System-to-User and User-to-System Adaptations in Binaural Audio

Lorenzo Picinali, Brian F. G. Katz

Chapter 4
System-to-User and User-to-System Adaptations in Binaural Audio

Lorenzo Picinali and Brian F. G. Katz

Abstract This chapter concerns concepts of adaption in a binaural audio context (i.e. headphone-based three-dimensional audio rendering and associated spatial hearing aspects), considering first the adaptation of the rendering system to the acoustic and perceptual properties of the user, and second the adaptation of the user to the rendering quality of the system. We start with an overview of the basic mechanisms of human sound source localisation, introducing expressions such as localisation cues and interaural differences, and the concept of the Head-Related Transfer Function (HRTF), which is the basis of most 3D spatialisation systems in VR. The chapter then moves to more complex concepts and processes, such as HRTF selection (system-to-user adaptation) and HRTF accommodation (user-to-system adaptation). State-of-the-art HRTF modelling and selection methods are presented, looking at various approaches and at how these have been evaluated. Similarly, the process of HRTF accommodation is detailed, with a case study employed as an example. Finally, the potential of these two approaches are discussed, considering their combined use in a practical context, as well as introducing a few open challenges for future research.

4.1 Introduction

Binaural technology is the solution for sound spatialisation which is the closest to real-life listening. It attempts to mimic the entirety of acoustic cues associated with the human localisation of sounds, reproducing the corresponding acoustic pressure signal at the entrance of the two ear canals of the listener (binaural literally means "related to two ears"). These two signals should be a complete and sufficient representation of the sound scene, since they are the only information that the auditory system requires in

L. Picinali (✉)
Imperial College London, South Kensington Campus, London SW7 2AZ, UK
e-mail: l.picinali@imperial.ac.uk

B. F. G. Katz
Sorbonne Université, CNRS, UMR 7190, Institut Jean Le Rond d'Alembert, Lutheries - Acoustique - Musique, Paris, France
e-mail: brian.katz@sorbonne-universite.fr

M. Geronazzo and S. Serafin (eds.), *Sonic Interactions in Virtual Environments*, Human—Computer Interaction Series, https://doi.org/10.1007/978-3-031-04021-4_4

order to identify the 3D location of a sound source. Thus, binaural rendering of spatial information is fundamentally based on the production (either through recording or synthesis) of localisation cues that are the consequence of the incident sound upon the listener's torso, head, and ears on the way to the ear canal, and subsequently to the eardrums. These cues are, namely, the ITD (interaural time difference), the ILD (interaural level difference) and spectral cues [48, 68]. Their combined effects are represented by the Head-Related Transfer Function (HRTF), which characterises the spectro-temporal filtering of a locus of source positions around a given head.[1]

4.1.1 Localisation Cues and Their Individual Nature

The ILD and ITD as a function of source position are determined principally by the size and shape of the head, as well as the position of the ears on the two sides. In order to better understand these localisation cues, Fig. 4.1 shows how ITD and ILD vary as a function of both distance (1.5–10 m) and azimuth. This comparison highlights potential effects of ITD/ILD mismatch, especially if they occur near the interaural axis where they can affect distance perception. The results were obtained by Boundary Element Method (BEM) simulation of the HRTF using the open-source `mesh2hrtf` software [110, 111]. The mesh employed was obtained from an MRI scan of a Neumann dummy recording head (model KU-100), previously used in HRTF computation [32] and measurement [4] comparisons. These cues vary as a function of frequency. For this example, the ITD was calculated using the *Threshold lp –30 dB* method (for a summary of various ITD estimation methods see [50]), which detects the first onset using a –30 dB relative threshold on a 3 kHz low-pass filtered version of the HRIR, as this has been shown to be the most perceptually relevant method for ITD estimation among 32 different estimation methods and variants [7, 50]. The ILD was calculated as the difference of left and right HRIR RMS values, after applying a 3kHz high-pass filter. The use of low-pass and high-pass filters for the two different acoustic cues is based on previous studies showing the frequency dependence of the different auditory cues [101], with ITD being dominated by low-frequency content (with interpretation of phase information being inconclusive for frequencies smaller than head dimensions) and ILD varying more significantly with high-frequency content (where the wavelength is less than the dimensions of the head). The application of a 2–3 kHz filter can be used to generally separate the contributions of the pinnae in the HRIR [50]. One can observe that ITD varies little over the simulated distance range, while becoming more vague and ambiguous near the

[1] We use the term *HRTF* to indicate the set of filters, each representing a pair of transfer functions from a point source in space at a given distance around a given head to the left and right ear, normalised by the transfer function with the body absent. The plural, *HRTFs*, therefore, represents a collection of more than one HRTF, typically for different heads or test conditions. The head-related impulse response or *HRIR* is the time domain transform of the HRTF.

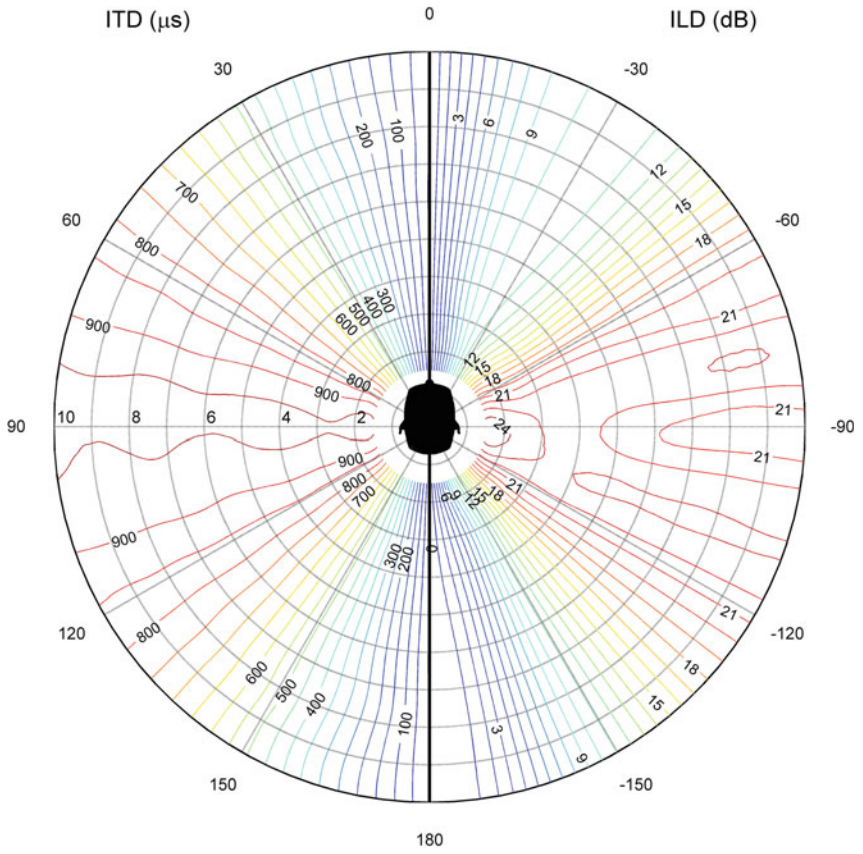


Fig. 4.1 Isocontours for ITD (left) and ILD (right) as a function of azimuth (in degrees) and radial distance (from 1.5 to 10 m) obtained via numerical simulation of the HRTF of a dummy head (not shown to scale). ITD (3 kHz low-pass Head-Related Impulse Response—HRIR, *Threshold*, –30 dB first onset method) 50 μs contours. ILD (3 kHz high-pass HRIR, RMS difference) 1 dB contours (from [48])

interaural axis. In contrast, the ILD varies with distance in the same interaural axis range of 70°–110°.

Other physical interactions between the sound wave and the torso, head, and pinnae (the external parts of the ear) introduce a range of spectral cues (principally through series of peaks and notches) which can be used to judge whether a sound source is e.g. above or below, to the front or rear of the listener, while ITD and ILD remain relatively unchanged. Considering the various morphological regions of the pinnae, as indicated later in Sect. 4.2.1—Fig. 4.2a, each of these is potentially related to specific characteristic of the HRTF filters. As such, individual morphological variations will result in different HRTFs. When reproducing binaural audio, it has been experimentally demonstrated that using an HRTF that does not match the one of

the listener has a detrimental effect on the accuracy and realism of virtual sound perception. For example, it has been noted that listeners are able to localise virtual sounds that have been spatialized using their own HRTFs with a similar accuracy to free field listening, though some studies have shown poorer elevation judgements and increased front-back confusions [67], which may be due to the idealised anechoic nature of HRTFs and the importance of slight head movements and associated dynamic cues [37, 102]. These errors can significantly increase when using someone else's HRTF [99]. Furthermore, using non-individual HRTFs (see Sect. 4.1.2) has been shown to affect various perceptual attributes when considering complex scenes, in addition to those associated with source localisation: i.e. Coloration, Externalisation, Immersion, Realism and Relief/Depth [87]. In this chapter, the primary focus is on localisation as the perceptual evaluation metric. Chapter 5 introduces and discusses other relevant metrics.

4.1.2 Minimising HRTF Mismatch Between the System and the Listener

Various means have been investigated to minimise erroneous or conflicting binaural acoustic localisation cues relative to the natural cues delivered to the auditory system and, as such, improve the quality of the resulting binaural rendering. Majority of research has focused on improving the similarity between the rendering systems' localisation cues and those of the individual listener. This is generally termed "individualisation" or "individualised" binaural rendering. To clarify questions of nomenclature, we propose the following terms:

- *individual* to identify the HRTF of the user;
- *individualised* or *personalised* to indicated an HRTF modified or selected to best accommodate the user;
- *non-individual* or *non-individualised* to indicate an HRTF that has not been tailored to the user and
- *dummy head* or so-called *generic* HRTF sets are specific instances of non-individual HRTFs, often designed with the goal of representing a certain pool of subjects.

While not exhaustive, a general overview of individualisation methods is discussed here.

Binaural Recordings and Synthesis

The first and most direct method to create an individual rendering is to perform the recording with binaural microphones placed in the ear canal of the listener. This is however, in most cases, an impractical solution. The second still rather direct method is to measure the HRTF of an individual for a collection of spatial positions and to then use this individual HRTF to produce an individual binaural synthesis

rendering through convolution of the sound source with the relevant incident direction HRTF [14, 105]. While this is the most common method employed to date, it is generally limited to those with the facilities and equipment to carry out such measurements [4].

The general pros and cons between binaural recordings and binaural synthesis merit mention. While individual binaural recordings provide arguably the most accurate 3D audio capture/reproduction method, they require the sonic environment and the individual to be situated accordingly. For any reasonable production, this would resemble a theatrical piece being performed around the individual in a first person context. The recording would capture the acoustic detail of the soundscape, including reflections from various surfaces, diffraction and scattering effects. However, the head orientation of the individual would be encoded into the recording, imposed on the listener at playback. If presented to another individual, the issues of HRTF mismatch are introduced, degrading the spatial audio quality to an unknown degree for each individual. In laboratory conditions, this method suffers additional difficulty, as the individual takes part in the recording, making the presentation of unfamiliar material difficult. In contrast, binaural synthesis allows for the scripting, manipulation and mixing of 3D scenarios without the intended listener present. With real-time synthesis, head tracking can be incorporated allowing freedom of movement by the individual, a basic requirement for VR applications. HRTF mismatch is alleviated through the use of individual HRTFs. However, the quality of the production is affected by the level of detail in the acoustic simulation of the environment, including elements such as source and surface properties. Highly complex scenes and acoustic environments can require significant computational resources (the interested reader can refer to Chap. 3 for further details on this topic). Spatial synthesis using HRTF data is also affected by the measurement conditions of the employed HRTF, predominantly the measurement distance. If sound sources are to be rendered at various distances, this requires either multiple HRTF datasets, or deformation of the individual HRTF data to approximate such changes in distance. Further discussion of these details is beyond the scope of this chapter. In continuing, the focus will be limited to questions concerning the individual nature of the HRTF as integrated into an auditory VR environment through binaural synthesis.

Introduction to System-to-User and User-to-System adaptation

A variety of alternative methods exist in order to improve the match between the HRTF used for the rendering and the specific HRTF of the listener. It is the aim of this chapter to present an overview of those approaches that have been evaluated and validated through experimental research. In order to map the various methods and at the same time simplify the narrative and facilitate the reading, the text has been organised in two separate sections. Section 4.2 presents research which looks at matching the rendering system to the specific listener (system-to-user adaptation), thus aiming to provide every individual with the best HRTF possible. Section 4.3 looks at the problem from a diametrically opposite point of view, introducing studies where the listener is trained in order to adapt to the rendering system (user-to-system

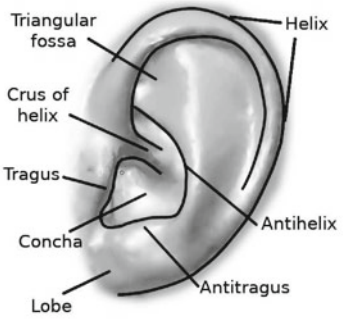
adaptation), therefore aiming at improving the performance of a specific individual when using non-individual HRTFs.

While a rather extensive number of studies exist on the topic of system-to-user adaptation, a more limited amount of research has been carried out focusing on user-to-system adaptation. For this reason, while Sect. 4.2 is presented as an extensive review of several research projects, Sect. 4.3, after an initial overview, then dives more in depth into one specific study carried out by this chapter's authors, giving details of the methodology and briefly discussing the results. Section 4.5 concludes by presenting a brief overview of open challenges on this topic.

4.2 System-to-User Adaptation: HRTF Synthesis and Selection

Two main approaches exist for obtaining individual (or at least personalised) HRTFs without having to measure them acoustically. The first one focuses on numerical simulations, therefore using mathematical methods to generate an HRTF for a given individual from 3D models of the head, torso, and pinnae. Techniques such as the Boundary Element Method (BEM), Finite Element Method (FEM), and Finite Difference Time Domain (FDTD) method which are commonly employed in diffraction, scattering, and resonance problems allow one to calculate the HRTF of a given individual based on precise geometrical data (e.g. coming from a 3D scan of the head and pinnae), which have been used for this purpose since the late 1990s, and have shown increased uptake and success in the past years thanks to technological advancements in domains such as high-performance computing and high-resolution 3D scanning. An example of such a resulting 3D mesh from a Neumann KU-100 dummy head can be seen in Fig. 4.2b. The second one relies on using HRTFs from available datasets, either transforming them in order to provide a better fit for a given listener or selecting a best fit considering, for example, preference or performance, e.g. using a sound localisation task or signal metric. Due to the relative independence between the ITD and the Spectral Cues, the HRTF can be decomposed and different elements addressed by different methods, e.g. an ITD structural model can be used with best fit selected Spectra Cues [22, 78].

As can be expected, each of these approaches comes with specific challenges. Moreover, the success in employing one or the other depends significantly on factors such as the available data (quantity and quality), the time constraints in order to run the tests and the calculations, and the context for which the rendering is needed (i.e. the requirements in terms of quality, interactivity, etc.). An overview of the various techniques and related challenges, including solutions found through state-of-the art research studies, is presented in the following sections.



(a) Diagram of the pinna with anthropometric labels.

(b) Mesh of Neumann KU-100 dummy head fitted with a parametric pinna model.

Fig. 4.2 Pinna morphology nomenclature and example BEM mesh (from [91])

4.2.1 HRTF Modelling

Various attempts have been made to investigate the function of the pinna, linking HRTFs to its morphology as well as that of the head and torso. Early work by Teranishi and Shaw [93] looked at creating a physical model of the pinnae and analysing the various excitation modes generated by a nearby point source. The model, based on very simple geometries, showed responses similar to those of real data, and represented one of the first steps towards better understanding the spatially varying acoustic role of the pinna. Similar work was done by Batteau [12], who created a mathematical representation of the acoustical transformation performed by the pinna and produced the first mathematically described theory of sound source localisation based on a reflection-diffraction model. These studies were the baseline of research carried out 30 and more years later, when the available computational power allowed to create more complex models, and to validate those by comparing them with experimental measures (e.g. [58]). Further modelling work was carried out looking at simplified models and approximations. Notable examples are those of Genuit [26] based on a structural simplification model of the pinnae, Algazi and colleagues [1] based on an approximation of the head and the torso using ellipsoidal and spherical models, and Spagnol and colleagues [89] looking at ray-tracing analysis of pinna reflection patterns. It is relevant to note that many of the early studies focused on models for understanding the various phenomena and principles involved, rather than models for binaural audio rendering. For these early studies, much of the research on spatial perception was carried out independently from acoustical/morphological studies regarding the details of the pinnae.

Structural Modelling

One of the first experiments using these techniques applied to HRTFs (including pinnae) was carried out by Katz [49, 51, 52]. This work focused on using BEM to calculate HRTFs by modifying various aspects of the geometrical models, for example, eliminating the pinna, changing the size and shape of the head, and accounting for hair acoustic impedance. Results from numerical simulations were then compared with experimental measures, validating the technique and improving our understanding of the role of the pinnae in modifying the incoming sound in a direction-dependent manner. Similar work was carried out in the same period by Kahana [44, 46]. Such simulations were initially limited, due to computational resources, to an upper frequency of 6 kHz, then extended to 10 and 20 kHz in later studies [32, 45]. Even in these cases the validation was performed comparing the numerical model results with experimental measurements showing a good match between the two, also in light of the variances observed between different HRTF measurement systems for the same individual [4, 47]. The computational complexity of these numerical methods was a major limitation in the early years of using this technique for generating HRTFs. Various optimisation techniques are being proposed [35, 55, 70], allowing significantly faster computation times with reasonable processing resources (i.e. no longer needing super computers). This led to the development of easy-to-use and open-source tools for the numerical calculation of HRTFs. A notable example is `mesh2hrtf` [110], a software package centred on a BEM solver, as well as tools for the pre-processing of geometry data, generation of evaluation grids and post-processing of calculation results. It is essential here to consider a major challenge to be tackled when approaching HRTF synthesis from geometrical models, which is the acquisition and processing of the 3D models from which the HRTFs are computed. Evaluations of various 3D scanning methods, specifically looking at capturing the geometry of the pinnae, have been carried out [44, 69, 80].

Numerical simulations also brought significant benefits with regard to repeatability, replicability and reproducibility. A comparison of different numerical tools for simulating an HRTF from scan data by Greff and Katz [32] (here employing the high-resolution scan of a Neumann KU-100 shown in Fig. 4.2b) showed little variance. In contrast, a similar comparison of acoustical HRTF measurements using the same head at different laboratories [4] showed significant variations between resulting HRTFs. Another significant advantage of numerically modelling HRTFs rather than measuring them is that with physical measurements on human subjects it is difficult or impossible to isolate the influence of different morphological characteristics on the actual HRTF filters.

Morphological Relationships

Exploring and modelling the relationship between geometrical features and filter characteristics is indeed a very important step for advancing our understanding of the spatial hearing processes. Research in this area was strongly advanced with the distribution of the CIPIC HRTF database [2], which included associated morpho-

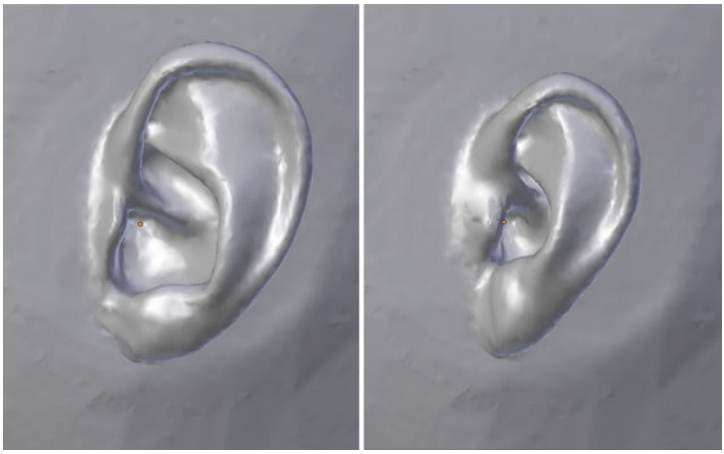


Fig. 4.3 Two pinna created with the parametric model developed in [91]

logical parameter data for most subjects. This effort was followed with the LISTEN HRTF database [98], providing similar data. Benefiting from the power of numerical simulation and controlled geometrical models, Katz and Stitt [91] investigated the effect of morphological changes by varying specific morphological parameters, an extension of the CIPIC set of morphological parameters to provide more unique solutions. In order to do this, they created a Parametric Pinna Model (PPM) and with BEM they investigated the sensitivity of the HRTF to specific morphological alterations. Examples of pinnae created using this PPM can be seen in Fig. 4.3. Evaluations included the use of auditory models [88] to identify those morphological changes most likely to affect spatial hearing perception. In line with previous studies, morphological features near to the rear of the helix were found to have little influence on HRTF objective metrics, while the dimension of the concha had a much more relevant impact, both looking at the directional and diffuse HRTF spectral components. [2] Other relevant findings include the importance of the region around the triangular fossa, which is often not considered when looking at HRTF personalisation, and the fact that the relief (or depth, directions parallel to the interaural axis) parameters were found to be at least as important as side-facing parameters, which are more frequently cited in morphological/HRTF studies.

Such interest in binaural audio, combined with major advancements in terms of available technologies, has encouraged the publication of large datasets of BEM-generated HRTFs and correspondent high-accuracy 3D geometrical models. An example is the Sydney York Morphological and Acoustic Recordings of Ears (SYMARE) database [42], which was then followed by other examples of either head-related or more reduced complexity pinnae-related datasets [18, 34]. The availability

[2] The diffuse field component is the spatial average of the HRTF. When removed from the HRTF, the result is a diffuse field equalised *directional transfer function* (DTF) [64].

of such large datasets opened the door to the use of machine learning approaches to tackle the issue of morphology-based HRTF personalisation. An example is the work by Grijalva and colleagues [33], where a non-linear dimensionality reduction technique is used to decompose and reconstruct the HRTF for individualisation, focusing on elements which vary the most between positions and across individuals. Results may offer improved performance over linear methods, such as principal component analysis (e.g. [81]).

HRTFs, Binaural Models and Perceptual Evaluations

It is evident that since the 1990s a large amount of work has been carried out looking at synthesising HRTFs and better understanding the relationship between these and morphological features of the pinnae, head and torso. Nevertheless, it must be reiterated that very few of the reviewed studies have included perceptual evaluations on the modelled HRTFs [18, 56], and that in no case such subject-based validations were extensive enough to fully support the use of synthesised HRTFs instead of measured ones. It is therefore clear that significant research is still needed in order to develop and validate models that can describe, classify and ultimately generate individual HRTFs from a reduced set of parameters.

While numerical assessments can be very useful when trying to better explain experimental results, they cannot be the only way to explore and validate the quality of the rendering choices. Binaural models (e.g. [88]) could become an invaluable tool to help overcome such limitations, as they offer a computational simulation of binaural auditory processing and, in certain cases, also allow to predict listeners' responses to binaural signals. Using them, it is possible to rapidly perform comprehensive evaluations that would be too time-consuming to implement as actual auditory experiments (e.g. [17]).

An example of this approach can be found in [29], where an anthropometry-based mismatch function between HRTF pairs, looking at the relationship between pinna geometry and localisation cues, was used to select an optimal HRTF for a given individual, specifically looking at vertical localisation. The outcome of the selection was then evaluated using an auditory model which computed a mapping between HRTF spectra and perceived spatial locations. While this study outlined that the best fitting HRTF selected with the proposed method was predicted to yield a significantly improved vertical localisation when compared to a selected generic HRTF, it must be reiterated that the reliability of perceptual models is still to be thoroughly validated, and potential biases can be identified and dealt with only through actual perceptual evaluations. Another similar application of binaural models has been recently published, focusing on the comparison between different Ambisonics-based binaural rendering methods [25]. The very large number of independent variables (e.g. each method was tested with Ambisonics orders from 1 to 44), as well as the complexity of the interactions between such variables, would make it very challenging to run perceptual evaluations with subjects. This study showed not only that models' predictions were consistent with previous perceptual data, but also contributed to validate the models' ability to predict user responses to binaural signals.

It is likely that models will never be able to provide 100% accurate assessments near to the zone of perfect reproduction, in part due to the difficulties in modelling processes such as cognitive loading and procedural/perceptual learning. However, it is reasonable to expect them to provide broadly correct predictions for larger errors. This means that they could be particularly useful when prototyping rendering algorithms and designing HRTF personalisation experiments, in order to rapidly reduce the number of conditions and variables which are subsequently assessed through real subject-based perceptual evaluations.

Artificial intelligence and machine learning should play an important role in such future research, looking at improving both HRTF synthesis and selection processes, as well as perceptual models accuracy and reliability.

4.2.2 HRTF Selection

A different approach for obtaining individual (or at least personalised) HRTFs without having to acoustically measure them is to rely on available HRTF databases, either transforming/tuning the transfer function according to certain subjective criteria, or designing a process for selecting the best fitting HRTF for a given subject. Regarding the first option, as mentioned at the beginning of this section, it is generally known that frequency-independent ITDs from a given HRTF can be modified and personalised according to e.g. the head circumference of a given listener [9]. Such a technique is implemented in a few binaural spatialisers [22, 78]. However, the personalisation of other HRTF features, such as monoaural and interaural Spectral Cues, presents more significant challenges. Early works in this direction looked at improving vertical localisation by scaling the HRTF in frequency [64, 65]. Other "simpler" approaches to tuning were found to be effective, for example, by manually modifying frequency and phase for every HRTF direction, for the left and right ears independently [86]. Hwang and colleagues [40] carried out a principal component analysis on the CIPIC HRTFs and used the output components to develop a customisation method based on subjective tuning of a generalised HRTF. Such customisation allowed listeners to perform significantly better in vertical perception and front-back discrimination tasks. The same approach was used to modify and personalise a KEMAR HRTF, resulting also in this case in significantly improved vertical localisation abilities [84].

HRTF Selection Methods

Methods for selecting a best fit HRTF based on subjective criteria can be grouped into two general categories: physical measurement-based matching and perceptual selection. The first pertains to selecting an HRTF from an existing set based on morphological measurements or sparse acoustical measurements. Of importance is the determination of the relevant morphological features, as they pertain to spatial hearing and HRTF-related cues, as examined by [91]. Zotkin and colleagues [112] looked at a selection strategy based on matching certain anthrophometric pinnae parameters of

the specific subject with those of HRTFs within a dataset, while providing associated low-frequency information using a "head-and-torso" model. Comparison between a non-personalised HRTF and the selected HRTF via this method showed heightened localisation accuracy and improved subjective perception of the virtual auditory scene when using the latter. A similar approach was used by [81], where advanced statistical methods were employed to create a subset of morphological parameters, which were then employed for predicting what might be the subject's preferred HRTF based on measurement matching. HRTFs selected using this method performed better than randomly selected ones. An alternate selection perspective was proposed in [30], where a reflection model was applied to the picture of the pinnae of the subject, facilitating the extraction of relevant anthropometric parameters which were then used for selecting one or more HRTFs from an existing database. This selection method resulted in a significant improvement in elevation localisation performances, as well as an enhancement of the perceived externalisation of the simulated sources. The relationship between features of the pinna shape and HRTF notches, focusing specifically on elevation perception, was successfully used in [27] for selecting a best fitting HRTF from pinna images. Interestingly, studies on Spectral Cues have suggested the importance of notches over peaks in the HRTF [31]. Another work from Geronazzo and colleagues [28] introduced a rather original approach by developing the *Mixed Structural Modelling* (MSM), a framework for HRTF individualisation which combines structural modelling and HRTF selection. The level of flexibility of this solution, which allows to mix modelled and recorded components (therefore HRTF selection and synthesis), is particularly promising when looking at the HRTF personalisation process.

HRTF Evaluation

It must be highlighted that whether selection is based on measured or perceptual data, the evaluation of said method is necessarily perceptual as the final application is a human-centred experience. With this in mind, a fundamental yet unanswered question is: "What determines the suitability of an HRTF for a given subject?" [48]. When establishing whether an HRTF is a good fit, should one look at how precisely sound sources can be localised using that HRTF (direct approaches), or should other subjective metrics (e.g. realism, spatial quality or overall preference) be employed [87]? In employing perceptual selection, the choice of protocol becomes more critical. In addition, as was observed with acoustical measurements, the repeatability of the measurement apparatus (here the response of human subjects) must be examined and taken into account. As an example, past studies using binaural audio rendering for applications other than spatial hearing research (e.g. [74]) relied on simple perceptually based HRTF selection procedures which, at a later stage, resulted in being less repeatable than originally thought [6]. Without extensive training as seen in some of the principal earlier studies, the reliability of naive listeners (those situations which are also more representative of applied uses of binaural audio rather than studies on fundamental auditory processing) must be taken into account. Early studies on HRTF selection through ratings [53, 74] assumed innate reliability in quality judgements.

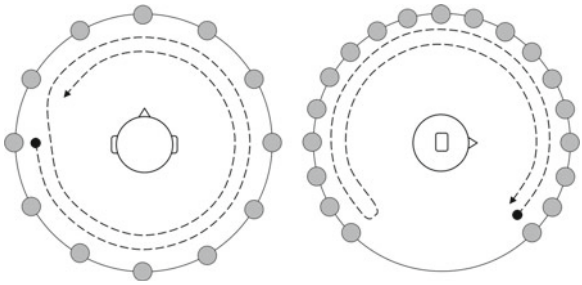


Fig. 4.4 Trajectory graphic description reference for HRTF quality ratings: horizontal (left) and median (right) plane trajectories indicating the start/stop position and trajectory direction (● - - →) (from [92])

More recently, studies have shown that such reliability cannot be assumed, but must be evaluated, with some listeners being highly repeatable while others are not [6].

It can be assumed that different HRTFs will, for a given subject, result in different performances in a sound source localisation task. From this we can infer that an optimal HRTF could be selected looking at such performances, for example, using metrics such as localisation errors and front-back and up-down confusion rates (see Sect. 4.3.2 for metric definitions). This assumption has been the baseline of several studies where an HRTF selection procedure was designed and evaluated based on localisation performances [41, 83, 96]. Such methods previously required specialised hardware, though current consumer Virtual Reality (VR) devices, thanks to their increasingly higher performance in terms of tracking capabilities (e.g. [43]), can now be employed for rendering and reporting the perceived direction of the sound source. However, these methods still remain rather time-consuming, as a large number of positions across the whole sphere should be evaluated in order to obtain reliable results.

Alternatively, HRTF selection can be the result of subjective evaluations based on indirect quality judgement approaches. Several research works have looked at asking listeners to rate HRTFs based on the perceived quality of some descriptive attributes, from the overall impression [106] to how well the auditory presentation matched specifically described locations or movements of the virtual source [53, 83, 85] (e.g. Fig. 4.4). Several methods have been introduced for ultimately being able to select one or more best performing HRTFs; these include ranking [83], rating on scales [6, 53, 82], multiple selection-elimination rounds [97] and pairwise comparisons [85, 106]. In general, there seems to be an agreement on the fact that expert assessors (as defined by [107]) perform significantly better (i.e. in a more reliable and repeatable manner) if compared with initiated assessors [6, 54]. To gain further insight into indirect method results, some work has been carried out to develop global perceptual distance metrics with the aim to describe both HRTF and listener similarities [8]. In addition to proposing and evaluating a set of perceptual metrics, this work encourages further research into novel experiment design which could help in minimising the need for data normalisation and, more importantly, outlines the need for further investigations on the stability of these perceptual experiments/evaluations, specifically looking at repeatability and training.

Methods Comparison

Few studies have examined the similarity between direct (i.e. localisation performances) and indirect HRTF selection methods. Using an immersive VR reporting system for the localisation test, results from [108] indicated a significant and positive mean correlation between HRTF selection based on localisation performance and HRTF ranking/selection based on quality judgement; the best HRTF selected according to one method had significantly better rating according to metrics in the other method. In contrast, using a gestalt reporting method through the use of an avatar representation of the listener's head, results from [54] showed no significant correlations. A number of protocol differences exist between these two studies, including the type of tasks used for both methods, the user interface (see [10, 11] regarding localisation reporting method effects), the stimuli signals, as well as the metrics evaluated in the quality judgement task.

4.3 User-to-System Adaptation: HRTF Accommodation

The previous section examined HRTF selection and individualisation methods in the signal domain. While such methods aim to provide every individual user with the best HRTF possible, such approaches are not always available in all conditions. However, evidence is increasingly available showing that the adult brain is adaptable to environmental changes. It has been demonstrated that this adaptability (or plasticity) regarding spatial auditory processing can lead to a reduction in localisation error over time in the case when a listener's normal localisation cues are significantly modified.

It has been established that one can adapt to modified HRTFs over time, with ear moulds inserted in the pinnae [19, 38, 94, 95], or with non-individual HRTFs through binaural rendering [73, 77, 90, 92, 99, 109]. Studies have shown that one can adapt to distorted HRTFs, e.g. in [60] where participants suffering from hearing loss learned to use HRTFs whose spectrum had been warped to move audio cues back into frequency bands they could perceive. HRTF learning is not only possible, but lasting in time [62, 92, 109]: users have been shown to retain performance improvements up to 4 months after training [109]. Given enough time, participants using non-individual HRTFs may achieve localisation performance on par with participants using their own individual HRTFs [73, 77, 92].

This concept has been successfully used to improve user localisation performance within virtual auditory environments when using non-individual HRTFs. Readers are referred to [61, 104] for more general reviews on the broader topic of HRTF learning.

4.3.1 Training Protocol Parameters

Learning methods explored in previous studies are often based on a localisation task. This type of learning is referred to as *explicit* learning [61], as opposed to *implicit* learning where the training task does not immediately focus participant attention on localisation cues [73, 92]. Performance-wise, there is no evidence to suggest either type is better than the other. Implicit learning gives more leeway for task design *gamification*. The technique is more and more applied to the design of HRTF learning methods [39, 73, 90, 92], and while its impact on HRTF learning rates remains uncertain [90], its benefit for learning, in general, is, however, well established [36]. On the other hand, explicit learning more readily produces training protocols where participants are *consciously* focusing on the learning process [63], potentially helping with the unconscious re-adjustment of auditory spatial mapping.

As much as the nature of the task, providing feedback can play an important role during learning. VR technologies are more and more relied upon to increase feedback density in the hope of increasing HRTF learning rates (in Chap. 10, the interested reader can find further insights on multisensory feedback in VR). While results encourage the use of a visual virtual environment [60], it has been reported that proprioceptive feedback alone can be used to improve learning rates [16, 73]. Direct comparison of experimental results suggests that active learning with direct feedback is more efficient (i.e. leads to faster improvement) than passive learning from sound exposure [61]. There is also a growing consensus on the use of adaptive (i.e. head-tracked) binaural rendering during training to improve learning rates [19], despite the generalised use of static head-locked localisation tasks to assess performance evolution [61]. It is not trivial to ascertain whether the benefit of head-tracked rendering comes from continuous situated feedback improving audio cue recalibration, or from unbalanced comparison, as static head-locked rendering creates user frustration and results in less sound exposure [90]).

Studies on the training stimulus indicate that learning extends to more than the signals used during learning [39, 90]. This result is likely dependent on specific characteristics of the stimuli and how these relate to auditory localisation mechanisms, i.e. whether they present the transient energy and broad frequency content necessary for auditory spatial discrimination [24, 57, 72].

There is no clear cut result on optimum training session duration and scheduling. Training session duration reported in previous studies ranges from ≈8 min [66] to ≈2h [60]. Comparative analysis argues in favour of several short training sessions over long ones [61]. Training session spread is also widely distributed in the literature, ranging from all sessions in one day [57] versus one every week or every other week [92]. Where results suggest spreading training over time benefits learning (all in 1 day versus spread over 7 days) [57] outcomes from [73, 92] indicate that weekly sessions and daily sessions result in the same overall performance improvement (for equal total training duration). There is some example of latent learning (improvement between sessions) in the literature [66], naturally encouraging the spread of training sessions. Regardless of duration and spread, studies have shown that learning sat-

uration occurs after a while. In [59], most of the training effect took place within the first 400 trials (≈160 min), a result comparable to that reported by [20] where saturation was reached after 252 to 324 trials.

One of the critical questions not fully answered to date is the role of the HRTF fit in the training process or how similar the training HRTF is to the actual HRTF of the individual. It would appear that a certain degree of affinity between a participant and the training HRTF facilitates learning [73, 92]. In contrast, lack of adaptation can occur if the HRTF to be learned is too different from one's own HRTF. This is evidenced by mixed adaptation results in studies where ill-suited HRTF matches were tested.

4.3.2 HRTF Accommodation Example

We present here as an example HRTF learning study by Stitt et al. [92], which examined the effect of adaptation to non-individual HRTFs. This study was chosen for this example as it provides a controlled study over a significant number of training sessions. As a "*worst-case*" real-world scenario, perceptually worst-rated non-individual HRTFs were chosen by each subject to allow for maximum potential for improvement, another factor of interest in its design. This study is part of a series of studies on the subject of user-to-system adaptation, providing continuity of comparisons [15, 73, 77]. The methodology consisted of a training game and a localisation test to evaluate performance carried out over 10 sessions. Subjects using non-individual HRTFs (group **W10**) were tested alongside control subjects using their own individual measured HRTFs (group **C10**).

Prior to any training, subjects were assigned non-individual HRTFs based on quality judgements of rendered sound object trajectories for 7 HRTF sets, taken as "*perceptually orthogonal*" [53]. These trajectories, shown in Fig. 4.4, were presented to subjects as a reference. Following the results of [8], which examined the reliability and repeatability of HRTF judgements by naive and experienced subjects, this rating task was performed three times, leading to a total of six ratings per subject, counting the two trajectories, with the overall judgement rating taken as the overall mean. The lowest rated HRTF for each subject was then used as that subject's *worst*-match HRTF. This method is an improvement over alternate methods which are either uncontrolled (e.g. a single HRTF used by all listeners) or limited in the extent of relative spectral changes presented to subjects when compared to their individual HRTFs.

The training procedure for the 10 sessions was devised as a simple game with a searching task in which the listener had to find a target at a hidden position in some direction (θ, ϕ), ignoring radial distance. Subjects searched for the hidden target by moving the motion-tracked hand-held object around their head (see concept in Fig. 4.5). For the duration of the search, alternating pink/white noise (50–20000 Hz) with an overall level of approximately 55 dBA measured at the ear was presented to the listener, positioned at the location of the tracked hand-held object relative to the

Fig. 4.5 Training game concept design

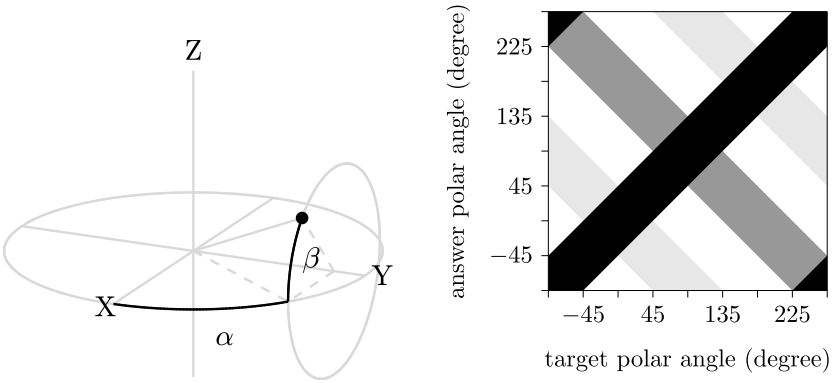
subject's head. This provided a link between the proprioceptively known position of the subject's own hand and spatial cues in the binaural rendering. The alternation *rate* of the pink/white noise bursts increased with increasing angular proximity to the target direction using a Geiger counter metaphor [71, 79]. Once the subject reached the intended target direction, a success sound would play, spatialised at the target's location. The training game lasted 12 min and subjects were instructed to find as many targets as possible in the time available. Sessions 1–4 occurred at 1-week interval, while the remaining sessions occurred at 2-week interval.

It should be emphasised that no auditory localisation on the part of the subject was actually required to accomplish this task, only tempo judgements of the alternation *rate* of the pink/white noise bursts and proprioceptive knowledge of one's hand position. HRTF adaptation was therefore an implicit result of game play, but not the task of the game as far as the participant was aware. This task was designed to facilitate learning with source positions outside of the visual field of view, as well as to function for individuals with visual impairments.

Performance Evaluation Metrics

The HRTF accommodation was evaluated via localisation tests. Subjects were presented a brief burst of noise (to limit the influence of any possible head movement during playback) and would subsequently point in the perceived direction of the sound using the hand-held object. No feedback was given to subjects regarding the target position. The noise burst consisted of a train of three, 40ms Gaussian broadband noise pulses (20000 Hz) with 2 ms raised cosine window applied at onset and offset and 30ms of silence between each burst [73]. There were 25 target directions with 5 repetitions of each target, resulting in the tested sphere including a full 360° of azimuth, and –40–90° of elevation.

Two types of metrics were used to analyse localisation errors: *angular* and *confusion* metrics. The interaural coordinate system defines a lateral and polar angle Fig. 4.6a. The lateral angle is the angle between the interaural axis and the line between the



(a) Interaural coordinate systems. Interaural lateral angle α defined in [-90:90], polar angle β in [-90:270]. Lateral angle is shifted by 90° compared to original definition [67]. Listeners facing X with their left ear pointing towards Y.

(b) Definition of the 4 different cone-of-confusion response classification zones: *Precision*, *Front-Back*, *Up-Down*, *Combined*(from [108]).

Fig. 4.6 Interaural polar coordinate system and associated polar angle cone-of-confusion zone definitions

origin and the source. The lateral angle approaches cones-of-confusion along which the interaural cues (ITD and ILD) are approximately equal. A cone-of-confusion is defined by the contour around the listener for a given ITD or ILD (see Fig. 4.1). For ITD, these contours can be generally represented by a hyperbolic function, where the difference in arrival time to the two ears is constant and the vertex is on the interaural axis, between the two ears. The intersection of the ITD and ILD cones-of-confusion for a given stimulus prescribes a closed curve (approaching a circle). The ITD and ILD are insufficient to resolve the localisation ambiguity, requiring further information, such as from Spectral Cues or head movements. The polar angle is the angle between the horizontal plane and a perpendicular line from the interaural axis to the point, such that the polar angle prescribes the source location on the cone-of-confusion. The polar angle is primarily linked with the monaural, Spectral Cues in the HRTF. This independence of binaural and Spectral Cues makes the interaural coordinate system a natural choice when looking at localisation performance. If the perceived ILD, ITD and Spectral Cues of a given source do not adequately coincide with the expectations of the auditory system for a single point in space, uncertainty in localisation response ensues. The most commonly referenced uncertainties are polar angle confusions.

Polar angle confusions are classified using a traditional segmentation of the cone-of-confusion [73, 92], revised in [108]. The classification results in three potential confusion types, front-back, up-down and combined, with a fourth type corresponding to precision errors, represented schematically in Fig. 4.6b. The *precision* category designates any response close enough to the real target so as not to be associated to

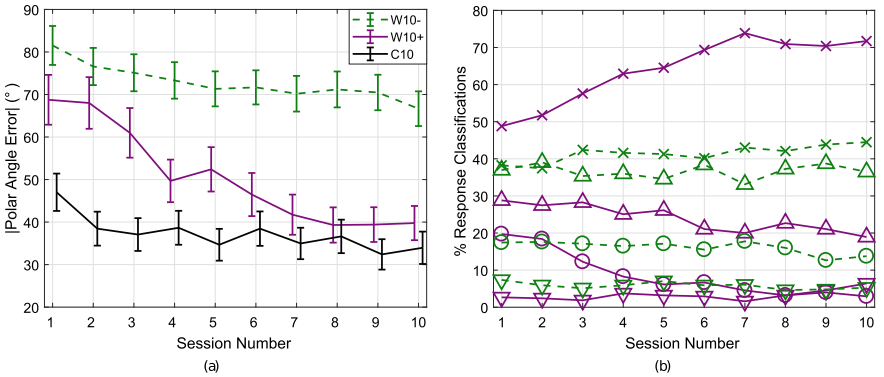


Fig. 4.7 Result analysis by subgroup. **a** Mean absolute polar angle error and 95% confidence intervals for groups **W10+**, **W10–** and **C10** across sessions 1–10. **b** Response classification analysis: Mean classification of results for group **W10** by type (*precision* (×), *front-back* error (◯), *up-down* error (▽) and *combined* error (△)) for groups **W10+** (—, 3 subjects) and **W10–** (**- -**, 5 subjects) over sessions 1–10 (from [92])

the other confusion types. In short, responses classified under *precision* are for those within ±45° of the target angle, *front-back* classified errors are responses reflected in the frontal plane, and those classified *up-down* are for those reflected in the transverse plane. Any responses that fall outside of these regions are classified as *combined* type errors.

Performance Evaluation Results

Results examined the evolution of polar angle error and confusion rates. As a measure of accommodation, the *rate of improvement* was defined as the gradient of the linear regression of polar angle error. The rates of improvement for the 8 subjects spanned values of 0.5° to 4.6°/session over sessions 5–10 (as results for initial sessions have been shown to be influenced by procedural learning effects [59]). In contrast, results for the control group over the same sessions spanned 0° to 2.2°/session. A clustering analysis of the test group relative to the control group, **C10**, separated those whose rate of improvement exceeded that of the control group (subgroup **W10+**) and the remaining subjects (**W10–**) who did not. This second group failed to exhibit clear HRTF adaptation results over and above that of the control group whose improvement can be considered primarily as procedural learning.

The polar errors are shown in Fig. 4.7a for groups **W10+**, **W10–** and **C10**. Group **W10+** approached a similar level of absolute performance to **C10**. This demonstrates that these subjects were able to adapt well to their *worst*-rated HRTF to a level approaching subjects using their individually measured one. It also shows clearly that, despite continuous training, some subjects (**W10–**) exhibited little or no improvement beyond the procedural learning seen in **C10**.

The response classification results for groups **W10+** and **W10–** are shown in Fig. 4.7b. At the outset of the study, it can be observed that *up-down* and *front-back* type error rates are comparable between the two subgroups, with **W10–** exhibiting more *combined* type errors. This metric could be a potential indicator for identifying poor HRTF adaptation conditions. Subsequently, it can be clearly seen that group **W10+** exhibits a steady increase in *precision* classified responses, with reductions in *front-back* errors over sessions 3–5 and subsequent reductions in *combined* errors. In contrast, group **W10–** exhibits generally consistent response classifications across sessions, with only small increases in *precision* classification mirrored by a decreasing trend in *front-back* errors. For all subjects, it can be noted that the occurrence of *up-down* errors is quite rare.

Results of this accommodation study show that adaptation to an individual's perceptually worst-rated HRTF can continue as long as training is provided, though the rate of improvement decreases after a certain amount of training. A subgroup achieving localisation performance levels approaching the control group with individual HRTFs. These performance levels were comparable to those observed in [73] with identical test protocol, where subjects performed only three training sessions using their perceptually *best rated* HRTF.

4.4 Discussion

It is clear that, while various methods and tools are available for selecting a best fit HRTF for a given listener, there is no established evaluation protocol to determine how well these methods work and compare with each other. While some work is advancing in proposing common methodologies and metrics [75], the lack of established methods raises some very relevant questions about the feasibility of a unique HRTF selection task which performs reliably and independently from factors such as the listeners expertise, the signals employed, the user interface, the context where the tests are carried out and, more in general, the task for which the final quality is judged. It seems evident that any major leap forward in this field is limited until two primary issues are addressed: (1) the establishment of pertinent metrics to perceptually assess HRTFs and (2) the relationship between these metrics and specific characteristics of the signal domain HRTF filters.

The use of HRTF adaptation, in examining the results of this and previous studies, has been shown to be a viable option to improve spatial audio rendering, at least with regard to localisation. The level of adaptation achievable is related to the initial suitability (perceptual similarity) between the system HRTF and the user's individual HRTF, with more suitable HRTFs showing more rapid adaptation. No significant effect has been found regarding the specific training intervals, though spreading out sessions is better than multiple sessions on the same day. The adaptation method could be integrated into a stand-alone game application, or as part of device setup and personalization configurations, typical of most VR devices to some degree. The major limitation, once the training HRTF is chosen, is the need for repeated training

Fig. 4.8 Example active HRTF learning training game. Training setup: (top-left) participant in the experiment room, (bottom-left) third person view of the training platform, (right) participant viewpoint during the training (from [77])

sessions, and this must be made clear to users so that they do not expect ideal results from the start.

The combination of user-to-system and system-to-user adaptation is a promising solution. While user-to-system adaptation appears limited by the initial training HRTF employed, system-to-user adaptation methods provide various means of providing, if not a perfect individual HRTF, a reasonable near approximation. As such, selection of a *pretty-good* HRTF match followed by user training could be a viable real-world solution.

An example of such a tailored HRTF training has been tested in [77]. In this work, as compared to the previous mentioned study in Sect. 4.3.2, the subject was aware of the goal of the training, with specific HRTF-based localisation difficulties presented with increasing difficulty (see Fig. 4.8). In addition, a best match HRTF condition was employed using an interactive exploration method, rather than the general ranking described in Sect. 4.2.2 and a worst-case selection scenario. Results indicated that the proposed training program led to improved learning rates compared to that of previous studies. A further addition of this study was the inclusion of a simulated room acoustic response, moving from the typical anechoic conditions of previous studies to a more natural acoustic for the user. Results showed that the addition of the room acoustics improved HRTF adaptation rate across sessions.

4.5 Conclusions and Future Directions

While binaural audio and spatial hearing have been studied for over 100 years, major advancements in these fields have occurred in the last two to three decades, possibly thanks to progress in real-time computing technologies. It has been extensively shown that everyone perceives spatial sound differently thanks to the particular shape of their ears, head and torso. For this reason, either high-quality simulations need to

be uniquely tailored to each individual listener, or the listener needs to adapt to the configuration (i.e. the HRTF on offer) of the rendering system, or again some combination of both using individualised HRTFs. This chapter has provided an overview of research aimed at systematically exploring, assessing and validating various aspects of these two approaches. But while there is a good level of agreement on certain notions and principles, e.g. that using non-individual HRTFs can result in impaired localisation performance which can however be improved through perceptual training, there are still open challenges in need of further investigation.

A rather general but very important question that has yet to be addressed is how we can measure whether a simulated immersive audio experience is suitable and of sufficient quality for a given individual. Previous work has established a certain level of standardisation for assessing general audio quality (e.g. related to telecommunication and audio compression algorithms), but equivalent work has yet to be carried out in the field of immersive audio. Objective and subjective metrics for assessing HRTF similarity have been explored and evaluated in the past [5], and recently published research suggests that additional metrics might exist, e.g. looking at speech understanding performance [21] or machine learning artificial localization tests [3, 13]. Nevertheless, extensive research is still needed in order to understand and model low-level psychophysical (sensory) as well as high-level psychological (cognitive) spatial hearing perception.

Factors other than choices related to binaural audio processing could also have an impact on the overall perception of the rendered scenes. The fact that high-quality, albeit non-interactive, immersive audio rendering can be achieved through recordings done with a simple binaural microphone, which by definition do not account for individualised HRTFs, can be considered an example of the major complexity and dimensionality of the problem. Matters such as the choice of audio content, the context of the rendered scene, as well as the experience of the listener (e.g. whether they have previously participated in immersive audio assessments) have been shown to be relevant when assessing the perceived quality of the immersive audio rendering [6, 54]. Such a discussion found a natural continuation in Chap. 5.

Looking more in depth at the need to quantify the individually perceived quality of the rendering, the understanding of the perceptual weighting of morphological factors contributing to spatial hearing becomes an essential target to be achieved. Data-based machine learning approaches may be a useful tool when tackling this, as well as challenges related to user-to-system adaptation. Examples include allowing a certain level of customisation of the training by individually and adaptively varying the difficulty of the challenge, maximising learning and at the same time avoiding an overload of sensory and cognitive capabilities. Further explorations on spatial hearing adaptation shall focus on exploring the transferability of the acquired training between different hearing skills (e.g. [100]) and examining to what extent spatial auditory training performed in VR is transferable to real-life tasks.

Another very relevant yet still under-explored area of research is employing cognitive and psycho-physiological measurements when trying to assess both the quality of rendered spatial hearing cues and the cognitive effort during HRTF training. In the first case, measures related with behavioural performance, as well as electroen-

cephalographic markers of selective attention, could be used to assess the suitability of immersive rendering choices [23], possibly opening the path towards passive perceptual-based HRTF selection. In the second case, similar metrics, with the addition of other measures of listening effort such as pupil dilation [103], could be employed for customising spatial hearing training routines, maximising outcomes while maintaining engagement and feasibility of the proposed tasks.

Final Thoughts

While most studies have focused on laboratory conditions to isolate specific perception elements, recent context-relevant studies have begun to examine the impact of spatial audio quality on task accomplishment. For example, [76] compared performance in a first-person-shooter VR game context with different HRTF conditions. Results showed performance for extreme elevation target positions was affected by the quality of HRTF matching. In addition, a subgroup of participants showed higher sensitivity to HRTF choice than others. At the same time, low-level sensory perception is only one of the dimensions where immersive audio simulations can have a significant impact. In order to significantly advance our understanding of the impact of HRTF personalisation in virtually rendered scenes and tasks, research needs to move beyond the evaluation of individual immersive audio tasks and metrics (e.g. sound localisation and/or perceived quality of the rendering), moving towards the evaluation of full experiences. The impact of immersive audio beyond perceptual metrics such as localisation, externalisation and immersion [87] is an as yet unexplored area of research, specifically when related with social interaction, entering the behavioural and cognitive realms.

In the past, several studies have been published in which auditory-based AR/VR interactions were created and evaluated without considering HRTF choice or using HRTF personalisation approaches that had not previously been appropriately validated from a perceptual point of view, or again ignoring the effects of HRTF accommodation, or blaming them in order to justify unexpected results. Considering our current knowledge and experience in immersive audio research, we are keen to recommend carrying out some level of personalisation of the spatial rendering when performing studies which involve auditory-based or multimodal interactions in AR/VR. As a baseline, ITDs can easily be customised to match the head circumference of the specific listener (as mentioned above, this function is already implemented in a few spatialisers, such as [22, 78]). Furthermore, HRTF selection routines, both perceptual and morphology based, could be very beneficial if carried out before the experiment, albeit it is important for the repeatability of such choices to be assessed with the specific subject (i.e. repeating the selection several times in order to verify the consistency across the trials, and possibly discard subjects/methods which do not show a sufficient level of repeatability). Regarding the use of synthesised HRTFs, until these are validated through extensive perceptual studies our advice is to use measured ones, possibly coming from the same dataset in order to avoid measurement-based differences.

In addition to these recommendations, it is important to emphasize that the future of immersive audio research will need to include studies focusing on different contexts (e.g. AR/VR interactions, virtual museum explorations and virtual assistant avatars), exploring the impact (and need) of HRTF personalisation on complex tasks such as interpersonal exchanges and distance learning in VR. Furthermore, in order to ensure a sufficient level of standardisation and consistently advance the achievements of research in this area, it seems evident that a concerted and coordinated effort across disciplines and research groups is highly desirable.

Acknowledgements Preparation of the chapter was made possible by support from SONICOM (www.sonicom.eu), a project that has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No. 101017743.

References

1. Algazi, V. R., Duda, R. O., Duraiswami, R., Gumerov, N. A., Tang, Z.: Approximating the head-related transfer function using simple geometric models of the head and torso. J Acoust Soc Am **112**, 2053–2064 (2002).
2. Algazi, V. R., Duda, R. O., Thompson, D. M., Avendano, C.: The cipic hrtf database in Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No. 01TH8575) (2001), 99–102.
3. Ananthabhotla, I., Ithapu, V. K., Brimijoin, W. O.: A framework for designing head-related transfer function distance metrics that capture localization perception. JASA Express Letters **1**, 044401:1–6 (2021).
4. Andreopoulou, A., Begault, D. R., Katz, B. F.: Inter-Laboratory Round Robin HRTF Measurement Comparison. IEEE J Selected Topics in Signal Processing **9**, 895–906 (2015).
5. Andreopoulou, A., Katz, B. F. G.: On the use of subjective HRTF evaluations for creating global perceptual similarity metrics of assessors and assessees in Intl Conf on Auditory Display (2015), 13–20.
6. Andreopoulou, A., Katz, B. F. G.: Investigation on Subjective HRTF Rating Repeatability in Audio Eng Soc Conv **140** (Paris, June 2016), 9597:1–10.
7. Andreopoulou, A., Katz, B. F.: Identification of perceptually relevant methods of inter-aural time difference estimation. J Acoust Soc Am **142**, 588–598 (2017).
8. Andreopoulou, A., Katz, B. F.: Subjective HRTF evaluations for obtaining global similarity metrics of assessors and assessees. Journal on Multimodal User Interfaces **10**, 259–271 (2016).
9. Aussal, M., Alouges, F., Katz, B. F.: ITD Interpolation and Personalization for Binaural Synthesis Using Spherical Harmonics in Audio Eng Soc UK Conf (York, UK, Mar. 2012), 04:01–10.
10. Bahu, H.: Localisation auditive en contexte de synthèse binaurale nonindividuelle PhD thesis (Université Pierre et Marie Curie-Paris VI, 2016).
11. Bahu, H., Carpentier, T., Noisternig, M.,Warusfel, O.: Comparison of different egocentric pointing methods for 3D sound localization experiments. Acta Acust **102**, 107–118 (2016).
12. Batteau, D. W.: The role of the pinna in human localization. Proceedings of the Royal Society of London. Series B. Biological Sciences **168**, 158–180 (1967).
13. Baumgartner, R., Majdak, P., Laback, B.: Modeling sound-source localization in sagittal planes for human listeners. J Acoust Soc Am **136**, 791–802 (2014).
14. Begault, D. R.: 3-D Sound for Virtual Reality and Multimedia (Academic Press, Cambridge, 1994).

15. Blum, A., Katz, B., Warusfel, O.: Eliciting adaptation to non-individual HRTF spectral cues with multi-modal training in 7ème Cong de la Soc Française d'Acoustique et 30ème congrès de la Soc Allemande d'Acoustique (CFA/DAGA) (Strasbourg, 2004), 1225–1226.
16. Bouchara, T., Bara, T.-G., Weiss, P.-L., Guilbert, A.: Influence of vision on short-termsound localization training with non-individualized HRTF in EAA Spatial Audio Signal Processing Symp (2019), 55–60.
17. Brinkmann, F., Weinzierl, S.: Comparison of Head-Related Transfer Functions Pre-Processing Techniques for Spherical Harmonics Decomposition English. in (Audio Engineering Society, Aug. 2018).
18. Brinkmann, F. et al.: A cross-evaluated database of measured and simulated HRTFs including 3D head meshes, anthropometric features, and headphone impulse responses. J Audio Eng Soc **67**, 705–718 (2019).
19. Carlile, S., Balachandar, K., Kelly, H.: Accommodating to new ears: the effects of sensory and sensory-motor feedback. J Acous Soc America **135**, 2002–2011 (2014).
20. Carlile, S., Leong, P., Hyams, S.: The nature and distribution of errors in sound localization by human listeners. Hearing Research **114**, 179–196 (1997).
21. Cuevas-Rodriguez, M., Gonzalez-Toledo, D., Reyes-Lecuona, A., Picinali, L.: Impact of non-individualised head related transfer functions on speechin- noise performances within a synthesised virtual environment. The JAcoust Soc Am **149**, 2573–2586 (2021).
22. Cuevas-Rodríguez, M. et al.: 3D Tune-In Toolkit: An open-source library for real-time binaural spatialisation. PloS one **14**, e0211899 (2019).
23. Deng, Y., Choi, I., Shinn-Cunningham, B., Baumgartner, R.: Impoverished auditory cues limit engagement of brain networks controlling spatial selective attention. NeuroImage **202**, 116151 (2019).
24. Dramas, F., Katz, B., Jouffrais, C.: Auditory-guided reaching movements in the peripersonal frontal space in Acoustics'08. 9e Congrèès Français d'Acoustique of the SFA. **123** (Acoustical Society of America, 2008), 3723.
25. Engel, I., Goodman, D. F. M., Picinali, L.: Assessing HRTF preprocessing methods for Ambisonics rendering through perceptual models. en. Acta Acustica **6**, 4 (2022).
26. Genuit, K.: A model for the description of outer-ear transmission characteristics PhD thesis (Rhenish-Westphalian Technical University, Düsseldorf, 1984), 220.
27. Geronazzo, M., Peruch, E., Prandoni, F.,Avanzini, F.: Applying a single-notch metric to image-guided head-related transfer function selection for improved vertical localization. Journal of the Audio Engineering Society **67**, 414–428 (2019).
28. Geronazzo, M., Spagnol, S., Avanzini, F.: Mixed structural modeling of headrelated transfer functions for customized binaural audio delivery in 2013 18th International Conference on Digital Signal Processing (DSP) (2013), 1–8.
29. Geronazzo, M., Spagnol, S., Avanzini, F.: Do we need individual head-related transfer functions for vertical localization? The case study of a spectral notch distance metric. IEEE/ACM Transactions on Audio, Speech, and Language Processing **26**, 1247–1260 (2018).
30. Geronazzo, M., Spagnol, S., Bedin, A.,Avanzini, F.: Enhancing vertical localization with image-guided selection of non-individual head-related transfer functions in 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2014), 4463–4467.
31. Greff, R., Katz, B.: Perceptual evaluation of HRTF notches versus peaks for vertical localisation in Intl Cong on Acoustics **19** (Madrid, Spain, 2007), 1–6.
32. Greff, R., Katz, B.: Round Robin comparison of HRTF simulation results : preliminary results. in Audio Eng Soc Conv **123** (New York, USA, 2007), 1–5.
33. Grijalva, F., Martini, L., Florencio, D., Goldenstein, S.: A manifold learning approach for personalizing HRTFs from anthropometric features. IEEE/ACM Transactions on Audio, Speech, and Language Processing **24**, 559–570 (2016).
34. Guezenoc, C., Seguier, R.: A wide dataset of ear shapes and pinna-related transfer functions generated by random ear drawings. J Acoust Soc Am **147**, 4087–4096 (2020).
35. Gumerov, N. A., O'Donovan, A. E., Duraiswami, R., Zotkin, D. N.: Computation of the head-related transfer function via the fast multipole accelerated boundary element method and its spherical harmonic representation. JAcoust Soc Am **127**, 370–386 (2010).

36. Hamari, J., Koivisto, J., Sarsa, H.: Does gamification work? A literature review of empirical studies on gamification in Intl Conf on System Sciences (2014), 3025–3034.
37. Hendrickx, E. et al.: Influence of head tracking on the externalization of speech stimuli for non-individualized binaural synthesis. J Acoust Soc Am **141**, 2011–2023 (2017).
38. Hofman, P. M., Van Riswick, J. G., Van Opstal, A. J.: Relearning sound localization with new ears. Nature Neuroscience **1**, 417–421 (1998).
39. Honda, A. et al.: Transfer effects on sound localization performances from playing a virtual three-dimensional auditory game. Applied Acoustics **68**, 885–896 (2007).
40. Hwang, S., Park, Y., Park, Y.-s.: Modeling and customization of head-related impulse responses based on general basis functions in time domain. Acta Acustica united with Acustica **94**, 965–980 (2008).
41. Iwaya,Y.: Individualization of head-related transfer functions with tournamentstyle listening test: Listening with other's ears. Acoustical Science & Technology **27**, 340–343 (2006).
42. Jin, C. T. et al.: Creating the Sydney York morphological and acoustic recordings of ears database. IEEE Transactions on Multimedia **16**, 37–46 (2013).
43. Jost, T. A., Nelson, B., Rylander, J.: Quantitative analysis of the Oculus Rift S in controlled movement. Disability and Rehabilitation: Assistive Technology, 1–5 (2019).
44. Kahana, Y.: Numerical modelling of the head-related transfer function PhD thesis (University of Southampton, 2000).
45. Kahana, Y., Nelson, P. A.: Boundary element simulations of the transfer function of human heads and baffled pinnae using accurate geometric models. Journal of Sound and Vibration **300**, 552–579 (2007).
46. Kahana, Y., Nelson, P. A., Petyt, M., Choi, S.: Boundary element simulation of HRTFs and sound fields produced by virtual acoustic imaging systems in Audio Engineering Society Convention 105 (1998).
47. Katz, B., Begault, D.: Round robin comparison of HRTF measurement systems : preliminary results. in Intl Cong on Acoustics **19** (Madrid, Spain, 2007), 1–6.
48. Katz, B., Nicol, R. in Sensory Evaluation of Sound (ed Zacharov, N.) 349–388 (CRC Press, Boca Raton, 2019).
49. Katz, B. F. G.: Measurement and Calculation of Individual Head-Related Transfer Functions Using a Boundary Element Model Including the Measurement and Effect of Skin and Hair Impedance PhD thesis (The Pennsylvania State University, 1998).
50. Katz, B. F. G., Noisternig, M.: A comparative study of interaural time delay estimation methods. J Acoust Soc Am **135**, 3530–3540 (2014).
51. Katz, B. F.: Boundary element method calculation of individual head-related transfer function. I. Rigid model calculation. J Acoust Soc Am **110**, 2440–2448 (2001).
52. Katz, B. F.: Boundary element method calculation of individual head-related transfer function. II. Impedance effects and comparisons to real measurements. J Acoust Soc Am **110**, 2449–2455 (2001).
53. Katz, B. F., Parseihian, G.: Perceptually based head-related transfer function database optimization. J Acoust Soc Am **131**, EL99–EL105 (2012).
54. Kim, C., Lim, V., Picinali, L.: Investigation Into Consistency of Subjective and Objective Perceptual Selection of Non-individual Head-Related Transfer Functions. J Audio Eng Soc **68**, 819–831 (2020).
55. Kreuzer, W., Majdak, P., Chen, Z.: Fast multipole boundary element method to calculate head-related transfer functions for a wide frequency range. J Acoust Soc Am **126**, 1280–1290 (2009).
56. Kreuzer, W., Majdak, P., Haider, A.: A boundary element model to calculate HRTFs. Comparison between calculated and measured data in Proceedings of the NAG-DAGA International Conference 2009 (2009), 196–199.
57. Kumpik, D. P., Kacelnik, O., King, A. J.: Adaptive reweighting of auditory localization cues in response to chronic unilateral earplugging in humans. J of Neuroscience **30**, 4883–4894 (2010).
58. Lopez-Poveda, E. A., Meddis, R.: A physical model of sound diffraction and reflections in the human concha. J Acoust Soc Am **100**, 3248–3259 (1996).

59. Majdak, P., Goupell, M. J., Laback, B.: 3-D localization of virtual sound sources: Effects of visual environment, pointing method, and training. Attention, Perception, & Psychophysics **72**, 454–469 (2010).
60. 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. J Acous Soc America **134**, 2148–2159 (2013).
61. Mendonça, C.: A review on auditory space adaptations to altered head-related cues. Frontiers in Neuroscience **8**, 219:1–14 (2014).
62. Mendonça, C., Campos, G., Dias, P., Santos, J. A.: Learning auditory space: Generalization and long-term effects. PloS One **8**, 1–14 (2013).
63. Mendonça, C. et al.: On the improvement of localization accuracy with nonindividualized HRTF-based sounds. J Audio Eng Soc **60**, 821–830 (2012).
64. Middlebrooks, J. C.: Individual differences in external-ear transfer functions reduced by scaling in frequency. J Acoust Soc Am **106**, 1480–1492 (1999).
65. Middlebrooks, J. C., Macpherson, E. A., Onsan, Z. A.: Psychophysical customization of directional transfer functions for virtual sound localization. J Acoust Soc Am **108**, 3088–3091 (2000).
66. Molloy, K., Moore, D. R., Sohoglu, E., Amitay, S.: Less is more: latent learning is maximized by shorter training sessions in auditory perceptual learning. PloS One **7**, 1–13 (2012).
67. Morimoto, M., Aokata, H.: Localization cues of sound sources in the upper hemisphere. J Acous Soc Japan **5**, 165–173 (1984).
68. Nicol, R.: Binaural Technology 77 (Audio Engineering Society, New York 2010).
69. Ospina, F. R., Emerit, M., Katz, B. F.: The 3D morphological database for spatial hearing research of the BiLi project in Proc. of Meetings on Acoustics **23** (Pittsburg, May 2015), 1–17.
70. Otani, M., Ise, S.: Fast calculation system specialized for head-related transfer function based on boundary element method. J Acoust Soc Am **119**, 2589–2598 (2006).
71. Parseihian, G., Katz, B., Conan, S.: Sound effect metaphors for near field distance sonification in Intl Conf on Auditory Display (Atlanta, June 2012), 6–13.
72. Parseihian, G., Katz, B. F. G.: Morphocons: A New Sonification Concept Based on Morphological Earcons. J Audio Eng Soc **60**, 409–418 (2012).
73. Parseihian, G., Katz, B. F. G.: Rapid head-related transfer function adaptation using a virtual auditory environment. J Acous Soc America **131**, 2948–2957 (2012).
74. Picinali, L., Afonso, A., Denis, M., Katz, B. F.: Exploration of architectural spaces by blind people using auditory virtual reality for the construction of spatial knowledge. International Journal of Human-Computer Studies **72**, 393–407 (2014).
75. Poirier-Quinot, D., Stitt, P., Katz, B. in Advances in Fundamental and Applied Research on Spatial Audio (eds Katz, B., Majdak, P.) (InTech, 2022).
76. Poirier-Quinot, D., Katz, B. F.: Assessing the impact of Head-Related Transfer Function individualization on task performance: Case of a virtual reality shooter game. J. Audio Eng. Soc **68**, 248–260 (2020).
77. Poirier-Quinot, D., Katz, B. F.: On the improvement of accommodation to non-individual HRTFs via VR active learning and inclusion of a 3D room response. Acta Acustica **5**, 1–17 (2021).
78. Poirier-Quinot, D., Katz, B. F.: The Anaglyph binaural audio engine in Audio Engineering Society Convention 144 (2018).
79. Poirier-Quinot, D., Parseihian, G., Katz, B. F.: Comparative study on the effect of Parameter Mapping Sonification on perceived instabilities, efficiency, and accuracy in real-time interactive exploration of noisy data streams. Displays **47**, 2–11 (2016).
80. Reichinger, A., Majdak, P., Sablatnig, R., Maierhofer, S.: Evaluation of methods for optical 3-D scanning of human pinnas in 2013 International Conference on 3D Vision-3DV 2013 (2013), 390–397.
81. Schonstein, D., Katz, B. F.: HRTF selection for binaural synthesis from a database using morphological parameters in International Congress on Acoustics (ICA) (2010).

82. Schönstein, D., Katz, B. F.: Variability in perceptual evaluation of HRTFs. Journal of the Audio Engineering Society **60**, 783–793 (2012).
83. Seeber, B. U., Fastl, H.: Subjective selection of non-individual head-related transfer functions in Proceedings of the 2003 Intl Conf on Auditory Display (ICAD) (2003), 259–262.
84. Shin, K. H., Park, Y.: Enhanced vertical perception through head-related impulse response customization based on pinna response tuning in the median plane. IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences **91**, 345–356 (2008).
85. Shukla, R., Stewart, R., Roginska, A., Sandler, M.: User selection of optimal HRTF sets via holistic comparative evaluation in the Audio Engineering Society Conference on Audio for Virtual and Augmented Reality (AVAR) 2018 (Audio Engineering Society, Redmond, WA, USA, 2018).
86. Silzle, A.: Selection and tuning of HRTFs in Audio Eng Soc Conv 112 (2002), 1–14.
87. Simon, L., Zacharov, N., Katz, B. F. G.: Perceptual attributes for the comparison of Head-Related Transfer Functions. J Acous Soc America **140**, 3623–3632 (Nov. 2016).
88. Søndergaard, P., Majdak, P. in The Technology of Binaural Listening (ed Blauert, J.) 33–56 (Springer, Berlin, Heidelberg, 2013).
89. Spagnol, S., Geronazzo, M., Avanzini, F.: On the relation between pinna reflection patterns and head-related transfer function features. IEEE transactions on audio, speech, and language processing **21**, 508–519 (2012).
90. Steadman, M. A., Kim, C., Lestang, J.-H., Goodman, D. F., Picinali, L.: Short-term effects of sound localization training in virtual reality. Scientific Reports **9**, 1–17 (2019).
91. Stitt, P., Katz, B. F.: Sensitivity analysis of pinna morphology on head-related transfer functions simulated via a parametric pinna model. J Acoust Soc Am **149**, 2559–2572 (2021).
92. Stitt, P., Picinali, L., Katz, B. F. G.: Auditory Accommodation to poorly MatchedNon-Individual spectral Localization Cues throughActive Learning. Scientific Reports **9**, 1063:1–14 (2019).
93. Teranishi, R., Shaw, E. A.: External-Ear Acoustic Models with Simple Geometry. J Acoust Soc Am **44**, 257–263 (1968).
94. Trapeau, R., Aubrais, V., Schönwiesner, M.: Fast and persistent adaptation to new spectral cues for sound localization suggests a many-to-one mapping mechanism. J Acous Soc America **140**, 879–890 (2016).
95. Van Wanrooij, M. M., Van Opstal, A. J.: Relearning sound localization with a new ear. J of Neuroscience **25**, 5413–5424 (2005).
96. Voong, T. M., Oehler, M.: Tournament Formats as Method for Determining Best-fitting HRTF Profiles for Individuals wearing Bone Conduction Headphones in Proceedings of the 23rd International Congress on Acoustics : integrating 4th EAA Euroregio 2019 : 9–13 September 2019 in Aachen, Germany (eds Ochmann, M., Vorländer, M., Fels, J.) (Berlin, Germany, Sept. 9, 2019), 4841–4847.
97. Wan, Y., Zare, A., McMullen, K.: Evaluating the consistency of subjectively selected head-related transfer functions (HRTFs) over time in Audio Engineering Society Conference: 55th International Conference: Spatial Audio (2014).
98. Warusfel, O.: IRCAM Listen HRTF database http://recherche.ircam.fr/equipes/salles/listen. 2003.
99. Wenzel, E. M., Arruda, M., Kistler, D. J.,Wightman, F. L.: Localization using nonindividualized head-related transfer functions. J Acous Soc America **94**, 111–123 (1993).
100. Whitton, J. P., Hancock, K. E., Shannon, J. M., Polley, D. B.: Audiomotor perceptual training enhances speech intelligibility in background noise. Current Biology **27**, 3237–3247 (2017).
101. Wightman, F. L., Kistler, D. J.: The dominant role of low-frequency interaural time differences in sound localization. J Acoust SocAm **91**, 1648–1661 (Mar. 1992).
102. Wightman, F. L., Kistler, D. J.: Resolution of front-back ambiguity in spatial hearing by listener and source movement. J Acoust Soc Am **105**, 2841–2853 (1999).
103. Winn, M. B., Wendt, D., Koelewijn, T., Kuchinsky, S. E.: Best practices and advice for using pupillometry to measure listening effort: An introduction for those who want to get started. Trends in hearing **22**, 1–32 (2018).

104. Wright, B. A., Zhang, Y.: A review of learning with normal and altered sound-localization cues in human adults. Intl J of Audiology **45**, 92–98 (2006).
105. Xie, B.: Head-Related Transfer Functions and Virtual Auditory Display 2nd ed. (J. Ross Publishing, Plantation, FL, USA, 2013).
106. Yairi, S., Iwaya,Y.,Yôiti, S.: Individualization feature of head-related transfer functions based on subjective evaluation in 14th Intl Conf onAuditory Display (Paris, 2008).
107. Zacharov, N., Lorho, G.: What are the requirements of a listening panel for evaluating spatial audio quality? in Proc. Int.Workshop on Spatial Audio and Sensory Evaluation Techniques (2006).
108. Zagala, F., Noisternig, M., Katz, B. F.: Comparison of direct and indirect perceptual head-related transfer function selection methods. J Acoust Soc Am **147**, 3376–3389 (2020).
109. Zahorik, P., Bangayan, P., Sundareswaran, V.,Wang, K., Tam, C.: Perceptual recalibration in human sound localization: Learning to remediate front-back reversals. J Acous Soc America **120**, 343–359 (2006).
110. Ziegelwanger, H., Kreuzer, W., Majdak, P.: Mesh2HRTF: An open-source software package for the numerical calculation of head-related transfer functions in 22nd International Congress on Sound and Vibration (2015).
111. Ziegelwanger, H., Majdak, P., Kreuzer,W.: Numerical calculation of listenerspecific head-related transfer functions and sound localization: Microphone model and mesh discretization. J Acoust Soc Am **138**, 208–222 (2015).
112. Zotkin, D., Hwang, J., Duraiswaini, R., Davis, L. S.: HRTF personalization using anthropometric measurements in 2003 IEEEWorkshop on Applications of Signal Processing to Audio and Acoustics (IEEE Cat. No. 03TH8684) (2003), 157–160.