Effect of Interpolation Artifacts on Perceived Stability of Nearby Sources in a Navigable Reverberant Virtual Environment

Julien De Muynke, David Poirier-Quinot, Brian F. G. Katz

OPEN ACCESS Freely available online

J. De Muynke, D. Poirier-Quinot, and B. F. G. Katz,
"Effect of Interpolation Artifacts on Perceived Stability
of Nearby Sources in a Navigable Reverberant Virtual Environment,"
J. Audio Eng. Soc., vol. 72, no. 10, pp. 664–678 (2024 Oct.).
https://doi.org/10.17743/jaes.2022.0158.

Effect of Interpolation Artifacts on Perceived Stability of Nearby Sources in a Navigable Reverberant Virtual Environment

JULIEN DE MUYNKE,[1, 2] *AES Student Member*, **DAVID POIRIER-QUINOT,**[1] *AES Associate Member* **AND**
(julien.de_muynke@sorbonne-universite.fr) (david.poirier-quinot@sorbonne-universite.fr)

BRIAN F. G. KATZ,[1] *AES Member*
(brian.katz@sorbonne-universite.fr)

[1] *Sorbonne Université, CNRS, Institut Jean Le Rond ∂'Alembert, UMR 75005, Paris, France*
[2] *Eurecat, Technology Center of Catalonia, Multimedia Technologies, Barcelona, 08005 Spain*

Convolution with spatial Room Impulse Responses can achieve realistic auralizations. When combined with interpolation between spatially distributed RIRs, this technique can be used to create navigable virtual environments. This study explores the impact of various interpolation parameters on the perceived auditory stability of a nearby static sound source with listener movements in a reverberant environment. The auditory scene was rendered via third-order Ambisonic RIR convolution combined with magnitude-least-squares binaural decoding using nonindividualized head-related transfer functions. First, the estimated direction of arrival as a function of the listener's position within a 2D grid of RIRs under various configurations is examined as an objective metric. The perceived stability of the auditory source is then assessed through a perceptual experiment. Participants freely explored a virtual scene reproduced over headphones and a tracked head-mounted display. They were asked to rate the stability of a nonvisual source under various conditions of RIR grid density, interpolation panning method, and reverberation time. Results indicate no need to use an RIR grid size finer than 1 m to optimize source stability when using a three-nearest-neighbor interpolation scheme.

0 INTRODUCTION

In recent years, auditory virtual and augmented reality experiences have been increasingly popular. Among possible public applications, they can be deployed in heritage sites for immersive audio-guided visits. In such situations, proximity sensors can be distributed throughout the visitor area to transition the reproduced audio content between zones automatically. Other sites employ tracked headphones to trigger audio content according to the visitor's head position and orientation [1], e.g., a character starting to talk when the visitor looks at his portrait [2]. In this context, the degree of immersion in the virtual element of the augmented auditory scene greatly depends on the plausibility of the auralized scene, i.e., on the perceived similarity of its sonic attributes with those of the real world in which it is overlaid.

To produce realistic auralizations, the implementation of such an experience often relies on room impulse response (RIR) convolution. With six-degrees-of-freedom listener tracking, auralizations are continuously adjusted to the listener's position within the scene in real time. This is achieved by generating a local representation of the RIR, a process referred to as *spatial interpolation*. Compared to rendering a static position using a unique RIR, interpolation can result in audible artifacts, potentially degrading the plausibility of the rendered environment. One generally used approach to such spatial interpolation is through spatial panning, based on existing RIRs distributed on a grid, where nearby RIRs are selected and combined according to various weighting schemes based on the listener position in the virtual environment. The current study aims to characterize the impact of RIR interpolation artifacts on the *perceived source stability* during free exploration of an audiovisual virtual environment for various panning interpolation methods and RIR grid densities in this context.

1 PREVIOUS WORK

Several techniques have been proposed to create dynamic auditory virtual environments, i.e., environments in which users can navigate freely while listening to virtual auditory

sources resonating in the space around them. Some target high localization precision and low CPU usage, using, e.g., anechoic binaural rendering coupled to an artificial reverberation [3]. Others might augment this workflow with a real-time image-source renderer, integrating coherent early reflections to increase authenticity while limiting the maximum number of concurrent audio sources in the scene [4]. For more demanding scenarios where RIRs cannot be wholly generated on the fly, e.g., echolocation training [5], archaeoacoustic studies [6], or concert hall acoustic evaluations [7], one generally relies on auralizations based on RIR convolutions. Enabling user motion in such systems requires a pre-existing discrete grid of RIRs paired with an interpolation scheme to facilitate the impression of continuous movement between uninterpolated RIR grid nodes.

Two main families of interpolation schemes are linear panning and parametric panning. Linear panning methods are generally simpler, computationally cost-efficient, and agnostic to the audio content being rendered. In contrast, parametric panning typically relies on the precomputation of acoustic features within the RIRs or the audio streams. Such precomputations are used to establish, for example, the predominant direction of arrival (DOA) of energy as a function of frequency, which is then used to generate a new interpolated audio stream [8]. They can also be used to identify perceptually relevant reflections in order to isolate, reposition, and synthesize a new position-appropriate RIR [9].

As an example of techniques based on the interpolation of spatially distributed Ambisonic streams, the interpolation method proposed in [10] is based on a regularized least-squares approach and was shown to introduce minimal spectral coloration through an objective evaluation. A different approach is proposed in [11], where the authors use variable weighting on spherical harmonic coefficients of different orders as a function of the distance of the interpolated position. The added value of the approach in terms of naturalness/realism and smoothness of the auditory image was assessed through a MUSHRA test, where the virtual scene was displayed on a computer screen using a prerendered cinematic along a predefined trajectory. In [12], the authors proposed an interactive binaural audio rendering system based on complex synthesized Ambisonic streams distributed along a given grid. Depending on the listener's position in the navigable space, the three closest Ambisonic streams were selectively redirected to the portable renderer and spatially mixed using a linear panning approach.

As an example of techniques that rely on the interpolation of RIRs prior to on-the-fly convolution, [13] proposed a perceptually informed method that interpolates a sparse grid of Higher Order Ambisonic RIRs through separate treatment of the direct sound, early reflections, and late reverberation. The method has been shown to be robust to changes in room acoustics between coupled rooms. In [14] and [15], the authors performed time-warping of sparse sets of measured omnidirectional and Ambisonic RIRs, respectively, in order to time-align early reflections prior to spatial up-sampling of the RIR set to reduce spatial blur. This upsampling improved localization accuracy for static listener scenarios, yet the method was not tested in navigation conditions. Finally, [16] proposed a method for up-sampling a grid of Ambisonic RIRs based on joint localization of early reflection peaks across RIRs and adjustment of their temporal and directional characteristics before linear interpolation. According to tests conducted on prerendered listener trajectories, this method achieved higher localization accuracy and less coloration artifacts than the more straightforward interpolation approach.

As previously noted, most parametric interpolation techniques require some prior knowledge of the RIR and/or audio content, relying on CPU intensive precomputation and rendering steps. As such, while they have been shown to outperform straightforward linear panning interpolation schemes when evaluated by seated participants focused on the auralization alone, it is worth considering that simpler schemes may be sufficient to create plausible auralizations in more dynamic navigable multi-modal scenes. In such cases, there is interest in investigating the effect of RIR grid density on the perceived continuity of the rendered sound scene in the context of navigable virtual reality auralization. This was previously studied in [17, 18], using a one-nearest-neighbor (1NN) selection, without panning, with binaural RIR (BRIR) convolution. The current study extends these previous studies by characterizing how RIR grid density, linear panning method, and room acoustics impact perceived stability during listener navigation.

2 RESEARCH CONTRIBUTIONS

2.1 Overview

Through objective and subjective evaluations, the current study aims to identify design criteria for creating artifact-free, high-fidelity auralizations for dynamic user navigation, particularly in augmented reality applications for museum visits and cultural heritage sites. This involves audio reproduction not tailored to individual users and is implemented on ordinary portable devices with limited computational and storage capacity.

The notion of quality was evaluated based on the criteria of source stability, chosen because it combines several spatial characteristics such as azimuth/elevation localization and perception of source distance. This relationship between an auditory source's perceived distance and stability has been studied in [19]. Assessing the minimum audible angle (MAA) induced by the listener's self-translation, the authors showed that the further away the source, the less impaired the absolute localization of stationary sound sources is. The perceived stability of nearby sources is thus expected to be reduced compared to distant sources.

Previous studies have also examined the impact of including room acoustics in virtual reality interactions [20]. Consequently, it may be expected that early reflections and room reverberation could impact source stability. A large room was preferred in this study to limit proximal surfaces and, hence, to minimize the contribution of early room reflections, which are highly specific to room geometry and surface acoustic parameters. Subsequently, modifying ma-

terial property definitions in the room model allows for variations in reverberation time while still limiting specific geometrical effects.

2.2 Focus

A commonly employed linear panning interpolation scheme was evaluated under various conditions using a grid of third-order Ambisonic RIRs. Ambisonic to binaural decoding was performed using the magnitude-least-squares (MagLS) binaural decoding scheme [21] and generic, nonindividualized head-related transfer function (HRTF).

Ambisonic RIR interpolation was selected over BRIRs [17, 18], because the use of Ambisonic RIRs facilitates dynamic listener head rotation at minimal computational and storage cost, though resulting in some decrease in spatial precision as a function of Ambisonic order. While a single Ambisonic RIR (third-order requiring 16 channels) at each listening position is sufficient to enable three-degrees-of-freedom listener head rotation, a significant number of BRIRs are required for the same listener position to account for dynamic rotation (e.g., a 15° solid angle spherical grid in azimuth and elevation requires more than 100 BRIRs). Additionally, in contrast to the direct BRIR approach, which encodes directly a given HRTF from the start, the Ambisonic RIR approach allows for the use of individualized HRTF applied at run-time, potentially increasing the rendering quality through HRTF individualization schemes [22], even though nonindividualized HRTF was used in the current formulation of the study. The choice of third-order Ambisonics was motivated by its current acceptance within real-world use cases, as it appears to be the standard Higher Order Ambisonic format supported by spatial audio platforms (e.g., Resonance Audio, dearVR Pro, VLC).

Various grid densities were compared, with higher densities expected to lead to an increase in perceived source stability as seen previously in [17]. In this study, the authors used a 1NN method on a grid of BRIRs, referred to as $1NN_{BRIR}$ in the following. The current study compared two three-nearest-neighbors (3NN) interpolation methods with different weighting schemes and a 1NN selection method across different grid sizes and room acoustics. Although 3NN interpolation methods are expected to generally outperform 1NN selection methods at the expense of higher complexity, this comparison aimed to unveil potential trade-offs between the computational cost of the employed method and the storage requirement on the reproduction device as a function of the RIR grid density.

The manuscript is organized as follows. SEC. 3 describes the various RIR grids and interpolation methods evaluated. SEC. 4 presents the results of a DOA error model based on direct sound interaural time difference (ITD) analysis as a first-order predictor of subjective responses. SEC. 5 details the perceptual test comparing the impact on perceived stability of different panning methods applied to the different RIR grid density conditions during free exploration. SEC. 6 presents the results of the perceptual experiment as compared to those of the DOA error prediction model, followed by the discussion in SEC. 7 including a comparison with results of [17] using $1NN_{BRIR}$. The comparison was conducted to test the validity of the proposed auralization method and to offer further insights on the cost-quality ratio of each tested method, thereby enabling educated choice for designing navigable auditory environments for real-world applications.

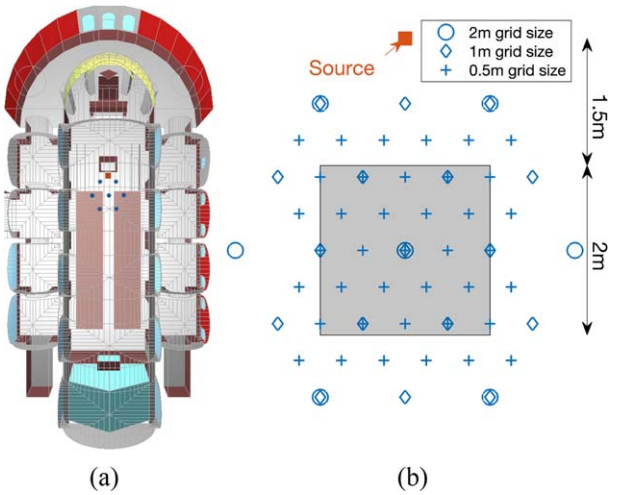


Fig. 1. (a) Top-down view of the GA model of St. Elisabeth Church, with the 2-m grid size of virtual receivers (circles) and the source (square) positioned near the altar. (b) Arrangement of the 2 × 2 m^2 navigation area (gray area) and the simulated RIR positions for all triangular grid sizes 0.5, 1, and 2 m.

3 MATERIALS AND METHODS

3.1 RIR Grids

A set of third-order Ambisonic RIRs was numerically simulated using a calibrated geometrical-acoustic (GA) model of the St. Elisabeth Church in Paris developed in CATT-Acoustic [23], illustrated in Fig. 1. A simulation was employed due to its practicality in generating various grids with exact positioning and orientation of individual receivers and its ability to simulate various room acoustics by simply modifying the room model. The acoustics modeling software was chosen because it has been shown to be capable of producing perceptually equivalent auralizations on well-calibrated models [6]. A well-behaved church model was chosen with regard to the previously mentioned conditions on room acoustics, with its size allowing for the RIR evaluation grids to be sufficiently distant from proximal walls or occluding elements.

An omnidirectional source was located in front of the altar at a height of 1.5 m with receivers distributed in the same horizontal plane. A navigation area, covering a 2 × 2 m^2 square, was located in the central nave, on the symmetry axis of the church, at a minimum distance of 1.5 m from the source position. RIR grids of various spatial densities, composed of equilateral triangular cells of various edge lengths—referred to as *grid sizes* in the following—were generated to cover the entire navigation area and aligned on the center node. Fig. 1 shows the source position and navigation area covered by grids of sizes 0.5, 1, and 2 m.

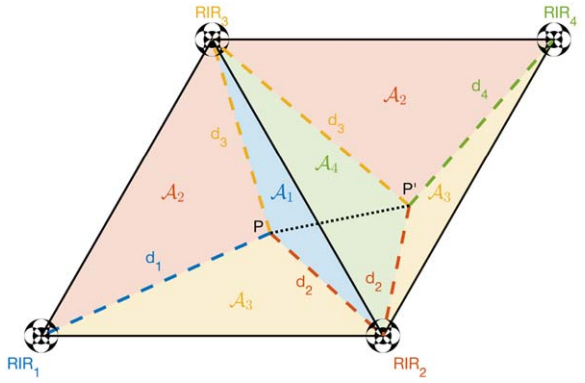


Fig. 2. Visualization of distances d_i and subtriangle surface areas A_i for target positions P in Cell_{123} and P' in Cell_{234}. The black dotted-line depicts the path $[PP']$ that crosses the edge between Cell_{123} and Cell_{234}.

A known issue with interpolating between spatially distributed RIRs is that of comb-filtering effects due to slight differences in time of arrival of the direct sound and reflections [16, SEC. 5.4]. All generated RIRs were time-aligned on the direct sound by trimming the leading zeros corresponding to the propagation time to avoid the most notable artifacts, although the reflections contained in the contributing RIRs remain non–time-aligned because no dynamic time warping is applied. It must be noted that in other auditory scenarios where early reflections prevail over the direct sound, e.g., a highly directional source pointing away from the listener in a room with strong early reflections, the time-alignment of the RIRs on the direct sound may lead to an alteration of various sound attributes including the perceived source stability itself. This study uses an omnidirectional source and a navigation area positioned relatively far from most reflecting surfaces, making the direct sound prevail over early reflections.

3.2 Panning Methods

Where more advanced interpolation methods exist, as presented in SEC. 1, the current study uses simple non-parametric RIR interpolation in the time domain without separate treatment of the direct sound, early reflections, and late reverberation. It compares two 3NN interpolation methods with different weighting schemes with a 1NN selection method.

In the following, a cell bounded within the triplet RIR_i, RIR_j, and RIR_k is denoted Cell_{ijk}. For any given target position contained in Cell_{ijk}, the panning method provides the amplitude weights applied to RIR_i, RIR_j, and RIR_k in the spatial interpolation. Two target positions P and P' in the two adjacent cells Cell_{123} and Cell_{234}, depicted in Fig. 2, are used to illustrate the difference between the three panning methods considered in this study.

- **1NN**: Only the nearest-neighbor RIR is selected. Its weight is set to 1 regardless of target position details. In Cell_{123}, RIR_2 is the closest to target position P. It follows:

$$RIR_P = RIR_2. \tag{1}$$

  1NN is employed in, e.g., [17].
- **3NN_{dist}**: the 3NN RIRs are selected. Their interpolation weights w_{dist_i} are inversely proportional to their respective distance to the target position d_i.

$$w_{dist_{i,i\in(1:3)}} \propto 1/d_{i,i\in(1:3)}. \tag{2}$$

  It follows:

$$RIR_P = \frac{RIR_1/d_1 + RIR_2/d_2 + RIR_3/d_3}{\sum_{j=1}^{3} 1/d_j}. \tag{3}$$

  3NN_{dist} is employed in, e.g., [13] and [15].
- **3NN_{area}**: The 3NN RIRs are selected. Their interpolation weights w_{area_i} are proportional to the surface area of the subtriangle formed by the two other selected RIRs and the target position (denoted A_i).

$$w_{area_{i,i\in(1:3)}} \propto A_{i,i\in(1:3)}. \tag{4}$$

  It follows:

$$RIR_P = \frac{A_1 \cdot RIR_1 + A_2 \cdot RIR_2 + A_3 \cdot RIR_3}{\sum_{j=1}^{3} A_j}. \tag{5}$$

  3NN_{area} is employed in, e.g., [24].

3NN_{dist} is often referred to as Inverse Distance Weighting in geographic information systems literature [25], whereas weights w_{area_i} provided by 3NN_{area} are named barycentric coordinates or sometimes areal coordinates in computer graphics literature [25]. They are calculated similarly to vector base amplitude panning gains [26] but without the equal-power constraint.

Although 3NN_{dist} and 3NN_{area} always select identical RIR triplets for any given target position, they may lead to different interpolation weights. At target position P, $d_{i,i\in(1:3)}$ are rather homogeneous, whereas A_1 seems to be significantly smaller than A_2 and A_3. It follows a more balanced contribution scheme between $\text{RIR}_{i,i\in(1:3)}$ for 3NN_{dist} compared with 3NN_{area}, which attributes a significantly larger weight to RIR_2 and RIR_3 compared with RIR_1.

When transitioning between adjacent cells, the RIR unique to the previous triplet is substituted by the RIR unique to the new triplet. For example, when transitioning between Cell_{123} and Cell_{234}, RIR_1 is substituted by RIR_4. Along the path $[PP']$, the subtriangle area A_1 in Cell_{123} decreases all the way down to 0 (the subtriangle becomes degenerate) before A_4 in Cell_{234} starts to increase from 0. This ensures a smooth RIR substitution during the transition. By contrast, on the transition edge d_1 evidently has a finite value, making the contribution of RIR_1 in the interpolation nonzero when it gets substituted by R_4, itself with a nonzero contribution. In practice, when navigating

from P to P' along $[PP']$ w_{dist_1} reaches 0.22 on the transition edge, so RIR_1 still contributes of up to 22% in the interpolation before being suddenly substituted by RIR_4. Consequently, $3NN_{area}$ may offer smoother transition (i.e., no abrupt switches between RIRs) than $3NN_{dist}$ when transitioning between adjacent cells.

For real-time applications, the computational complexity of the reproduction system is a point of interest in application design. The use of interpolation between a triplet of RIRs, a more complex processing technique than the simpler selection of the nearest-neighbor RIR, generally entails a higher computational cost. Furthermore, 3NN methods can be implemented in various ways, differing in the order sequence in which they perform interpolation and convolution processing steps. These different implementations ideally produce identical output signals yet lead to different computational costs and potential artifacts.

One particular implementation, well suited to static auralizations, first performs the interpolation between the selected RIRs before convolution with the input signal, requiring only a single convolution operation with the resulting interpolated RIR. An alternative implementation, well suited to dynamic auralizations with time-varying filters, first performs the convolution of the input signal for each of the three selected RIRs, followed by interpolation between the resulting audio streams. This implementation comes at a higher computational cost due to the additional convolutions, though with potentially less audible artifacts during movement, as the convolution filters are not continuously updated at every position change. Further details on the various implementations of the 3NN methods can be found in [27, SEC. 4.2].

3.3 Audio Rendering

The rendering was accomplished using the RoomZ plugin [27] that performs spatial interpolation and uniformly partitioned convolution with the input source stimuli. The different interpolation grids and room acoustics were combined and defined as independent source-locked scenarios in a single RoomZ configuration XML file, each specifying the position of the source and of the third-order Ambisonic RIRs of the corresponding grid. The different panning methods were set using the *Neighboring RIRs selection* parameter: 1NN with $KNN = 1$, $3NN_{dist}$ with $KNN = 3$, and $3NN_{area}$ with *Delaunay* selection, using the barycentric coordinates for the weights of the selected RIRs. By combining a given scenario with a given value of the *Neighboring RIRs selection* parameter, it was then possible to simulate all considered interpolation scenarios.

RoomZ was configured with multi-threading convolution to spread the CPU consumption across cores and set to use *post-convolution interpolation*, performing the convolution before the interpolation to support dynamic auralizations better as discussed in SEC. 3.2. The plugin cross-fade time, i.e., the time it took to cross-fade between an old and a new convolution line when one of the neighboring RIR changed, was set to 50 ms. This value was selected as it provided a zipper noise-free cross-fade while keeping the plugin responsive enough to update RIRs during navigation, even for the 0.5-m grid. The binaural decoding was performed using the IEM BinauralDecoder plugin employing the MagLS decoding scheme and the nonindividual Neumann KU100 artificial head HRTFs.[1]

4 DOA PREDICTION ANALYSIS

To maintain the perceived source stability during navigation, the perceived DOA of a static source must change consistently with the moving listener's position. At any arbitrary static position, interpolation between DOA-specific RIRs may result in a DOA error with respect to the same listening situation in real conditions. In the dynamic listener case, the variation of DOA error may lead to perceived shifts or even jumps of the source position, detrimental for the perceived source stability. This was examined using a DOA prediction model of the direct sound of a nearby static source based on ITD estimation across the tested navigation area. The model is limited to the direct sound in order to avoid any disturbances due to room reflections. By comparing the reference DOA with the DOA resulting from the RIRs interpolation, the model can predict perceived source stability in the navigation area as a function of grid size and panning method.

4.1 DOA Prediction Method

The ITD is an important cue for sound source localization, especially for lateral displacements of auditory events [28, p.141]. ITD values may range from 0 µs for a source in the median plane to about 800 µs depending on the head size for a source on the interaural axis [22, Fig. 11.1]. In the horizontal plane, the ITD is directly mapped to the source azimuth. Because the MAA has been reported to be as low as 1° in certain regions of space such as the frontal area [29, 30], a slight change of ITD conveys a change in the perceived DOA.

The DOA error of the direct sound was computed as the difference between the reference DOA resulting from a real listening situation in anechoic conditions and the interpolated DOA estimated on the anechoic BRIRs obtained after binaural decoding of the interpolated Ambisonic RIRs. At each node of a grid covering the navigation area (grid spacing = 10 cm), the reference DOA was calculated as the sound angle of incidence with a standard trigonomet-

[1]Concerning HRTF, the term *individual* identifies the HRTF of the user, *individualized* or *personalized* is used to indicate an HRTF modified or selected to best accommodate the user, and *nonindividual* or *nonindividualized* to indicate an HRTF that has not been tailored to the user. A *generic* or *artificial-head* HRTF is a specific instance of nonindividual HRTF, often designed with the goal of representing a certain pool of subjects.

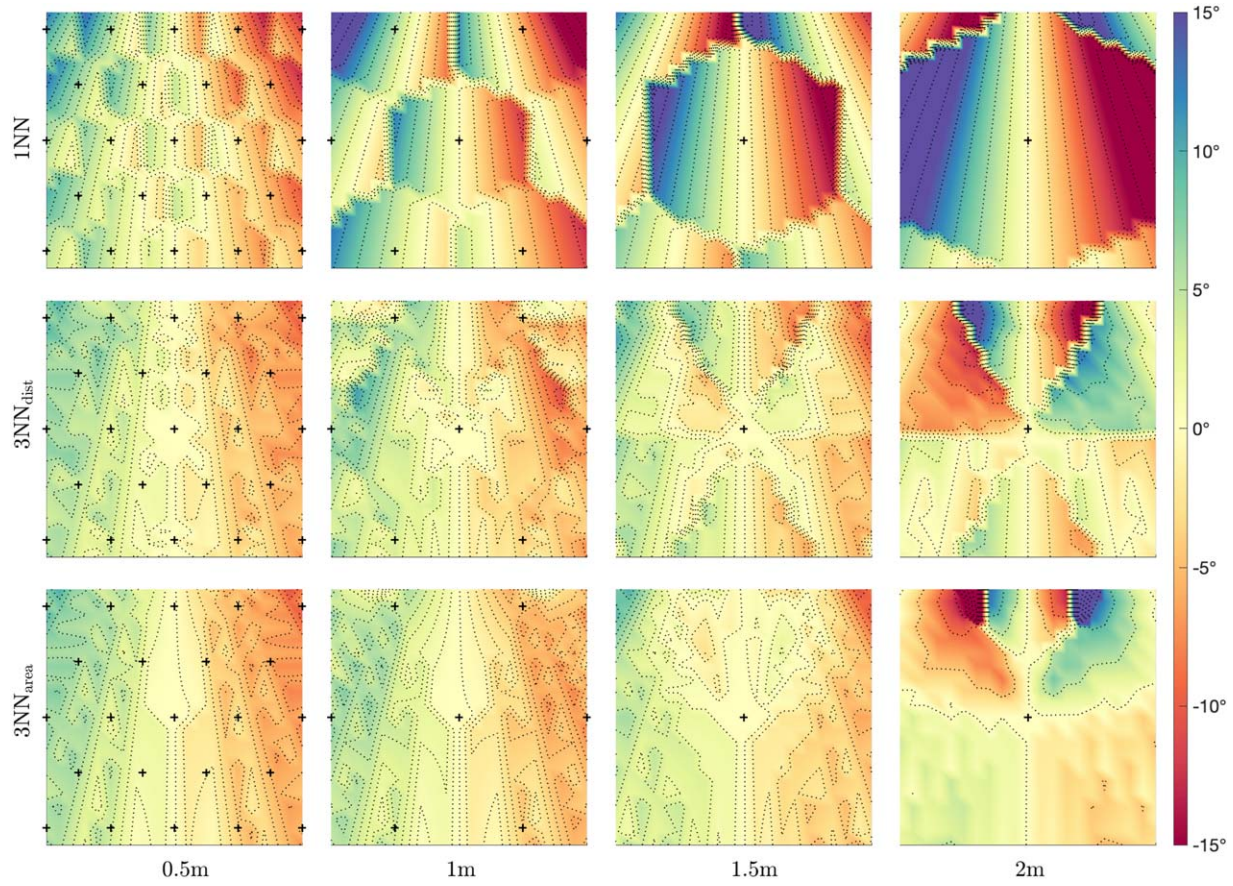


Fig. 3. Maps of predicted DOA error for a frontal static source in anechoic conditions at a minimum distance of 1.5 m from the navigation area, for various (rows) panning methods and (columns) RIR grid densities. iso-azimuth contour lines of DOA error are shown in 2° steps. Solid lines indicate positive DOA error, dotted lines indicate negative DOA error. RIR grid nodes within the navigation area are depicted by crosses.

ric approach. The interpolated DOA was estimated using the onset threshold detection ITD model applied to a low-pass filtered (3 kHz) version of the anechoic interpolated BRIRs. This method estimates the ITD as the difference of first time of arrival of the incident sound between the left and right ear signals. A threshold of −3 dB was used, because −20 or −30 dB thresholds recommended by [31] lead to erroneous ITD estimation due to pre-ringing present in BRIRs obtained from low Ambisonic orders decoding [32].

For each panning method and grid size, the anechoic interpolated BRIRs were generated by sending impulses to the RoomZ plugin loaded with "anechoic" Ambisonic RIRs, calculated with the same GA model with all reflection coefficients set to zero. The varying position along the sampling grid was controlled by changing the receiver position in RoomZ through an automation track.

Finally, resulting ITD values were paired with their corresponding DOA values through a mapping function computed by analyzing a 1° resolution horizontal subset of the KU100 artificial head HRTF (the same head used in the IEM BinauralDecoder), HRIR_CIRC360.sofa [33] stored in SOFA format [34]. This provided an estimation of the resulting DOA after interpolation to be compared against the reference DOA across the entire navigation area.

4.2 DOA Error Maps

The DOA error of the direct sound was estimated in the navigation area for all considered panning methods and for triangular grids with horizontal inter-RIR distances of 0.5, 1, 1.5, and 2 m. Maps of Fig. 3 show the DOA error for a frontal orientation, i.e., listener looking in front toward the region of space containing the source, where lighter areas denote a null DOA error and darker areas denote a maximum DOA error magnitude of 15° (blue: positive DOA error, i.e., to the left of the reference source position; red: negative DOA error, i.e., to the right of the reference source position). Dashed lines are iso-azimuth error contours with an interval of 3°. A high density of iso-azimuth error contours denotes a fast variation of the DOA error, potentially resulting in noticeable translation movements or even jumps of the perceived source position.

The highest DOA errors are obtained for 1NN with all grid sizes and for 2-m grid size with all panning methods. Regardless of the interpolation parameters, the DOA error is consistently low on the central symmetry axis of the navigation area, except for 1NN that exhibits DOA error variations in some regions. This means that the source is expected to be stable while the listener navigates along the central axis, i.e., to perceptually remain in front. More-

over, except for 2-m grid size and 1NN, the DOA error is generally positive in the left half of the navigation area and generally negative in the right half of the navigation area. This means that when the listener moves laterally, the source position is expected to shift laterally to the same direction. DOA error is generally higher in the front half than in the rear half of the navigation area, meaning that the closer the listener is to the simulated source position, the higher the lateral deviation of the perceived source position.

It must be noted that, in the absence of source visual cues, smooth DOA error variations during navigation may not necessarily result in perceived source instability since it may be harder to identify the actual reference position of the source. In contrast, abrupt DOA error variations are likely to be perceived as source instabilities despite the absence of source visual cues.

In Fig. 3, the density of iso-azimuth error contours indicates how fast the DOA error may change while the listener moves throughout the navigation area. For 1NN, the contours tile the DOA error maps in hexagonal cells centered on the nodes of the RIR grid. The lower the grid density, the larger the hexagonal cells and the higher the extent of DOA error within each tile. Inside the hexagons, iso-azimuth error contours form straight lines homogeneously distributed and converging as they approach the source. This means that the DOA error is expected to vary smoothly as long as the listener remains inside the hexagon, and within a range of lower extent in the region of the hexagon that is more distant from the source. When transitioning to an adjacent hexagon, the listener crosses a high density of iso-azimuth error contours, provoking a potentially abrupt variation of the DOA error, which is all the greater as the grid density decreases. The most abrupt variation of DOA error for 1NN is observed when navigating between the two front corners of the navigation area for the 2-m grid size.

In general, iso-azimuth error contours for $3NN_{dist}$ and $3NN_{area}$ are more homogeneously spread across the navigation area compared to 1NN, leading to less frequent abrupt changes of the source position for the 3NN methods than for the 1NN method. With the highest grid density, contour distributions are similar for $3NN_{dist}$ and $3NN_{area}$, leading to comparable source stability for these two panning methods, which decreases as the grid density decreases. Moreover, for $3NN_{area}$ the distribution is maintained across 0.5-m and 1-m grid sizes.

For grid sizes of 1.5 m and above, unlike $3NN_{area}$, $3NN_{dist}$ exhibits high concentrations of iso-azimuth error contours forming edges connecting the center node to the other RIRs distributed outside the navigation area. This may be explained by the abrupt switch inherent to $3NN_{dist}$ between the substituted and the substituting RIRs from adjacent cells, which does not occur with $3NN_{area}$ as discussed in SEC. 3.2. For the 2-m grid size, whereas the rear half of the navigation area exhibits a relatively low density of contours, both $3NN_{dist}$ and $3NN_{area}$ exhibit high concentrations of iso-azimuth error contours near the front edge of the navigation area connecting the two front corners. This similarity between the two tested 3NN methods indicate that the expected difference in smoothness during cell transitions between the two methods is less pronounced as the distance to the source decreases and as the grid size decreases.

The front corners of the navigation area exhibit the highest DOA error for various interpolation conditions and are sometimes separated by a high density of iso-azimuth error contours. As a result, the front corners seem to be the most critical zones for the perceived source stability, as much as the front edge connecting those two corners seems to be a critical region.

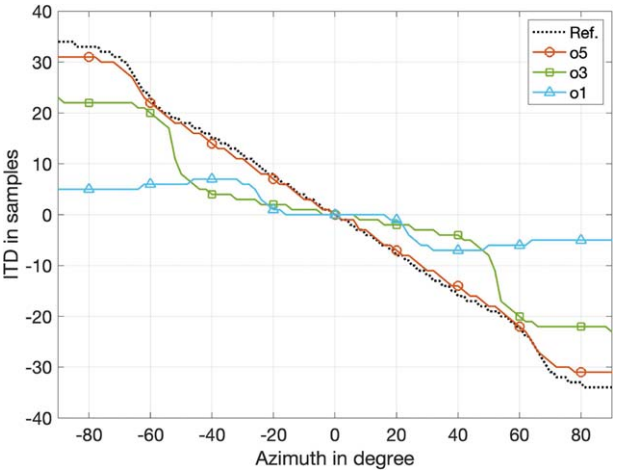


Fig. 4. Estimated ITD values as a function of azimuth, for the original HRTF (Ref.) and the anechoic BRIRs obtained through MagLS binaural decoding of Ambisonic encoded impulses for fifth (o5), third (o3), and first-order (o1) Ambisonics.

4.3 ITD Compression Effect

Because 1NN uses only the 1NN RIR for any given target position, the DOA error may be expected to be null on all grid nodes of 1NN maps. However, only the grid nodes on the central symmetry axis of the navigation area exhibit a null DOA error. All other grid nodes exhibit a nonzero DOA error, because the color gradient inside their respective hexagonal cells is shifted toward the central axis, as it can be observed on maps of 1NN with 1-m and 0.5-m grid sizes. For example, from listening positions situated in the left half of the navigation area, the source is geometrically located on the right of the listener. The grid nodes from the left half of the navigation area exhibit a positive DOA error, meaning that from these positions, the source appears somewhat to the left of the reference source position, and vice versa for grid nodes from the right half of the navigation area.

ITD values were estimated on anechoic BRIRs obtained by binaural decoding of Ambisonic encoded impulses distributed in the horizontal plane for different Ambisonic orders. The encoding in Ambisonics of different orders was done using the *ambix_encoder* VST plugin from the AmbiX plugin suite [35], and the source position was controlled through the automation track of the *azimuth* parameter in Reaper. The binaural decoding was done using the MagLS method employed by the IEM BinauralDecoder. Resulting ITD values are shown in Fig. 4, together with the reference

Table 1. RT_{60}(s) as a function of octave bands and Direct-to-Reverberant ratio (DRR; dB) calculated on the W channel of all Ambisonic RIRs of the 0.5-m grid of the St. Elisabeth model for the two considered acoustic conditions.

	RT_{60} (s) / Octave band (Hz)						DRR (dB)		
Room condition	125	250	500	1,000	2,000	4,000	Avg.	Min.	Max.
Reverberant	3.0	3.1	3.2	3.2	3.0	2.6	3.3	−3.4	12.7
Damped	1.1	1.2	1.2	1.2	1.1	1.0	4.8	−2.6	15.1

ITD values estimated on the original HRIRs, that were already used for the generation of the mapping function mentioned in SEC. 4.1.

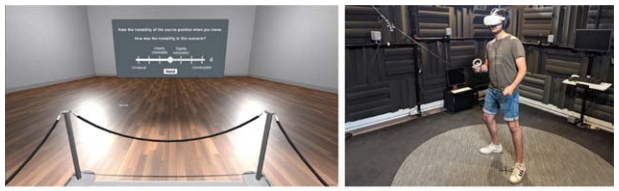
It can be observed that the ITD values estimated using the onset threshold detection method after applying a low-pass filter (3 kHz) on anechoic BRIRs issued from Ambisonic encoding undergo a compression effect in certain regions of space, which increases as the Ambisonic order decreases. Whereas fifth-order Ambisonics does not affect the estimated ITD much compared to the reference, for third-order Ambisonics, the ITD compression effect is significant in the frontal region between −55° and 55° and on the sides and is even more significant in all regions of space for first-order Ambisonics. This observation is consistent with the ITD degradation induced by MagLS binaural decoding of low-order Ambisonics reported in [36, Fig. 6], although the ITD values reported herein are significantly different from those reported in Fig. 4, probably resulting from a different ITD estimation method (MaxIACC$_e$). The impact of Ambisonic order on estimated ITD values observed in Fig. 4 was also reported in [37, Fig. 3a], evaluated using a different binaural decoding method.

Because a reduced ITD value leads to a DOA deviation toward the median plane, the ITD compression effect provoked by third-order Ambisonic encoding in the frontal region may explain why the DOA error is nonzero on RIR grid nodes for 1NN in Fig. 3. If fifth-order Ambisonic RIRs were used for the DOA error prediction, the color gradient inside individual hexagonal tiles for 1NN would probably be more centered on the RIR grid nodes.

5 PERCEPTUAL VALIDATION

A perceptual test was carried out to assess how perceived source stability was impacted by the considered interpolation parameters using the same audio rendering engine components, navigation area, and acoustic scenario (static omnidirectional source located on the central axis at a minimum distance of 1.5 m from the navigation area). The test examined the impact of the three proposed panning methods and the proposed grids of sizes 0.5, 1, and 2 m, as well as two room acoustic conditions to assess how a change in reverberation time would affect stability ratings.

The two room acoustics used in this study were generated using the same geometrical room model, but they were attributed different acoustic materials. The "reverberant" acoustic condition corresponds to the acoustics of the actual church, calibrated based on measurements performed in St. Elisabeth. The acoustic calibration was done following the calibration procedure published in [38]. The "damped" acoustic condition was generated from the same geometrical model, though more absorbent materials were applied. The RT_{60} reverberation time averaged over the listening zone and Direct-to-Reverberant ratios are shown in Table 1 for both considered room acoustics. These two conditions were constructed so that the compared RIRs have similar temporal structure and spatial characteristics, differing only in energy and reverberation time.



(a) (b)

Fig. 5. (a) Screenshot of the visual environment. (b) Experimental test setup.

5.1 Experimental Setup

The visual environment and user interface were developed in Unity and displayed in a Meta Quest 2 head-mounted display (HMD). The audio scene was rendered in parallel in Max/MSP and reproduced over Sennheiser HD 600 headphones. The listener's position and head orientation were tracked via the built-in cameras of the HMD and sent to the audio engine at a rate of 100 Hz via a local WiFi network using the Open Sound Control protocol. These data were logged to allow for further analysis of the navigation. Spatial interpolation and convolution were performed using the RoomZ plugin, Ambisonic rotation compensating for head orientation was done using the IEM SceneRotator plugin, and final binaural decoding was done using the IEM BinauralDecoder plugin. The input/output buffer length in Max/MSP was set to 1,024 samples at 48 kHz.

5.2 Evaluation Protocol

As shown in Fig. 5, the virtual visual environment included the 2 × 2 m^2 navigation area in the center of a virtual shoebox room. The room was kept empty, with realistic though relatively simple textures to minimize any impact the visuals might have on the perceived auditory scene. The navigation area was depicted by a carpet on the floor, enclosed by museum ropes at waist height attached to poles in each corner. The ropes aided the participants in understanding the extent of the navigation area without having to look at the floor.

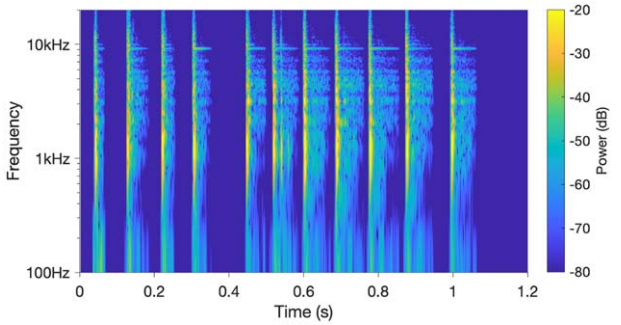


Fig. 6. Spectrogram of the first 1.2 s of the stimulus input signal, highlighting the sharp broadband attack of each impulsive sound.

The experiment took place in the acoustically dry MotionCapture/VirtualReality room at the Institute Jean Le Rond ∂'Alembert. Before the experiment, participants were briefed on the position of the nonvisual virtual auditory source. They were given explicit instructions encouraging them to explore the full extent of the navigation area during the experiment by walking all the way to the visual poles and ropes. The actual test was preceded by a tutorial session where they could train for the task and get familiar with the user interface. However, no reference trial condition with associated rating was presented to the participants.

Prior to any audio playback, the participants had to stand in the center of the navigation area to start the audio loop. This ensured that all participants had the same perceived reference source position before they started to navigate, and that the perceived reference source position before navigation was similar across conditions. The audio source was temporarily muted if participants left the navigation area to prevent unwanted auralization artifacts. Each trial consisted of two consecutive loops of 20 s of the same stimulus and condition. The stimulus used was that of a crank, composed of a nonperiodic sequence of percussive sounds, chosen to optimize localization ability [39]. The signal spectrogram is shown in Fig. 6. After the two repetitions, they rated the overall perceived instability of the source position during navigation by answering the following question using a seven-point Likert scale: "In this scenario, how would you judge the instability of the source position when you navigate?" The odd-numbered rating marks were labeled (1) "Unnatural," (3) "Clearly noticeable," (5) "Slightly noticeable," and (7) "Unnoticeable" in ascending order.

The conditions were randomized and repeated twice to gauge participants' repeatability and to compensate for a potential training effect. After the test, participants answered a questionnaire to evaluate their level of fatigue, level of self-confidence in their rating, and experience with such evaluation tests and to report other audio artifacts they might have perceived during the navigation.

5.3 Participants

A total of 22 paid subjects (18 males, 4 females) with an average age of 32.2 years participated in the experiment. A total of 32% of them had already participated in at least three sound localization tests, and as such, they are considered as expert listeners during the analysis. All participants stated having normal hearing abilities. The average duration of the experiment was 32.5 min, and about 85% of the participants reported at least a bit of fatigue after completing the test. The ratings of an extra 23rd participant were removed from the statistical analysis because of a low repeatability rate across repetitions of the same conditions.

6 RESULTS

Analyses of variances of participants' ratings were conducted to assess the effect of the different factors of panning method, room, grid size, critical listening expertise, and the first-order interaction terms between them. Statistical significance was determined for p values below a 0.05 threshold. The notation $p < \varepsilon$ is adopted to indicate p values below 10^{-3}. Post hoc pairwise comparisons for significant factors were made with Tukey-Kramer adjusted p values, or with Wilcoxon rank-sum p values for unbalanced comparisons.

6.1 Impact of the Panning Method and Grid Size

The panning method had a significant impact on participants' ratings ($F = 65.0$, $p < \varepsilon$). 1NN was rated overall significantly below $3NN_{dist}$ (3.2 vs. 4.6, $p < \varepsilon$), which was itself rated below $3NN_{area}$ (4.6 vs. 5.0, $p = 0.017$). Those ratings correspond to auditory source position instabilities rated on average as "clearly perceptible" for 1NN and "slightly perceptible" for both $3NN_{dist}$ and $3NN_{area}$.

The RIR grid size also had a significant impact on participants' ratings ($F = 101.9$, $p < \varepsilon$). Those overall significantly decreased with increasing grid size: 0.5-m grid size was rated as more stable than 1-m grid size (5.1 vs. 4.8, $p = 0.013$), which was itself rated as more stable than 2-m grid size (4.8 vs. 3.0, $p < \varepsilon$). Those ratings correspond to instabilities judged as "slightly perceptible" for 0.5-m and 1-m grid sizes and as "clearly perceptible" for 2-m grid size. No significant impact of the room condition or critical listening expertise was observed on participants' ratings.

6.2 Further Interactions

Analysis revealed a significant interaction between the grid size and panning method regarding participants' ratings ($F = 9.9$, $p < \varepsilon$), illustrated in Fig. 7. As expected, ratings overall increase across panning methods ($1NN < 3NN_{dist} < 3NN_{area}$) and with decreasing grid size (2 m < 1 m < 0.5 m). The decomposition, however, indicates that the difference observed between the 0.5-m and 1-m grid size conditions only held for 1NN panning condition (4.7 vs. 3.4, $p < \varepsilon$) and was nonsignificant otherwise. It also revealed that $3NN_{area}$ panning method was actually rated higher than $3NN_{dist}$ (4.1 vs. 3.4, $p < \varepsilon$) when using the 2-m grid size.

Interestingly, there was a significant difference between how self-reported critical listening experts and nonexperts rated the different panning methods ($F = 4.3$, $p = 0.014$), as seen in Fig. 8. $3NN_{area}$ was overall rated higher than $3NN_{dist}$ by the experts (5.6 vs. 4.9, $p = 0.007$), while the

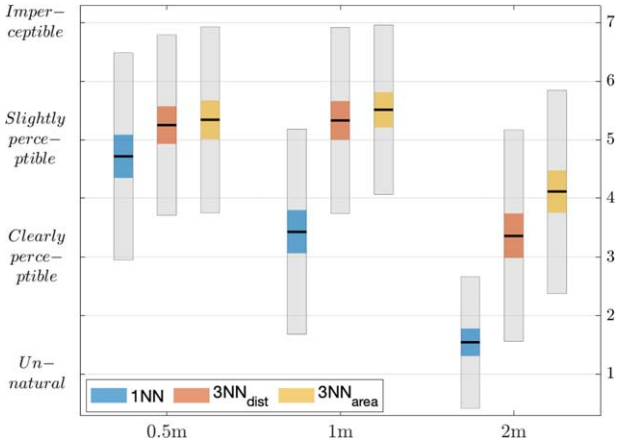


Fig. 7. Mean (—), 95% confidence intervals (darker area), and standard deviation (lighter gray area) of ratings of perceived source instability vs. grid size, aggregated over panning method. Boxes with nonoverlapping confidence intervals indicate that the difference between the associated results is statistically significant.

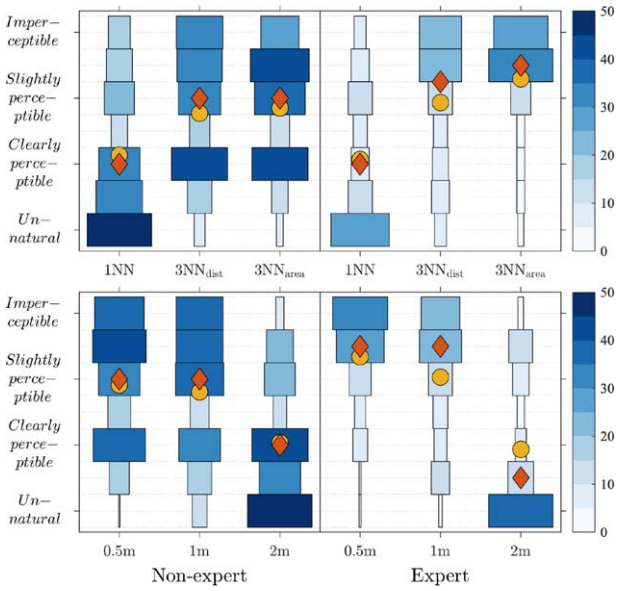


Fig. 8. Histogram distributions of participant ratings across (top) panning method and (bottom) grid size conditions for the (left) nonexpert and (right) expert groups. Diamonds and circles respectively indicate median and mean values.

nonexperts did not perceive any difference between these panning methods. Similarly, there was a significant interaction between expertise and how participants rated the various grid sizes ($F = 5.0$, $p = 0.007$). The added value of using a 0.5-m grid size compared to a 1-m grid size was only perceived by critical listening experts (5.7 vs. 5.1, $p = 0.014$). Moreover, averaged ratings for the 3NN$_{area}$ interpolation method and the 0.5-m grid size are higher for experts than for nonexperts. As seen in the histograms of Fig. 8, this is due to expert ratings being more in agreement with one another compared to nonexpert ratings distributed all over the Likert scale, suggesting that the former had a more uniform understanding of the task at hand.

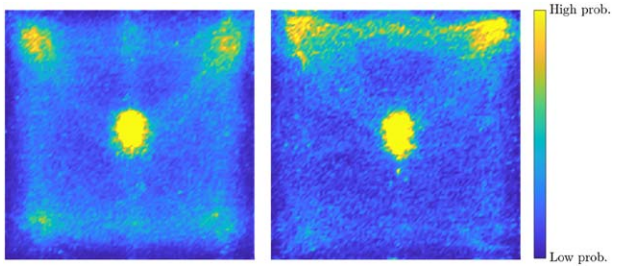


Fig. 9. Heat maps of the probability of presence of the participants in the navigation area, for (left) nonexperts and (right) experts. Darker color indicates a low probability, and lighter color indicates a high probability. The high probability observed in the center of the navigation zone is mainly due to the fact that participants had to start in the center, resulting in little significance regarding the navigation strategy.

6.3 Navigation Strategy

An analysis of participant positions within the navigation area during playback was carried out. This allowed for an assessment of whether or not some participants developed a navigation strategy to optimize the observation of source instabilities by exploring some regions of the navigation area more than others. Fig. 9 shows the probability of the presence of participants in the navigation area, for self-reported experts and nonexperts. It must be noted that the center of the navigation area exhibits a high probability of presence since this is where the participants had to stand to start the playback of each condition. Consequently, the high probability observed in the center of the navigation area has little significance regarding the navigation strategy.

It can be observed that the front half of the navigation area was overall more explored than the rear half. This preference is more pronounced for experts than for nonexperts. Such a strategy increases the chances of detecting source instabilities according to the DOA error prediction model presented in SEC. 4. Moreover, a high probability of presence for all participants was observed in the two front corners, which are expected to be critical zones for most conditions. Nonexperts explored indiscriminately the central axis and the edges of the navigation area, including the rear edge. In contrast, experts did not navigate much along the central axis and opted more for the front edge that connects the two front corners. Again, this strategy was probably developed to ease the detection of source instabilities occurring between the front corners of the navigation area.

7 DISCUSSION

As initially expected and predicted by the DOA error model, the 1NN method led to lower perceived source stability than the other tested methods, regardless of grid density. The 3NN$_{dist}$ and 3NN$_{area}$ methods performed similarly for the two highest grid densities, while the latter resulted in a more stable rendering for the lowest density (2-m grid size). Those observations suggest that if the reproduction

device can support any of the equally computationally intensive 3NN panning methods, $3NN_{area}$ is the best choice overall.

$3NN_{area}$ and $3NN_{dist}$ maintained their performance for a grid size below 1 m, meaning that in the given configuration, the perceived source stability did not benefit from grid densities higher than 1-m grid size threshold when using either of the panning methods. This result is consistent with the predicted DOA error. A similar plateau effect was observed for a different grid size threshold, for the $1NN_{BRIR}$ selection scheme used in [17].

The 1NN method used on a 0.5-m grid size led to a similar perceived stability as the other two methods used on a 2-m grid size. This suggests that for a reproduction device with limited computational power, the 1NN panning method could produce a comparable level of stability as the other two methods at the cost of a higher RIR grid density requirement, at least for the scenarios tested here. This subsequently entails a higher storage requirement on the reproduction device.

Most participants reported that they preferred navigation in the front half of the navigation area, i.e., closer to the auditory source, as it made source instability detection easier, as confirmed by the analysis of participant positions during navigation and observed on the DOA error model. This further confirms the self-translation MAA theory stating that source stability will increase as the source distance increases [19].

The position analysis further revealed that most participants explored much the two front corners, expected to be critical zones for the perceived source stability according to the DOA error model and that experts mostly explored the front edge of the navigation area joining the two front corners, which was expected to be a critical trajectory because it crosses high densities of iso-azimuth error contours as predicted by the DOA error model. Additionally, most participants reported that they were mostly looking toward the source when navigating, because it eased the detection of source instability. This could be related to the fact that the MAA is smaller in the frontal listening area than on the sides [29].

Fig. 10 compares the full-scale normalized ratings of the current study with those on perceived continuity reported in [17]. As a reminder, [17] used an interpolation scheme ($1NN_{BRIR}$) similar to the 1NN on BRIR grids of size 1 m and below. The figure shows that the current 1NN condition and $1NN_{BRIR}$ yield similar results on the two common grid sizes (0.5 and 1 m). In general, the 3NN methods outperformed $1NN_{BRIR}$ results on these two grid sizes, particularly $3NN_{area}$ with 1-m grid size. Moreover, despite a reduced RIR grid density, $3NN_{area}$ with the 2-m grid size yielded results comparable to $1NN_{BRIR}$ with the 0.5-m grid size. Because $3NN_{area}$ requires only one Ambisonic RIR at each individual grid node as opposed to $1NN_{BRIR}$, it offers a similar source stability combined with fluid listener head rotation even with RIR datasets of much reduced complexity, at the cost of an increased number of convolutions (three Ambisonic RIRs of 16 channels each vs. one BRIR of two channels).

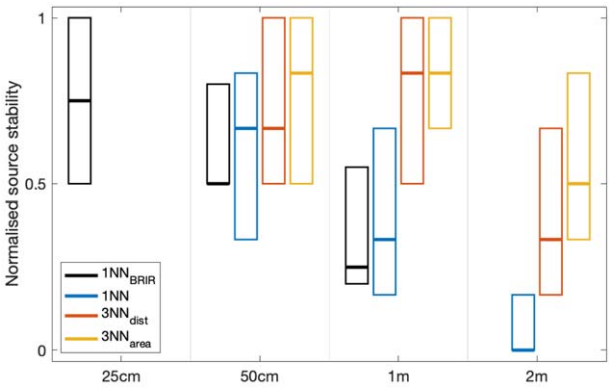


Fig. 10. Comparison between continuity ratings of [17] for the audio stimulus "Solo saxophone" and stability ratings in the current study as a function of grid size. Thick lines represent median values; rectangle boxes represent interquartile range. Likert scale ratings used in each study have been normalized between 0 and 1 to ease results comparison.

Looking forward, these results might serve as a starting point for studies on systems using different Ambisonic orders. It might be that as the order increases, participants will be more sensitive to shifts in spatial position because of increased spatial resolution. It is unclear how using a different Ambisonic decoding scheme will impact the results. However, previous literature suggests that other decoding schemes will result in less pronounced ITD compression [36], potentially leading to a source position perceived as more stable during listener head rotation. Additionally, given the task's focus on horizontal plane localization, using individual HRTF, or at least ITD-matched generic HRTF, could likely improve stability ratings.

Results suggest that reverberation duration does not impact stability ratings, at least in the range of 1.2–3.2 s for the rooms tested. It is possible that those results will not be generalizable to smaller rooms, where spatial perception becomes more geometry and specific surface material dependent due to higher prominence, and hence perceptual salience, of early reflections. Regarding source distance, because stability rating is related to the perceived angle to the source, it is likely that the constraint on grid size will go down for auralization scenarios based on more distant sources.

8 CONCLUSION

This study examined the impact of third-order Ambisonic RIR homogeneous grid spatial density and interpolation panning method on the stability of a nonvisual nearby static source with listener movements in the context of RIR convolution-based navigable auralizations. Rendering was achieved via MagLS binaural decoding and nonindividual HRTF. A DOA error model based on ITD estimation of the direct sound was proposed to predict the expected variation of the source position during navigation and was evaluated through a perceptual test where participants freely explored

an audiovisual scene rendered over an HMD and a pair of headphones.

Results of objective predictions and perceptual tests showed that perceived auditory source stability generally increased with increasing grid density, a result expected and in line with that reported by [17, 18], who used 1NN BRIR selection rather than the Ambisonic RIR panning tested here. Perceived stability reached a response plateau for the grid sizes of 1 m and below for all but the simplest 1NN panning method. This result is consistent with the similar plateau effect observed for $1NN_{BRIR}$ in [17], though with a lower grid size threshold of 10–25 cm depending on the nature of the stimuli. Results of the perceptual test also indicated that 1NN was systematically rated below the other two panning methods and that the $3NN_{area}$ method outperformed the $3NN_{dist}$ method for an Ambisonic RIR grid size of 2 m.

Self-reported expert listeners proved to further benefit from a higher grid density and, on average, preferred the $3NN_{area}$ over the $3NN_{dist}$ panning method. No significant impact of the room acoustics (different reverberation times, same geometry) on the perceived stability of the auditory source was observed.

In general, DOA error of the direct sound predicted by the ITD model was consistent with the ratings of the participants on source stability in echoic conditions, further confirming the lack of impact of the room acoustics as included in this study. The DOA error maps also revealed an ITD compression effect, probably due to the use of third-order Ambisonics and MagLS binaural decoding, resulting in an apparent shift of the source position toward the median plane.

Analysis of participant positions during navigation revealed that they developed a navigation strategy to facilitate the detection of source instabilities. Whereas they generally preferred the front half of the navigation area compared to the rear half, experts mostly focused on the front edge connecting the two front corners, which were expected to be the most critical regions for source stability according to the DOA error maps.

Future work could focus on the perceptual evaluation of other auditory attributes impacting the plausibility of the auralization, such as apparent source width or sound coloration, in similar conditions. Moreover, the impact of Ambisonic order and binaural decoding scheme on the perceived DOA of an auditory source warrants investigation. Further studies on multi-modal interactions, such as the impact of visuals on perceived stability, are necessary to better understand how to best deploy RIR-based auralizations in mixed-reality environments.

The results presented here serve as a guideline for the design of navigable auditory scenes used in general public applications such as immersive audio-guides. Such designs usually balance auditory scene quality against CPU load and rendering device storage capacity. On one hand, the design of high-density Ambisonic RIR grids requires a longer time for either simulations or measurements and a higher storage capacity on the rendering device. On the other hand, using a panning method that requires three Ambisonic RIR convolutions will be more CPU demanding and potentially impossible on some devices, compared to a simpler single Ambisonic RIR selection rendering method.

This dilemma was illustrated by the rating comparison between the 1NN method with a 0.5-m grid size and the $3NN_{area}$ method with a 2-m grid size. The 1NN method required one-third the CPU than the $3NN_{area}$. In contrast, the 0.5-m grid size required five times more storage than the 2-m grid size (39 Ambisonic RIRs vs. 7 Ambisonic RIRs with triangular grids) to cover the 2×2 m^2 navigation area used in this study.

9 ACKNOWLEDGMENT

The authors want to thank the participants who took part in the listening experiment. Funding has been provided by the European Union's Joint Programming Initiative on Cultural Heritage project PHE (The Past Has Ears; Grant No. 20-JPIC-0002-FS; phe.pasthasears.eu) and the French project PHEND (The Past Has Ears at Notre-Dame; Grant No. ANR-20-CE38-0014; phend.pasthasears.eu). Additional resources have been provided by the SONICOM project (www.sonicom.eu) that has received funding from the European Union's Horizon 2020 research and innovation program under Grant No. 101017743.

10 REFERENCES

[1] A. Zimmermann and A. Lorenz, "LISTEN: A User-Adaptive Audio-Augmented Museum Guide," *User Model. User-Adapt. Interact.*, vol. 18, no. 5, pp. 389–416 (2008 Nov.). https://doi.org/10.1007/s11257-008-9049-x.

[2] O. Delerue and O. Warusfel, "Mixage Mobile," in *Proceedings of the 18th Conference on l'Interaction Homme-Machine*, pp. 75–82 (Montreal, Canada) (2006 Apr.). https://doi.org/10.1145/1132736.1132746.

[3] V. Välimäki, J. D. Parker, L. Savioja, J. O. Smith, and J. S. Abel, "Fifty Years of Artificial Reverberation," *IEEE Trans. Audio Speech Lang. Process.*, vol. 20, no. 5, pp. 1421–1448 (2012 Jul.). https://doi.org/10.1109/TASL.2012.2189567.

[4] M. Gospodarek, O. Warusfel, P. Ripollés, and A. Roginska, "Methodology for Perceptual Evaluation of Plausibility With Self-Translation of the Listener," in *Proceedings of the AES International Audio for Virtual and Augmented Reality Conference* (2022 Aug.), paper 44. https://www.aes.org/e-lib/browse.cfm?elib=21874.

[5] D. Pelegrín-García, E. De Sena, T. van Waterschoot, M. Rychtáriková, and C. Glorieux, "Localization of a Virtual Wall by Means of Active Echolocation by Untrained Sighted Persons," *Appl. Acoust.*, vol. 139, pp. 82–92 (2018 Oct.). https://doi.org/10.1016/j.apacoust.2018.04.018.

[6] B. N. J. Postma and B. F. G. Katz, "Perceptive and Objective Evaluation of Calibrated Room Acoustic Simulation Auralizations," *J. Acoust. Soc. Am.*, vol. 140, no. 6, pp. 4326–4337 (2016 Dec.). https://doi.org/10.1121/1.4971422.

[7] J. Pätynen and T. Lokki, "Evaluation of Concert Hall Auralization With Virtual Symphony Orchestra," *Build. Acoust.*, vol. 18, no. 3-4, pp. 349–366 (2011 Dec.). https://doi.org/10.1260/1351-010X.18.3-4.349.

[8] O. Olgun, E. Erdem, and H. Hacıhabiboğlu, "Sound Field Interpolation via Sparse Plane Wave Decomposition for 6DoF Immersive Audio," in *Proceedings of the International Conference on Immersive and 3D Audio: From Architecture to Automotive (I3DA)*, pp. 126–135 (Bologna, Italy) (2023 Sep.). https://doi.org/10.1109/I3DA57090.2023.10319880.

[9] T. Deppisch, S. Amengual Garí, P. Calamia, and J. Ahrens, "Perceptual Evaluation of Spatial Room Impulse Response Extrapolation by Direct and Residual Subspace Decomposition," in *Proceedings of the AES International Conference on Audio for Virtual and Augmented Reality* (2022 Aug.), paper 14. http://www.aes.org/e-lib/browse.cfm?elib=21844.

[10] J. G. Tylka and E. Choueiri, "Soundfield Navigation Using an Array of Higher-Order Ambisonics Microphones," in *Proceedings of the AES International Conference on Audio for Virtual and Augmented Reality* (2016 Sep.), paper 4-2. https://www.aes.org/e-lib/browse.cfm?elib=18502.

[11] E. Patricio, "Toward Six Degrees of Freedom Audio Recording and Playback Using Multiple Ambisonics Sound Fields," presented at the *146th Convention Audio Engineering Society* (2019 Mar.), paper 10141. https://www.aes.org/e-lib/browse.cfm?elib=20274.

[12] N. Mariette, B. F. G. Katz, K. Boussetta, and O. Guillerminet, "SoundDelta: A Study of Audio Augmented Reality Using WiFi-Distributed Ambisonic Cell Rendering," presented at the *128th Convention Audio Engineering Society* (2010 May), paper 8123. https://www.aes.org/e-lib/browse.cfm?elib=15420.

[13] T. McKenzie, N. Meyer-Kahlen, R. Daugintis, et al., "Perceptually Informed Interpolation and Rendering of Spatial Room Impulse Responses for Room Transitions," in *Proceedings of the 24th International Congress on Acoustics*, paper 439 (Gyeongju, South Korea) (2022 Oct.). https://www.researchgate.net/publication/364829625_Perceptually_informed_interpolation_and_rendering_of_spatial_room_impulse_responses_for_room_transitions.

[14] G. Kearney, C. Masterson, S. Adams, and F. Boland, "Dynamic Time Warping for Acoustic Response Interpolation: Possibilities and Limitations," in *Proceedings of the 17th European Signal Processing Conference*, pp. 705–709 (Glasgow, UK) (2009 Aug.). https://ieeexplore.ieee.org/abstract/document/7077851.

[15] C. Masterson, G. Kearney, and F. Boland, "Acoustic Impulse Response Interpolation for Multichannel Systems Using Dynamic Time Warping," in *Proceedings of the 35th AES International Conference on Audio for Games* (2009 Feb.), paper 34. http://www.aes.org/e-lib/browse.cfm?elib=15188.

[16] K. Müller and F. Zotter, "Auralization Based on Multi-Perspective Ambisonic Room Impulse Responses," *Acta Acust.*, vol. 4, no. 6, paper 25 (2020 Nov.). https://doi.org/10.1051/aacus/2020024.

[17] A. Neidhardt and B. Reif, "Minimum BRIR Grid Resolution for Interactive Position Changes in Dynamic Binaural Synthesis," presented at the *148th Convention Audio Engineering Society* (2020 May), paper 10371. https://www.aes.org/e-lib/browse.cfm?elib=20788.

[18] S. Werner, F. Klein, and G. Götz, "Investigation on Spatial Auditory Perception Using Non-Uniform Spatial Distribution of Binaural Room Impulse Responses," in *Proceedings of the International Conference on Spatial Audio*, pp. 137–144 (Ilmenau, Germany) (2019 Sep.). https://doi.org/10.22032/dbt.39967.

[19] O. S. Rummukainen, S. J. Schlecht, and E. A. P. Habets, "Self-Translation Induced Minimum Audible Angle," *J. Acoust. Soc. Am.*, vol. 144, no. 4, pp. EL340–EL345 (2018 Oct.). https://doi.org/10.1121/1.5064957.

[20] D. Poirier-Quinot and B. F. G. Katz, "On the Improvement of Accommodation to Non-Individual HRTFs via VR Active Learning and Inclusion of a 3D Room Response," *Acta Acust.*, vol. 5, paper 25 (2021 Jun.). https://doi.org/10.1051/aacus/2021019.

[21] C. Schörkhuber, M. Zaunschirm, and R. Höldrich, "Binaural Rendering of Ambisonic Signals via Magnitude Least Squares," in *Proceedings of the German Annual Conference on Acoustics (DAGA)*, vol. 44, pp. 339–342 (Munich, Germany) (2018 Mar.). https://pub.dega-akustik.de/DAGA_2018/data/articles/000301.pdf.

[22] B. F. G. Katz and R. Nicol, "Binaural Spatial Reproduction," in N. Zacharov (Ed.), *Sensory Evaluation of Sound*, pp. 349–388 (CRC Press, Boca Raton, FL, 2018), 1st ed. https://doi.org/10.1201/9780429429422-11.

[23] E. K. Canfield-Dafilou and B. F. G. Katz, "Comparing Virtual Source Configurations for Pipe Organ Auralization," presented at the *157th Convention Audio Engineering Society* (2023 Oct.), paper 159. https://www.aes.org/e-lib/browse.cfm?elib=22313.

[24] L. McCormack, A. Politis, T. McKenzie, C. Hold, and V. Pulkki, "Object-Based Six-Degrees-of-Freedom Rendering of Sound Scenes Captured With Multiple Ambisonic Receivers," *J. Audio Eng. Soc.*, vol. 70, no. 5, pp. 355–372 (2022 May). https://doi.org/10.17743/jaes.2022.0010.

[25] K. Hormann and N. Sukumar (Eds.), *Generalized Barycentric Coordinates in Computer Graphics and Computational Mechanics* (CRC Press, Boca Raton, FL, 2017), 1st ed. https://doi.org/10.1201/9781315153452.

[26] V. Pulkki, "Virtual Sound Source Positioning Using Vector Base Amplitude Panning," *J. Audio Eng. Soc.*, vol. 45, pp. 456–466 (1997 Jun.). https://www.aes.org/e-lib/browse.cfm?elib=7853.

[27] D. Poirier-Quinot, P. Stitt, and B. F. G. Katz, "RoomZ: Spatial Panning Plugin for Dynamic Auralisations Based on RIR Convolution," in *Proceedings of the AES International Conference on Spatial and Immersive Audio* (2023 Aug.), paper 19. https://www.aes.org/e-lib/browse.cfm?elib=22199.

[28] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization* (MIT Press, Cam-

bridge, MA, 1997), revised ed. https://doi.org/10.7551/mitpress/6391.001.0001.

[29] A. W. Mills, "On the Minimum Audible Angle," *J. Acoust. Soc. Am.*, vol. 30, pp. 237–246 (1958 Apr.). https://doi.org/10.1121/1.1909553.

[30] D. R. Perrott and A. D. Musicant, "Minimum Auditory Movement Angle: Binaural Localization of Moving Sound Sources," *J. Acoust. Soc. Am.*, vol. 62, pp. 1463–1466 (1977 Dec.). https://doi.org/10.1121/1.381675.

[31] A. Andreopoulou and B. F. G. Katz, "Identification of Perceptually Relevant Methods of Inter-Aural Time Difference Estimation," *J. Acoust. Soc. Am.*, vol. 142, no. 2, pp. 588–598 (2017 Aug.). https://doi.org/10.1121/1.4996457.

[32] C. Pörschmann, J. M. Arend, and F. Brinkmann, "Directional Equalization of Sparse Head-Related Transfer Function Sets for Spatial Upsampling," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 27, no. 6, pp. 1060–1071 (2019 Jun.). https://doi.org/10.1109/TASLP.2019.2908057.

[33] B. Bernschütz, "A Spherical Far Field HRIR/HRTF Compilation of the Neumann KU 100," in *Proceedings of the 40th Italian Annual Conference on Acoustics and 39th German Annual Conference on Acoustics (AIA-DAGA)*, pp. 592–595 (Merano, Italy) (2013 Mar.). https://audiogroup.web.th-koeln.de/FILES/AIA-DAGA2013_HRIRs.pdf.

[34] P. Majdak, F. Zotter, F. Brinkmann, et al., "Spatially Oriented Format for Acoustics 2.1: Introduction and Recent Advances," *J. Audio Eng. Soc.*, vol. 70, no. 7/8, pp. 565–584 (2022 Jul.). https://doi.org/10.17743/jaes.2022.0026.

[35] M. Kronlachner, "Plug-In Suite for Mastering the Production and Playback in Surround Sound and Ambisonics," Gold Award at the AES Students Design Competition, Category 2 - Graduate Level, *136th Convention Audio Engineering Society* (2014 Apr.). https://api.semanticscholar.org/CorpusID:189811753.

[36] I. Engel, D. F. M. Goodman, and L. Picinali, "Improving Binaural Rendering With Bilateral Ambisonics and MagLS," in *Proceedings of the 47th German Annual Conference on Acoustics (DAGA)*, pp. 1608–1611 (Vienna, Austria) (2021 Aug.). https://pub.dega-akustik.de/DAGA_2021/data/articles/000533.pdf.

[37] I. Engel, D. F. M. Goodman, and L. Picinali, "Assessing HRTF Preprocessing Methods for Ambisonics Rendering Through Perceptual Models," *Acta Acust.*, vol. 6, paper 4 (2022 Jan.). https://doi.org/10.1051/aacus/2021055.

[38] B. N. J. Postma and B. F. G. Katz, "Creation and Calibration Method of Acoustical Models for Historic Virtual Reality Auralizations," *Virtual Real.*, vol. 19, pp. 161–180 (2015 Nov.). https://doi.org/10.1007/s10055-015-0275-3.

[29] D. R. Begault, E. M. Wenzel, and M. R. Anderson, "Direct Comparison of the Impact of Head Tracking, Reverberation, and Individualized Head-Related Transfer Functions on the Spatial Perception of a Virtual Speech Source," *J. Audio Eng. Soc.*, vol. 49, no. 10, pp. 904–916 (2001 Oct.). https://www.aes.org/e-lib/browse.cfm?elib=10175.

THE AUTHORS


Julien De Muynke


David Poirier-Quinot


Brian F. G. Katz

Julien De Muynke is a Ph.D. candidate in the field of audio virtual reality and heritage room acoustics at the Sorbonne Université/CNRS ∂'Alembert Institute, Paris. His Ph.D. is funded through the EU JPI-CH PHE project. Since 2015, he has also been a member of the spatial audio research group at Eurecat, Technology Center of Catalonia, Barcelona, Spain. His fields of interest include spatial audio perception, room acoustics, and virtual and augmented realities. He obtained his M.Eng. in signal processing and computer telecommunications in 2005 from the ENSEA graduate school of Electrical Engineering, France, and carried out his graduation project in the Acoustics Department of Aalborg University, Denmark. Before joining the ∂'Alembert Institute and Eurecat, he worked as an audio signal processing engineer in the consumer electronics industry sector.

•

David Poirier-Quinot is a researcher, presently focused on sound spatialization, perception, and room acoustics simulation for virtual and augmented realities. He studied these fields along with signal processing and computer sciences at the ∂'Alembert Institute, Imperial College London, IRCAM, LIMSI, and ETIS labs. With a background in mathematics, physics, and chemistry, he obtained an M.Eng. in signal processing and telecommunications from the ENSEA graduate school of Electrical Engineering, France, in 2011, and received a Ph.D. degree in acoustics, signal processing, and computer science from Sorbonne Université, Paris VI, France, in May 2015.

•

Brian F. G. Katz is a CNRS Research Director at the Sorbonne Université/CNRS ∂'Alembert Institute and coordinator of the Sound & Space research theme. His fields of interest include spatial 3D audio rendering and perception and room acoustics. With a background in physics and philosophy, he obtained his Ph.D. in Acoustics from Penn State in 1998 and his H.D.R. in Engineering Sciences from UPMC in 2011. Before joining CNRS, he worked for various acoustic consulting firms, including Artec Consultants Inc., ARUP & Partners, and Kahle Acoustics. He has also worked at LIMSI-CNRS and IRCAM.