussions in the analysis, as the participants can initiate more head movements to facilitate auditory localisation during the presentation of longer stimuli [45]. While adaptive rendering (*i.e.* dynamic cues) was disabled during stimulus presentation in **exp-parseihian**, **exp-stitt**, and **exp-poirier**, this is not explicitly stated in **exp-majdak** and **exp-steadman**.

4.2.3 Assessing the global extent of localisation error

The evolution of great-circle angle error across studies and training sessions is reported in **Figure 6**. Besides the clear benefit of training observed in all studies, the metric also highlights the overall positive impact of HRTF quality on initial performance. Interestingly, while the results from **exp-parseihian** suggest a similar intra-HRTF quality/performance relationship, it reports larger great-circle angle errors compared to those of the other experiments. This point already illustrates how differences in evaluation protocols or inter-participant variations may complicate the comparison of results across studies, as discussed in Section 4.3.

4.2.4 Assessing the critical localisation confusions

Much like the great-circle error, precision confusion rates can be used to assess performance evolution during training, as illustrated in **Figure 7**. Trends observed on initial precision rates and their evolution reflect the observation made on the great-circle error analysis. Precision rates and great-circle angle values are indeed highly correlated across training sessions, with correlation coefficients in $[-1.0; -0.9]$ for all studies. As each confusion rate aggregates all the responses of a participant during an evaluation session however, their CI is by construction often wide enough to confuse the analysis compared to that based on great-circle errors.

This widening of the CIs is particularly apparent in the comparison of the other confusion rates, reported in **Figure 8** for the evaluation that took place after the first training session. While a trend indeed suggests that the amount of confusions increases with decreasing HRTF quality, overlapping CIs often prevent any definite conclusion. Observing these rates can still help inform the analysis, as the poor performance of **grp-parse-indiv** on great-circle error observed in the previous section can be partly attributed to their high in-cone confusion rates, while their off-cone confusion rate is on par with that of **grp-stitt-indiv** and **grp-majdak-indiv**.

Maybe the most interesting use of confusion rates is to decompose the overall performance evolution. As illustrated by its confusion rate evolution in **Figure 9**, **grp-stitt-worst** performance evolution observed in **Figure 6** should, confusion wise, mainly be attributed to improvements in front-back confusions during training.

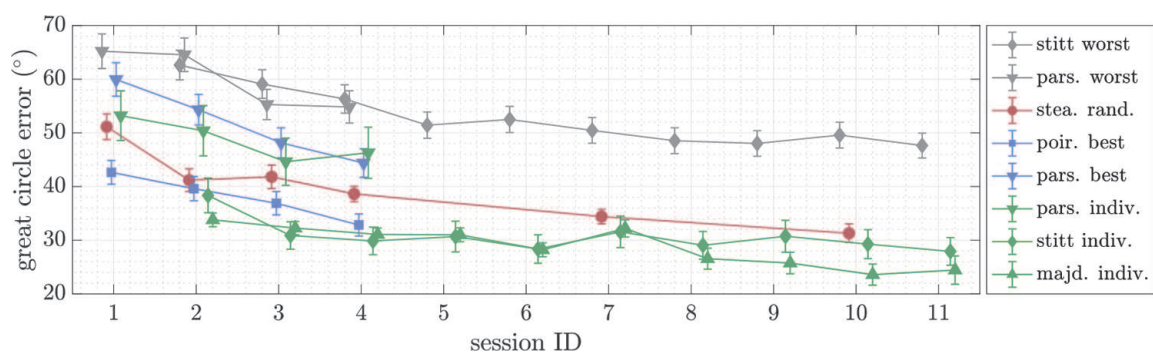


Figure 6.

Great-circle error mean and CI evolution across sessions and experiments. The great-circle error value for random responses is of 90° for all experiments.

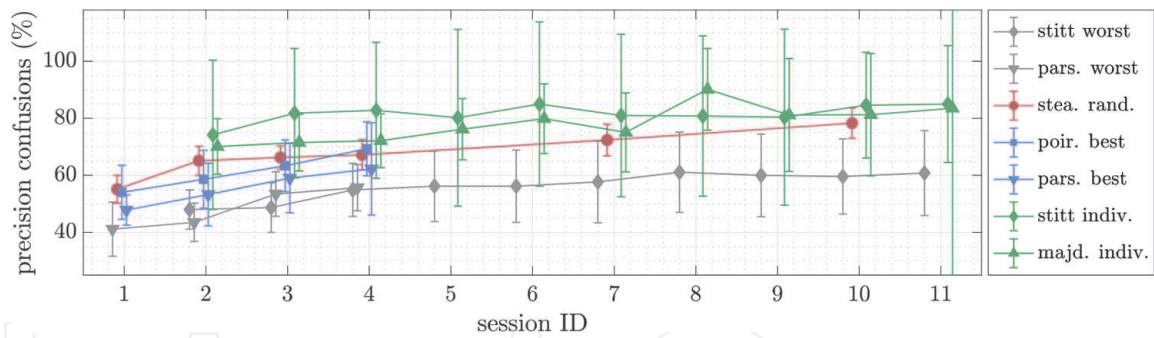


Figure 7. Precision confusion rates mean and CI evolution across sessions and experiments. *Grp-parse-indiv* was removed from the figure, composed of only 2 participants, resulting in a CI so large it confused the whole plot.

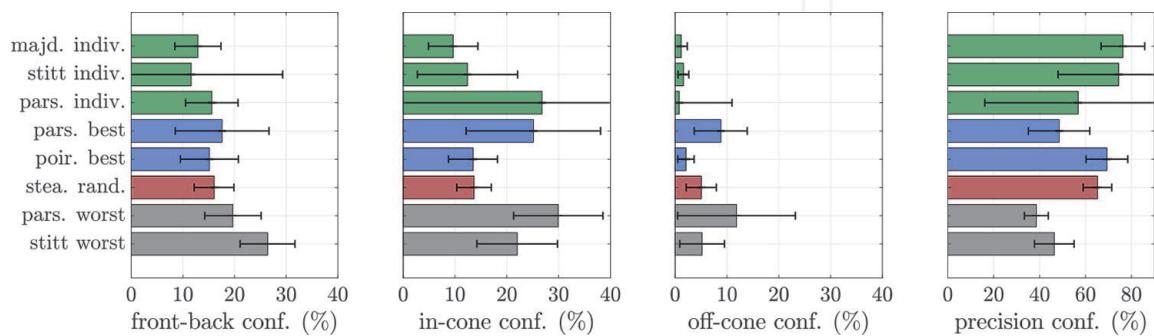


Figure 8. Confusion rates after the first training session across experiments.

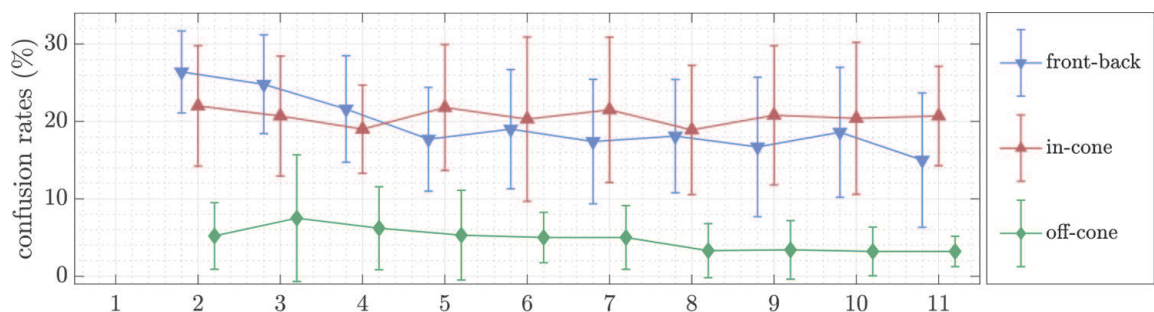


Figure 9. Confusion rates mean and CI evolution across sessions for *grp-stitt-worst*.

4.2.5 Assessing the local extent of localisation error

Results of the confusion classification indicate that roughly 50% of responses were within the vicinity of the target (precision errors) after the first training session across experiments. The analysis here focuses on these responses, assessing local accuracy issues to complete that on localisation confusions.

Figure 10 reports local great-circle errors across training sessions and experiments. Looking once more at **grp-stitt-worst**, their local accuracy did not improve during training, oscillating around 25°. The improvement seen on overall great-circle error for that group can therefore be solely attributed to the reduction in front-back confusions reported in the previous section. Likewise, the 10° improvement on overall great-circle error observed for **grp-parse-worst** between sessions 2 and 3 can be attributed to a reduction in confusion rates, as it does not appear on local great-circle error. Separating the contribution of confusions from that of local accuracy also reveals a significant difference between **grp-stitt-indiv** and **grp-**

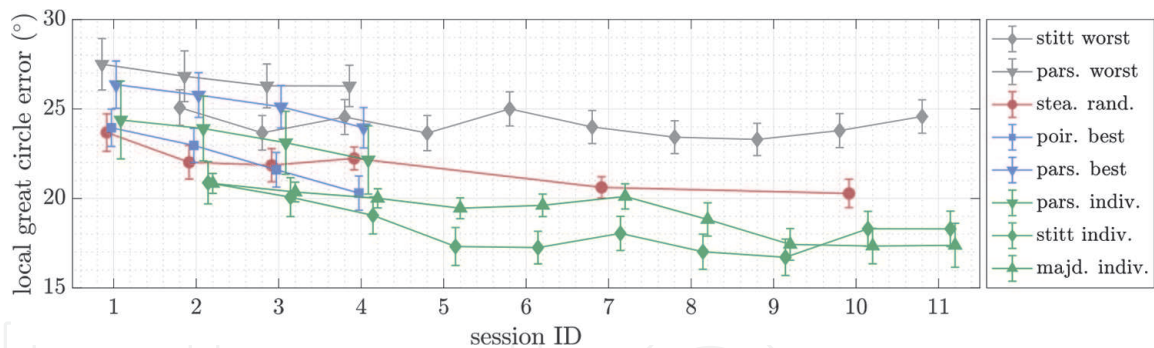


Figure 10.
Local great-circle error mean and CI evolution across sessions and experiments.

majdak-indiv improvement of local great-circle error between sessions 2 and 6, not visible on global great-circle error.

4.2.6 Horizontal and vertical decomposition of the localisation error

Local lateral error evolution across sessions for all experiments is reported in **Figure 11a**. As expected, initial performances indicate that participants using individual HRTF were quite apt at lateral localisation, accustomed as they were to the presented ITD and ILD cues. **Exp-poirier**, **exp-stitt**, and **exp-parseihian** used a similar ITD adjustment scheme, slightly improved in its last iteration for **exp-poirier** compared to that of **exp-stitt**, itself an incrementation on that of **exp-parseihian**. As such, the progression of initial lateral errors between **grp-parse-worst**, **grp-stitt-worst**, and **grp-poirier-best** can be expected. The performance of **grp-steadman-random**, on par with that of participants using ITD-adjusted or individual HRTFs, could be either attributed to the small number of evaluation positions (similar to that used during training), or to the 1.6 s burst and voice stimulus used as compared to the 180 ms to 500 ms burst trains used in the other experiments.

Participants trained with individual HRTF did not improve much on local lateral error overall, starting at $\approx 11^\circ$ after the first training session and only improving to at $\approx 9^\circ$ after the last. Comparison of performance evolution between groups training with a worst-match HRTF (**grp-parse-worst** and **grp-stitt-worst**) against that of groups training with a best-match HRTF (**grp-parse-best** and **grp-poirier-best**) suggests a positive impact of HRTF quality on potential local lateral error improvement. It would also seem that the ITD adjustment applied in **exp-parseihian** and **exp-stitt** was not sufficient to compensate for poor HRTF quality regarding lateral localisation accuracy.

Focusing on local lateral compression evolution, **Figure 11b** reveals a systematic over-estimation of the lateral angle across experiments, *i.e.* participants overall reported targets closer to the inter-aural axis poles than they truly were. Analysis of session 2, after the first training session, indicates that 62% of the 73 participants presented an overall lateral compression of less than -5° , against only 4% presenting one above 5° .

Local polar error evolution across sessions for all experiments is reported in **Figure 12a**. Overall performance was still a function of HRTF quality, but for **grp-parse-indiv** poor performance prior to training and **grp-steadman-random**, on par with **exp-stitt** and **exp-majdak** control groups using individual HRTFs. The impact of training is hardly more pronounced than that observed on local lateral error. Training still helped lower local polar error overall, with even participants using individual HRTFs slightly improving during training: **grp-stitt-indiv** and **grp-**

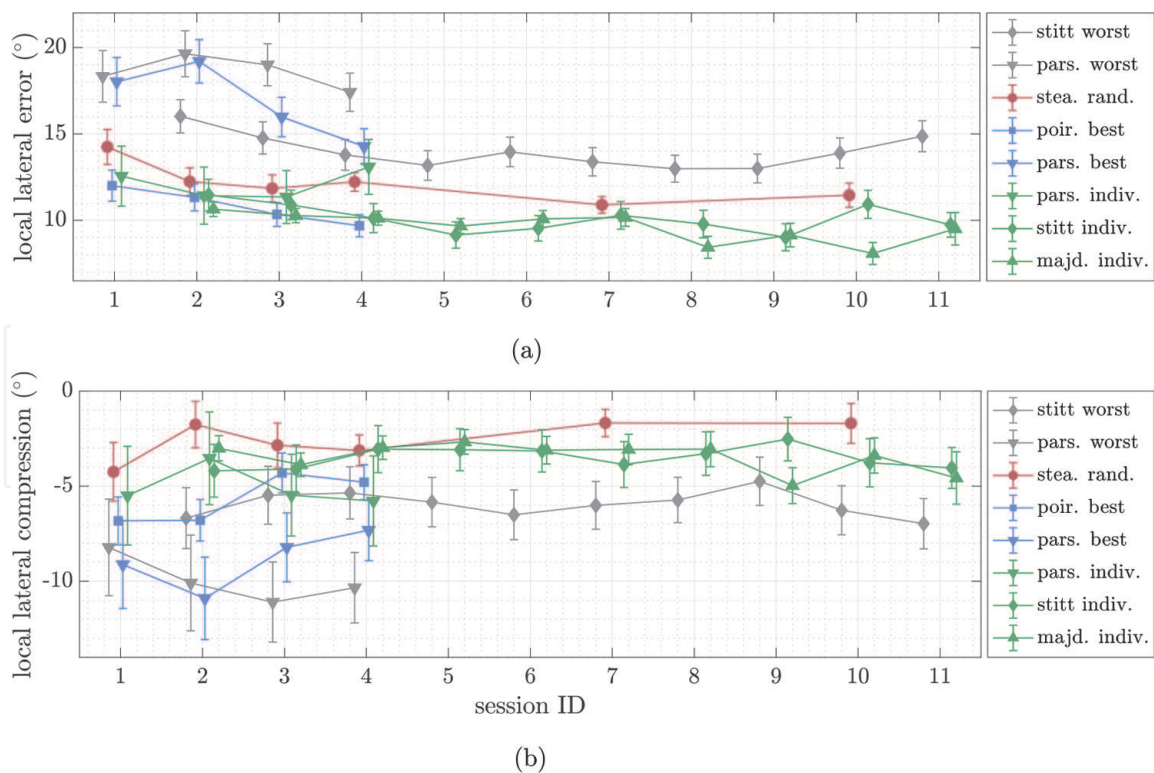


Figure 11. (a) local lateral error, and (b) local lateral compression evolution across sessions and experiments.

majdak-indiv gained $\approx 3^\circ$ in local polar accuracy over the course of training, roughly identical to the improvement observed on local lateral accuracy. Note here that an analysis based on the overall polar error, *i.e.* taking into account confusions, would have suggested $\approx 12^\circ$ improvement after training for these two groups. Finally, most of the improvement on local polar error occurred during the early stage of the training, decreasing of $\approx 7^\circ$ between sessions 1 and 2 in average over all experiments, not considering **exp-stitt** and **exp-majdak** as participants were not tested prior to training, and of only $\approx 7^\circ$ between sessions 2 and 4.

The analysis of local elevation compression also reveals a stronger tendency to under-estimate target elevation, *i.e.* responses closer to the horizontal plane than the true target, than that observed on local lateral compression. Across experiments, 38% of the 73 participants presented a local elevation compression of more than 5° after the first training session, compared to 14% for elevation dilation. A trend suggests that local elevation compression is quickly corrected during the first training session and remains at a relatively constant value regardless of the method or number of training sessions. The surprisingly high plateau reached by **grp-majdak-indiv** compared to **grp-stitt-indiv**, also training on individual HRTFs, could be attributed to the difference in tested grid positions: **exp-majdak** presented far more targets near the 90° elevation pole than **exp-stitt**.

4.2.7 Decompose the analysis across sphere regions

This section illustrates how splitting results analysis across sphere regions might highlight spatial imbalances in performance. To avoid further cluttering the chapter, only two example decompositions will be presented: confusion rates based on sphere regions, and local great-circle error based on individual target locations.

Decomposition of confusion rates based on the regions defined in Section 3.1.6 is illustrated **Figure 13**. Results displayed are aggregated over all five studies, to focus the analysis on general binaural localisation behaviours. The first noticeable result is

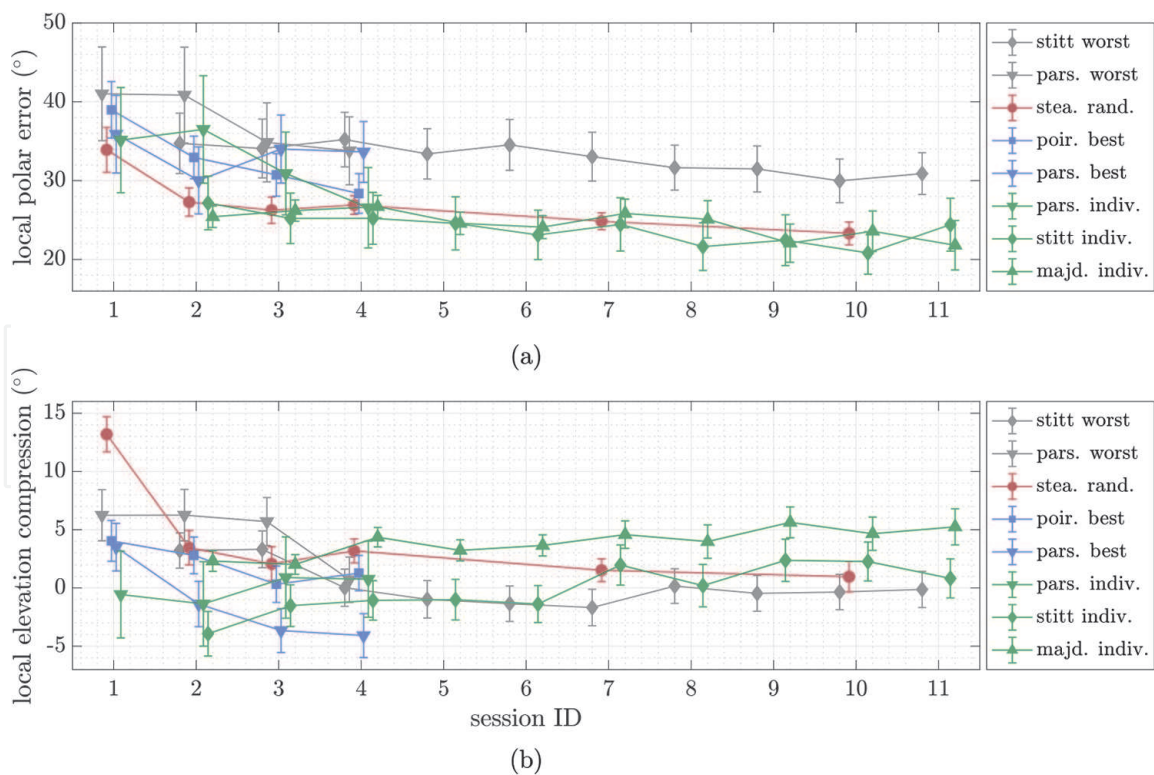


Figure 12. Participants (a) local polar error, and (b) local elevation compression across training sessions and experiments.

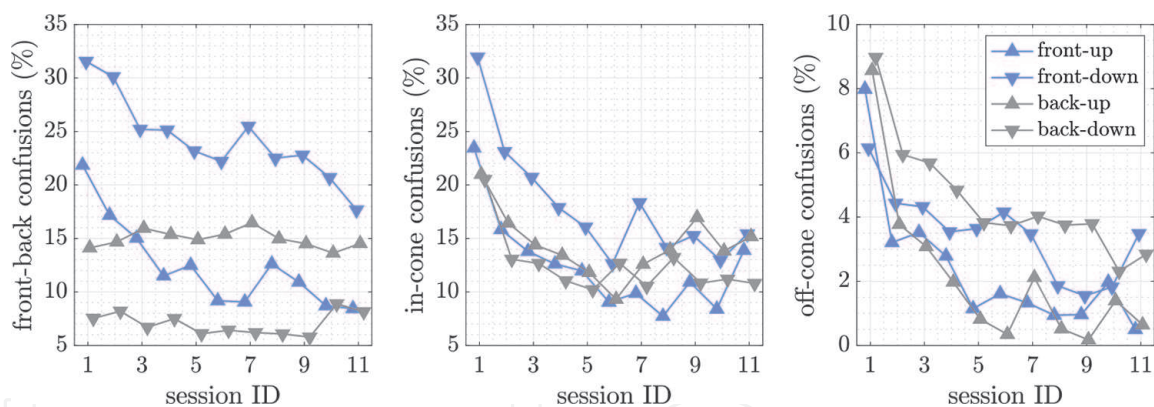


Figure 13. Evolution of confusion rates across sessions, decomposed based on sphere regions, aggregated over all experiments.

that targets in the front-down region were the most susceptible to front-back and in-cone confusions initially, resulting in a very low precision rate (30% vs. 47% and more for the other regions) prior to the first training session. Interestingly, confusion rates in the front-down region were systematically higher than those in the front-up region, for all but off-cone confusions. The initial rate of front-back confusions of targets in front of participants, more than twice that of targets behind them, is likely due to the absence of visual feedback during the localisation task, increasing likelihood of perceiving a sound as behind if they cannot see its source, regardless of HRTF cues.

A second interesting result is the negligible evolution of front-back confusions for targets in the back regions throughout training (*i.e.* back-to-front). While the precision rate of all regions increased, and front-back confusions dropped for front regions, training seemed to have no impact on front-back rates in the back region. Analysis of per-region accuracy however revealed that the local great-circle error decreased evenly across regions, from $\approx 25^\circ$ in session 1 to $\approx 21^\circ$ in session 11.

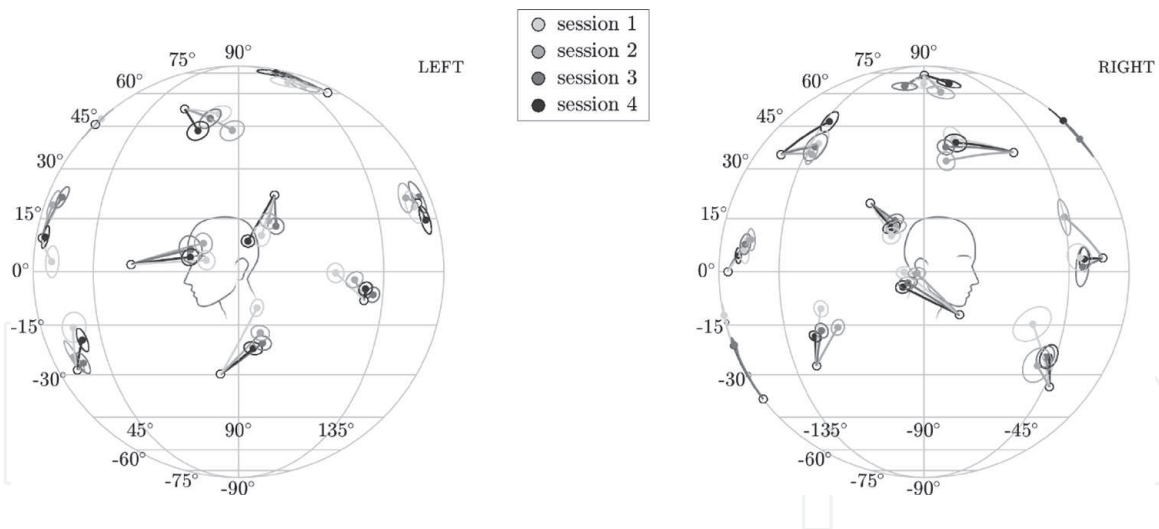


Figure 14. Evolution of mean response locations across targets and sessions in *exp-poirier*. Hollow circles represent target positions. Filled circles represent mean response locations, surrounded by standard error ellipses computed using Kent distributions.

These observations suggest that future training programs could be improved by focusing slightly more on reducing front-back and in-cone confusions in the front-down region. Stagnating rates, such as that of front-back confusions in the back-up region, around 15% across sessions, would also suggest that there is room for improvement in the design of didactic training programs that would aid participants towards reaching 0% confusion rates.

Further refining the analysis, **Figure 14** focuses on the assessment of mean response locations for each target presented in *exp-poirier*. Mean response locations were obtained by summing local great-circle error vectors as discussed in Section 3.2.6. Their positions relative to targets, and the evolution of these positions during training, provides a thorough characterisation of participant's local accuracy evolution on the sphere. Additionally, the lateral and elevation compression effects observed in Section 4.2.6 are clearly visible, where mean responses are generally biased towards the interaural axes and/or the horizontal plane.

4.2.8 Handling initial performance offsets

This additional step in the analysis can be seen as an extension of the evaluation task characterisation proposed in Section 4.2.2 specific to the assessment of localisation performance *evolution*. It presents some of the techniques that exist to compare said evolution despite unbalanced initial conditions across studies or groups of participants.

Techniques have been proposed to conduct training efficiency analysis on unbalanced initial conditions. Stitt et al. [10] for example applied per-participant arithmetic normalisation, based on group baseline performances. Realigning initial conditions, this technique allows to focus the analysis on relative improvement, as illustrated in **Figure 15**.

Another technique for relative improvement comparison, used for example by Majdak et al. [31] and Poirier-Quinot and Katz [9], is to compare the coefficients of a regression applied on performance evolution. As mentioned in Section 2.3.2, two main regression models have been adopted to fit said evolution depending on the training stages represented in the data. **Figure 16** illustrates how both can be fitted to local great-circle error evolution across experiments. Groups performance evolution was first fitted to the exponential form in **Figure 16a**, resorting to the linear

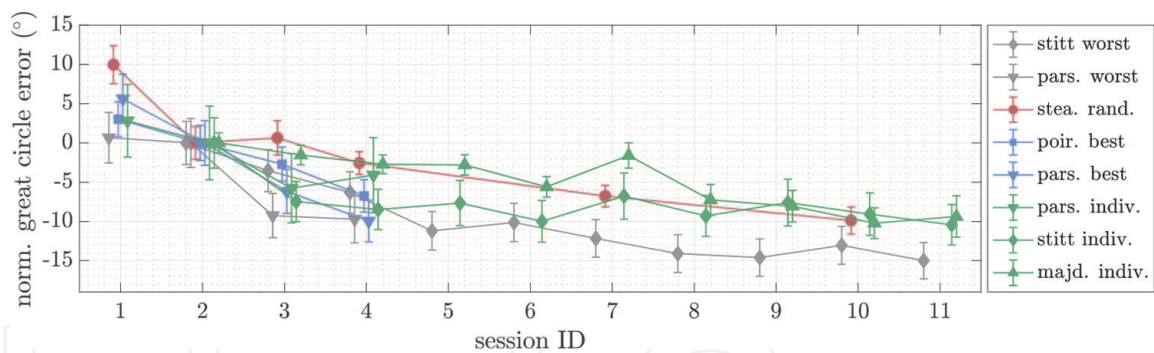


Figure 15. Great-circle error evolution across sessions and experiments. Data normalised (subtraction) with group mean results of session 2 as reference.

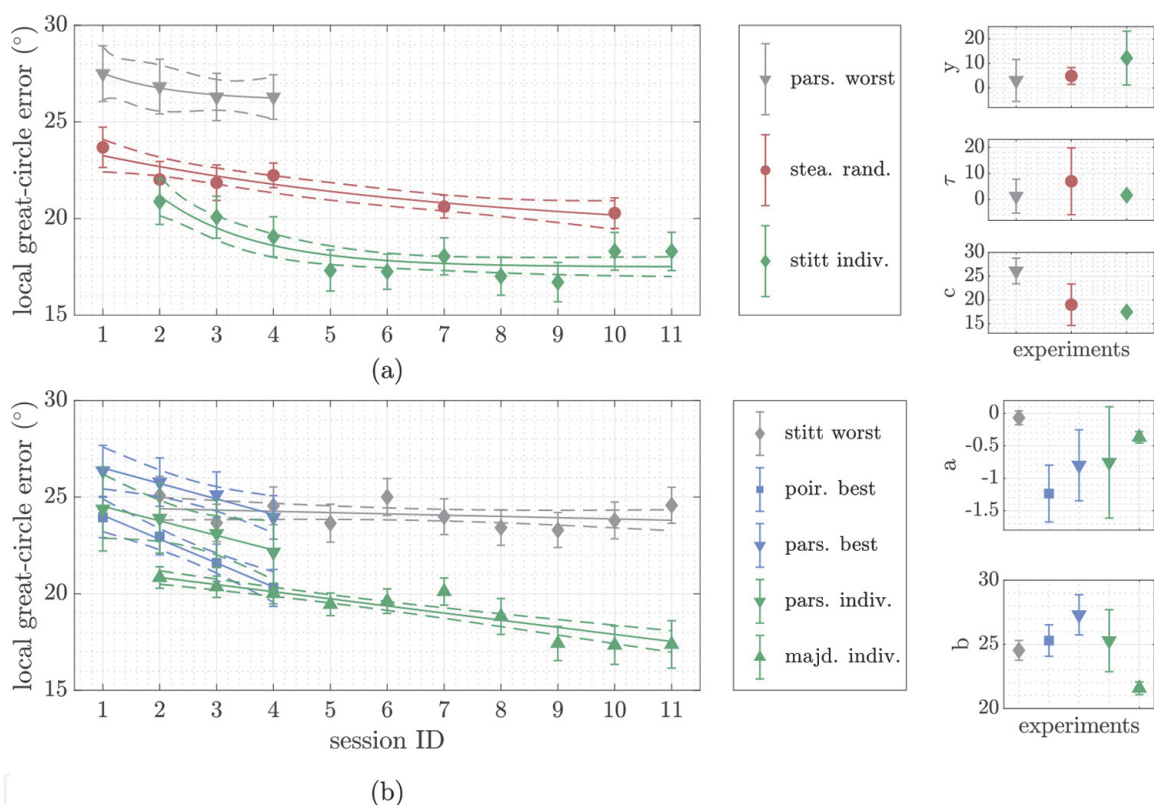


Figure 16. Regressions on local great-circle error evolution across training and experiments, (a) exponential regression “ $y_0 \times \exp(-\text{session}_{ID}/\tau) + c$ ”, and (b) linear regression “ $a \times \text{session}_{ID} + b$ ”. y_0 represents the initial performance, τ the improvement time constant, and c the long term performance. b represents the initial performance, a the improvement rate.

form in (b) when the evolution did not follow an exponential form, resulting in regression parameters CIs so wide as to prevent any meaningful interpretation. The use of a regression is particularly attractive, as it reduces the performance evolution analysis to a simple high level coefficient comparison, coefficients that can usually be interpreted in simple terms such as initial performance or improvement rate.

As mentioned, these techniques are generally applied to compensate for unbalanced initial performance. Although they are perfectly valid to assess the impact of HRTF quality or training efficiency on *relative* improvement, the scope of any conclusion made using them is greatly limited as the potential improvement margin naturally depends on initial performance.

4.3 Discussion

As illustrated throughout Section 4.2, drawing clear cut conclusions from the comparison of results from several studies is difficult at best. Most of the time, it is simply impossible, generally because of uncontrolled variations across test conditions. These variations, limiting both intra- and inter-study analysis, are discussed in this section.

4.3.1 Evaluation task

Variations in the evaluation protocols and procedures between studies in the literature present a challenge for comparing the multiple experiments. Different experimental design choices, such as reporting method, spectral content and duration of the stimulus, and evaluation grid, have a direct impact on the baseline performance of participants [32]. For example, given the choice by **exp-steadman** to use a random-match HRTF, the notable results of **grp-steadman-random** compared to those of the other groups could be attributed to the training program. However, the 1.6 sec stimulus (that may have enabled the use of head movements during the evaluation) may also have contributed to the improved performance of **grp-steadman-random** compared to the other studies that used 180 or 500 ms bursts [46].

The use of a unique grid for localisation tasks across studies would assuredly simplify results comparisons. Said grid could, for example, be designed to be homogeneously distributed on the sphere [35]. For more flexible test conditions, a series of test grids of increasing point densities could be defined, where test positions of any given grid would be present on its higher density neighbours, easing down-sampling for comparison. Regarding the stimulus used or the reporting method, a simple solution would be to settle on those that respectively optimise localisation accuracy [47] and minimise reporting bias [32]. Pending the adoption of common practices, the bias induced by those design choices could technically be assessed from the results of a control group using individual HRTFs.

Another issue when comparing performance evolution across studies is the alignment of the evaluation sessions for fair comparison. As proposed in Section 4.2.1, a simple solution is to align them based on training duration. Time alignment would seem a better option than its alternative, based on the number of positions presented during the training. Time is of direct interest for end-users, and an alignment based on presented positions would bias the analysis in favour of slower exploratory training paradigms.

Finally, the merging of both evaluation and training sessions, as used in **exp-majdak**, is not ideal in the context of inter-study comparison. Although this practice allows for a more granular analysis of performance evolution, it systematically leads to confusing analysis compared to studies alternating between training and evaluation sessions. Additionally, it would seem that the alternating design imposes a lesser constraint on the training paradigm itself, allowing for implicit learning strategies not focused on target localisation [48].

4.3.2 Intra- and inter-participant variations

Variations between participants' performance is an issue common to most psychophysical studies. Two aspects of these variations can become critical in the context of HRTF learning studies.

The first aspect concerns imbalances in initial participant performance across tested conditions. As discussed in Section 4.2.8, such imbalance is likely to weaken

or void conclusions resulting from the analysis. For within experiment comparisons, a simple solution is to run a pre-training evaluation session, to then create groups of equivalent performance based on the metrics used in the analysis. The problem naturally worsens when dealing with inter-study analysis. The use of a control group using individual HRTF is again advised to serve as a baseline reference for the comparative analysis.

The second aspect concerns the difference in participants' immediate sensitivity to HRTF quality, and their ability to adapt to a non-individual HRTF. Both have been discussed in previous studies, where some participants were more prone to instantly benefit from a best-match HRTF [49] or to adapt to a poorly matched HRTF [10]. To avoid missing out on interesting behaviours due to the variance introduced by some participants, it is recommended to conduct a second pass of the analysis on sub-groups, for example aggregated based on their improvement rate [10]. Although the conclusions from the sub-group analysis may be weaker compared to an overall analysis, the technique provides readers with a more thorough understanding of the training as well as the potential advantages and limitations of the tested conditions.

4.3.3 Procedural versus perceptual learning

In the present context, procedural learning refers to participants becoming familiar with the various aspects of the localisation task, resulting in a performance improvement that is not due to an accommodation to HRTF specific cues (perceptual learning). As of yet, there exists no model for *a posteriori* dissociating the contribution of both types of learning to performance evolution. Intra-study comparisons would most likely not be affected since one could generally assume that the procedural learning has a similar impact on all tested conditions. However, by not allowing the procedural learning to plateau before the first evaluation, the generalisation of a study conclusions become problematic when one needs to compare the results from various studies based on different protocols.

Results of control groups generally prove extremely valuable during inter-study comparison. Participants only taking part in the evaluation and not the training, as in **exp-steadman**, can provide a good insight on the impact of the evaluation task implementation on performance across experiments. Even better, the inclusion of a control group using their own HRTF, as in **exp-stitt** and **exp-parseihian**, provides a solid baseline to dissociate procedural from perceptual learning during both intra- and inter-study analysis.

Additionally, simple experimental design choices can be applied to avoid having to deal with certain forms of procedural training. The proprioceptive adjustment required for accurately reporting perceived positions [14] can for example be greatly accelerated by using a natural 3D reporting method coupled to a visual pointer [9], as well as providing a reference grid to help orientation in the sphere [31]. Thorough beta testing can further eliminate design flaws that participants can exploit to improve their performance, such as the use of too small a set of test positions, or unconstrained tracking allowing for small head movements during the stimulus presentation phase of the localisation task.

Other aspects of procedural training, such as having participants focus on the listening task, can only be removed by introducing a pre-experimental training session. Such a session was applied in **exp-majdak**, where participants trained for approximately 30 min on a localisation task coupling visual feedback and stereo panning. This pre-experimental training likely contributed to the smooth improvement in great-circle error by **grp-majdak-indiv** from session 2 onward compared to the disjointed improvement observed for **grp-stitt-indiv** between sessions 2 and 3

in **Figure 6**. Paradoxically, the only limitation of the pre-training proposed in **exp-majdak**, which did not use actual binaural signals, is that it does not familiarise participants with binaural rendering. Pending formal evidence, one may assume that there exists an adaptation process during which participants will grow consistent in their localisation estimation, even in the absence of feedback, much like the effect observed on HRTF quality ratings reported by Andreopoulou and Katz [50]. Regardless of whether this adaptation should be labelled as perceptual or procedural training, it will still interfere with the evaluation of training efficiency itself.

Overall, it is reasonable to assume that one could design a pre-training session that accommodates procedural learning in roughly 15 min, even taking into account this last point, and relaxing the time constraint imposed in **exp-majdak**. This session however still takes a non-negligible amount of time, which will contribute to participant fatigue and loss of focus. Because of this, it is likely that most experimental designs will continue to include aspects of procedural learning as a shared effect, equally impacting all tested conditions. An alternative solution would be to conduct a set of studies to measure and model the various aspects of procedural learning in the present context, so that its contribution to performance evolution could be dissociated from that of perceptual improvement even in the absence of a pre-training session.

5. Conclusion

This chapter presented a methodology for the assessment of auditory localisation accuracy in the context of HRTF selection and learning tasks. Based on existing metrics and decomposition schemes, the methodology consists of a series of steps guiding analysis towards the creation of comprehensive and repeatable performance assessments. A collected case-study was then proposed that compared the results of five contemporary experiments on HRTF learning and illustrates how the methodology can be applied to better understand participant performances and their evolution.

The initial intent of this chapter was to propose a set of metrics and an analysis workflow that would be adopted and adapted by the community to standardise the evaluation of localisation performance. In time, the standardisation would help simplify the comparison of results from different studies, allowing to assess hypotheses and draw conclusions beyond the scope of the constituting studies. While the proposed case-study provides a glimpse at the benefits of such standardisation, it is limited by one of, if not the most, major issue of inter-study comparison: the lack of a reference between tested conditions. Without this reference, conclusions drawn from the analysis can hardly be generalised, much like those that would result from a comparison between language learning techniques without *a priori* knowledge of participants learning abilities, or how different is the language learnt compared to their mother tongue.

As of now, the only applicable solution to provide such reference across studies is to systematically add a control group composed of participants using their own HRTF to the experiment. A large enough group composed of experts and novices alike would indeed provide a stable reference that can be used to assert a certain equivalence in *e.g.* the evaluation task before proceeding to inter-study performance comparison. However, this solution is rarely practical due to the complexity of the HRTF measurement process, which is the main incentive for HRTF learning in the first place. A somewhat less constraining, yet highly unlikely, scenario would be the creation and adoption of a unique evaluation platform, shared across all studies to formalise future HRTF selection methods and training program comparisons.

With luck, the issue will solve itself as the next generation of HRTF individualisation techniques render selection and training obsolete. In the meantime, methodologies such as the one proposed here should help improve the rigour of studies and consequently the understanding of the fundamental issues regarding auditory localisation and spatial hearing accommodation to non-individual HRTFs and their applications.

Acknowledgements

This work was funded in part through a fundamental research collaboration partnership between Sorbonne Université, CNRS, Institut d'Alembert and Facebook Reality Labs. This work was funded in part by the RASPUTIN project (ANR-18-CE38-0004, <https://rasputin.lam.jussieu.fr>) and an associated “Innov’up Faisabilité” grant from the Région Île de France. Portions of this work have been carried out in the context of the Sonicom project, that has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No. 101017743.

Author details

David Poirier-Quinot^{1†}, Martin S. Lawless^{2†}, Peter Stitt^{2†} and Brian F.G. Katz^{2*†}


1 Sciences et Technologies de la Musique et du Son (STMS)—IRCAM, CNRS, Sorbonne Université, Paris, France

2 Sorbonne Université, Paris, France

*Address all correspondence to: brian.katz@sorbonne-universite.fr

† The listed authors contributed equally.

IntechOpen

© 2022 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/3.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 

References

- [1] Letowski T, Letowski S. Localization error: Accuracy and precision of auditory localization. In: *Advances in Sound Localization*, Chapter 4. London: IntechOpen; 2011. DOI: 10.5772/15652
- [2] Morimoto M, Aokata H. Localization cues of sound sources in the upper hemisphere. *The Journal of the Acoustical Society of America*. 1984; **5**(3):165-173. DOI: 10.1250/ast.5.165
- [3] Blauert J. *Spatial Hearing: The Psychophysics of Human Sound Localization*. Cambridge, Massachusetts, United States: MIT Press; 1997
- [4] Katz B, Nicol R. *Sensory Evaluation of Sound*, Chapter Binaural Spatial Reproduction. Boca Raton: CRC Press; 2019. pp. 349-388
- [5] Makous JC, Middlebrooks JC. Two-dimensional sound localization by human listeners. *The Journal of the Acoustical Society of America*. 1990; **87**(5):2188-2200
- [6] Kistler DJ, Wightman FL. A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction. *The Journal of the Acoustical Society of America*. 1992; **91**(3):1637-1647
- [7] Bouchara T, Bara T-G, Weiss P-L, Guilbert A. Influence of vision on short-term sound localization training with non-individualized HRTF. In: *EAA Spatial Audio Signal Processing Symposium*. London: IEEE; 2019. pp. 55-60. DOI: 10.25836/sasp.2019.04
- [8] Parsehian G, Katz BFG. Rapid head-related transfer function adaptation using a virtual auditory environment. *The Journal of the Acoustical Society of America*. 2012; **131**(4):2948-2957. DOI: 10.1121/1.3687448
- [9] Poirier-Quinot D, Katz BF. On the improvement of accommodation to non-individual HRTFs via VR active learning and inclusion of a 3D room response. *Acta Acustica*. 2021; **5**(25):1-17. DOI: 10.1051/aacus/2021019
- [10] Stitt P, Picinali L, Katz BFG. Auditory accommodation to poorly matched non-individual spectral localization cues through active learning. *Scientific Reports*. 2019; **9**(1): 1063:1-1063:106314. DOI: 10.1038/s41598-018-37873-0
- [11] Heffner HE, Heffner RS. The sound-localization ability of cats. *Journal of Neurophysiology*. 2005; **94**(5): 3653-3655. DOI: 10.1152/jn.00720.2005
- [12] Middlebrooks JC. Virtual localization improved by scaling nonindividualized external-ear transfer functions in frequency. *The Journal of the Acoustical Society of America*. 1999; **106**(3):1493-1510
- [13] Carlile S, Balachandar K, Kelly H. Accommodating to new ears: The effects of sensory and sensory-motor feedback. *The Journal of the Acoustical Society of America*. 2014; **135**(4):2002-2011. DOI: 10.1121/1.4868369
- [14] Majdak P, Goupell MJ, Laback B. 3-D localization of virtual sound sources: Effects of visual environment, pointing method, and training. *Attention, Perception, & Psychophysics*. 2010; **72**(2):454-469. DOI: 10.3758/APP.72.2.454
- [15] Steadman MA, Kim C, Lestang J-H, Goodman DF, Picinali L. Short-term effects of sound localization training in virtual reality. *Scientific Reports*. 2019; **9**(1):1-17. DOI: 10.1038/s41598-019-54811-w

- [16] Zagala F, Noisternig M, Katz BFG. Comparison of direct and indirect perceptual head-related transfer function selection methods. *The Journal of the Acoustical Society of America*. 2020;**147**(5):3376-3389. DOI: 10.1121/10.0001183
- [17] Leong P, Carlile S. Methods for spherical data analysis and visualization. *Journal of Neuroscience Methods*. 1998; **80**(2):191-200
- [18] Wightman FL, Kistler DJ. Headphone simulation of free-field listening. II: Psychophysical validation. *The Journal of the Acoustical Society of America*. 1989;**85**(2):868-878
- [19] Edmondson-Jones AM, Irving S, Moore DR, Hall DA. Planar localisation analyses: A novel application of a Centre of mass approach. *Hearing Research*. 2010;**267**(1-2):4-11
- [20] Irving S, Moore DR. Training sound localization in normal hearing listeners with and without a unilateral ear plug. *Hearing Research*. 2011;**280**(1-2): 100-108
- [21] Carlile S, Leong P, Hyams S. The nature and distribution of errors in sound localization by human listeners. *Hearing Research*. 1997;**114**:179-196. DOI: 10.1016/S0378-5955(97)00161-5
- [22] Jammalamadaka SR, Sengupta A. *Topics in Circular Statistics*. Vol. 5. Singapore: World Scientific; 2001. DOI: 10.5772/15652
- [23] Hofman PM, Van Riswick JG, Van Opstal AJ. Relearning sound localization with new ears. *Nature Neuroscience*. 1998;**1**(5):417-421. DOI: 10.1038/1633
- [24] Trapeau R, Aubrais V, Schönwiesner M. 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*. 2016; **140**(2):879-890. DOI: 10.1121/1.4960568
- [25] Van Wanrooij MM, Van Opstal AJ. Relearning sound localization with a new ear. *The Journal of Neuroscience*. 2005;**25**(22):5413-5424. DOI: 10.1523/JNEUROSCI.0850-05.2005
- [26] Wenzel EM, Arruda M, Kistler DJ, Wightman FL. Localization using nonindividualized head-related transfer functions. *The Journal of the Acoustical Society of America*. 1993;**94**(1):111-123. DOI: 10.1121/1.407089
- [27] Zahorik P, Bangayan P, Sundareswaran V, Wang K, Tam C. Perceptual recalibration in human sound localization: Learning to remediate front-back reversals. *The Journal of the Acoustical Society of America*. 2006;**120**(1):343-359. DOI: 10.1121/1.2208429
- [28] Honda A, Shibata H, Gyoba J, Saitou K, Iwaya Y, Suzuki Y. Transfer effects on sound localization performances from playing a virtual three-dimensional auditory game. *Applied Acoustics*. 2007;**68**(8): 885-896. DOI: 10.1016/j.apacoust.2006.08.007
- [29] Martin RL, McAnally KI, Senova MA. Free-field equivalent localization of virtual audio. *Journal of the Audio Engineering Society*. 2001; **49**(1/2):14-22
- [30] Yamagishi D, Ozawa K. Effects of timbre on learning to remediate sound localization in the horizontal plane. In: *Principles and Applications of Spatial Hearing*. Singapore: World Scientific; 2011. pp. 61-70
- [31] Majdak P, Walder T, Laback B. 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. 2013; **134**(3):2148-2159. DOI: 10.1121/1.4816543
- [32] Bahu H, Carpentier T, Noisternig M, Warusfel O. Comparison of different egocentric pointing methods for 3D sound localization experiments. *Acta Acustica*. 2016;**102**(1):107-118. DOI: 10.3813/AAA.918928
- [33] Klein F, Werner S. Auditory adaptation to non-individual HRTF cues in binaural audio reproduction. *Journal of the Audio Engineering Society*. 2016; **64**(1/2):45-54
- [34] Van Oosterom A, Strackee J. The solid angle of a plane triangle. *IEEE Transactions on Biomedical Engineering*. 1983;**2**:125-126. DOI: 10.1109/TBME.1983.325207
- [35] Saff EB, Kuijlaars AB. Distributing many points on a sphere. *The Mathematical Intelligencer*. 1997;**19**(1): 5-11
- [36] Middlebrooks JC, Green DM. Sound localization by human listeners. *Annual Review of Psychology*. 1991;**42**(1): 135-159
- [37] Rayleigh L. XII. On our perception of sound direction. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*. 1907; **13**(74):214-232. DOI: 10.1080/14786440709463595
- [38] Best V, Brungart D, Carlile S, Jin C, Macpherson E, Martin R, et al. A meta-analysis of localization errors made in the anechoic free field. In: *Principles and Applications of Spatial Hearing*. Singapore: World Scientific; 2011. pp. 14-23
- [39] Cumming G. The new statistics: Why and how. *Psychological Science*. 2014;**25**(1):7-29. DOI: 10.1177/0956797613504966
- [40] Andreopoulou A, Katz BFG. Subjective HRTF evaluations for obtaining global similarity metrics of assessors and assessees. *Journal of Multimodal User Interfaces*. 2016b; **10**(3):259-271. DOI: 10.1007/s12193-016-0214-y
- [41] Katz BFG, Parseihian G. Perceptually based head-related transfer function database optimization. *The Journal of the Acoustical Society of America*. 2012;**131**(2):99-105. DOI: 10.1121/1.3672641
- [42] Warusfel O. IRCAM Listen HRTF Database. 2003. Available from: <http://recherche.ircam.fr/equipes/salles/listen> [Accessed: September 29, 2018]
- [43] Dramas F, Katz BFG, Jouffrais C. Auditory-guided reaching movements in the peripersonal frontal space. *The Journal of the Acoustical Society of America*. 2008;**123**(5):3723-3723. DOI: 10.1121/1.2935195
- [44] Kumpik DP, Kacelnik O, King AJ. Adaptive reweighting of auditory localization cues in response to chronic unilateral earplugging in humans. *The Journal of Neuroscience*. 2010;**30**(14): 4883-4894. DOI: 10.1523/JNEUROSCI.5488-09.2010
- [45] Woodworth RS, Schlosberg H. *Experimental Psychology*. Rev. ed. Oxford, England: Holt; 1954
- [46] Wallach H. The role of head movements and vestibular and visual cues in sound localization. *Journal of Experimental Psychology*. 1940;**27**(4): 339-368
- [47] Begault DR, Wenzel EM, Anderson MR. 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*. 2001; **49**(10):904-916

[48] Mendonça C. A review on auditory space adaptations to altered head-related cues. *Frontiers in Neuroscience*. 2014;8(219):1-14. DOI: 10.3389/fnins.2014.00219

[49] Poirier-Quinot D, Katz BFG. Assessing the impact of head-related transfer function individualization on performance: Case of a virtual reality shooter game. *The Journal of the Audio Engineering Society*. 2020;68(4): 248-260. DOI: 10.17743/jaes.2020.0004

[50] Andreopoulou A, Katz B. Investigation on subjective HRTF rating repeatability. In: *Audio Engineering Society Convention*. Vol. 140. New York, United States: Audio Engineering Society; 2016. pp. 9597:1-9597:959710

IntechOpen