J Multimodal User Interfaces (2016) 10:273–284
DOI 10.1007/s12193-016-0221-z

ORIGINAL PAPER

Auditory navigation with a tubular acoustic model for interactive distance cues and personalized head-related transfer functions

An auditory target-reaching task

Michele Geronazzo[1] · Federico Avanzini[1] · Federico Fontana[2]

Received: 9 November 2015 / Accepted: 4 May 2016 / Published online: 19 May 2016

Abstract This paper presents a novel spatial auditory display that combines a virtual environment based on a Digital Waveguide Mesh (DWM) model of a small tubular shape with a binaural rendering system with personalized head-related transfer functions (HRTFs) allowing interactive selection of absolute 3D spatial cues of direction as well as egocentric distance. The tube metaphor in particular minimizes loudness changes with distance, providing mainly direct-to-reverberant and spectral cues. The proposed display was assessed through a target-reaching task where participants explore a 2D virtual map with a pen tablet and hit a sound source (the target) using auditory information only; subjective time to hit and traveled distance were analyzed for three experiments. The first one aimed at assessing the proposed HRTF selection method for personalization and dimensionality of the reaching task, with particular attention to elevation perception; we showed that most subjects performed better when they had to reach a vertically unbounded (2D) rather then an elevated (3D) target. The second experiment analyzed interaction between the tube metaphor and HRTF showing a dominant effect of DWM model over binaural rendering. In the last experiment, participants using absolute distance cues from the tube model performed comparably well to when they could rely on more robust, although relative, intensity cues. These results suggest that participants made proficient use of both binaural and reverberation cues during the task, displayed as part of a coherent 3D sound model, in spite of the known complexity of use of both such cues. HRTF personalization was beneficial for participants who were able to perceive vertical dimension of a virtual sound. Further work is needed to add full physical consistency to the proposed auditory display.

Keywords Head-related transfer function · Auditory distance rendering · Digital waveguide mesh · Perceptual model individualization · Target-reaching task · Human spatial navigation · Auditory display

1 Introduction

Accurate acoustic rendering of sound source distance is a difficult task; auditory cues of egocentric distance have been shown to be essentially unreliable since they depend on several factors, which can be hardly kept under experimental control. Researchers have found psychophysical maps, usually in the form of perceived vs. real distance functions, showing a strong dependence on the experimental conditions [43]. Besides this dependence, a broad variability of the distance evaluations across subjects has been observed in most of the tests [40]; this variability is mainly explained by the level of familiarity with the sound source: the more unfamiliar a sound is, the more difficult is for a listener to disaggregate acoustic source information from the environmental cues.

In our research, we focus on *absolute* cues, i.e., those which are not a function of the source sound: loudness, direct-to-reverberant energy ratio, spectrum, and binaural differences when the source is nearby the listener's head. This approach has a threefold aim: (1) to preserve the sonic signature of the sound source, particularly its loudness, (2) to avoid cannibalization of otherwise informative additional cues, and (3) to maintain sufficient ecological consistency of

✉ Michele Geronazzo
geronazzo@dei.unipd.it

[1] Department of Information Engineering, University of Padova, via Gradenigo 6B, 35131 Padova, Italy

[2] Department of Mathematics, Computer Science and Physics, University of Udine, via delle Scienze 206, 33100 Udine, Italy

the auditory scene. Together, these three properties in principle allow the sound designer to make use of the resulting distance rendering tool regardless of the type of source sound employed with it, as well as to neglect potential interferences coming from concurrent sonification models running in parallel with the same tool, for instance in the context of an auditory interface displaying a rich dataset.

Digital Waveguide Mesh (DWM) models and similar computational schemes have been employed offline to render auditory distance cues [7,10], allowing for moving source and listener positions everywhere inside a 3D shape. Interactivity, however, requires to make a leap forward: the model needs to be computed in real time and must be robust against abrupt movements of the source and/or listening points. Nowadays machines are able to compute DWMs counting some thousand nodes in real time, hence ensuring interactive control of the corresponding virtual scene: based on this assumption, a DWM-inspired model has been used to enable interactive reverberation for computer game applications [8].

In this work we propose a real-time spatial sound rendering architecture that combines binaural (individualized, HRTF based) rendering with a virtual (non-individualized, DWM based) environment simulating a tubular shape. This choice is supported by an experiment on HRTFs with embedded distance cues [41], which showed that directional cues were highly individual whereas distance cues were not. Hence, by decoupling the rendering of directional and distance cues, we expect that environmental effects simulated through the DWM model can improve listeners' performance in distance estimation, while preserving their ability to estimate direction, as HRTF-related cues should not be degraded or distorted by this simplified environment.

The technical features of both binaural rendering and the DWM model are illustrated in Sect. 3. Section 4 describes an experimental task aimed at assessing the validity of the proposed approach using different rendering strategies: it is a target-reaching task, in which subjects have to explore a 2D virtual map through a stylus on a tablet, and to hit an elevated sound source (the target) in the map using auditory information. The experimental scenario describes an egocentric view of the virtual map in which the pointer corresponds to the listener's head, and follows the "ears in hand" ecological metaphor [26]. Experimental results are presented in Sect. 5 and discussed in Sect. 6; they show that participants who were able to exploit 3D, HRTF-personalized display and absolute distance cues, achieved a first level of spatial knowledge [39] by performing comparably to (1) when they reached a 2D (i.e., vertically unbounded) instead of 3D (i.e., bounded and vertically offset) target, and (2) when they relied on relative (i.e., intensity) instead of absolute (i.e., direct-to-reverberant energy and spectral) cues of distance.

These two results are particularly interesting, considered the known unreliability of the monaural cues of elevation [4] as well as the complexity of the absolute cues of distance. Taken together, they suggest that the perceptual impact of otherwise less informative cues of space may become significant if the auditory display reproduces such cues as part of an experience which is sufficiently natural and valid in an ecological sense.

2 Related works

The ambiguity about the origin (either source- or environment-based) of the auditory cues related to distance makes the perception of a moving sound source especially interesting to investigate: by listening to dynamic cues humans receive a range of psychophysical information about the source sound in relation to its continuous modifications due to the environment. However, literature on distance recognition experiments involving moving sound sources is sparse and mostly limited to virtual acoustic setups; furthermore, due to the complexity of the dynamic rendering models this literature mixes psychological issues with arguments of sound processing: Lu et al. describe a model capable of rendering motion parallax and acoustic τ (time-to-contact), already identified by Spiegle and Loomis as two salient cues for the positional recognition in a moving listener and source scenario [25,36]. Moreover, moving sources can evoke so-called "looming" effects causing localization bias, such as when a tonal stimulus is displayed by a loudspeaker approaching the listener [31].

Moreover, near-field distance has been sonified using auditory metaphors, too [32]: by rendering robust effects (such as the repetition rate of a beep) that are disjoint with the sound source properties, this approach may produce reliable distance estimations as soon as listeners learn and get used to the proposed sonification.

When rendering is not limited to nearby sources, direct-to-reverberant energy ratio and spectrum form a typical pair of absolute distance cues. The former has been shown to provide significant, although coarse coding of distance [42]; the latter introduces audible changes in the sound "color", with association of increased high-frequency content to closer source positions. More generally, these cues impact spatial auditory perception in two respects: while a listener's ability in perceiving distance is enhanced, the ability in perceiving sound source direction is degraded in a complementary fashion [34]. This is due to the fact that reverberation corrupts and distorts directional cues, regarded as both binaural cues along azimuth (especially interaural time differences) and monaural cues along elevation (pinna reflections and resonances). The degradation in localization performance is particularly evident when the environment is unknown to the listener.

Direct-to-reverberant energy ratio and spectral cues together are effective even in uncommon/unrealistic environments. In an experiment where a loudspeaker could be moved inside a long, narrow pipe, listeners were able to build a consistent psychophysical map of distance in absence of loudness changes [11]; this map was in good accordance with the prediction model proposed by Bronkhorst and Houtgast [6], although compressed and non-linear. Later experiments made use of virtual rather than real environments, using distributed computational models, and extended the tubular model to other simple 3D shapes, such as cones and pyramids, in an effort to identify a shape capable of evoking psychophysical maps with a good degree of linearity [9].

In spite of its unreliability and subjective dependency, perception of egocentric distance remains highly interesting for auditory display purposes as an informative dimension having immediate physical interpretation and, hence, strong ecological meaning. Zahorik suggested design guidelines that are of great help for realizing accurate auditory displays, given specific technological constraints [41]. Designing similar guidelines in the case of moving sources is an even more challenging task, and still a matter of discussion.

3 3D sound rendering

Spatial audio technologies through headphones usually involve binaural room impulse responses (BRIRs) to render a sound source in space. BRIR can be split in two separate components: room impulse response (RIR), which defines room acoustic properties, and head related impulse response (HRIR), which acoustically describes the individual contributions of listener's head, pinna, torso and shoulders. In this paper, the latter acoustic contribution was implemented through an HRTF selection technique based on listener anthropometry, while virtual room acoustic properties and distance cues were delivered through an acoustic tube metaphor.

3.1 HRTF-based spatialization

Recording individual HRIRs/HRTFs is both time- and resource-consuming, and binaural audio technologies usually employ non optimal choice of a pre-defined HRTF set (e.g., recorded on a dummy head, such as the KEMAR mannequin [13]) for any possible listener. However, individual anthropometric features heavily affect the perception and the quality of the rendering [35]. Accordingly, advanced HRTF selection techniques aim at providing a listener with his/her "best matching" HRTF set extracted from a HRTF database, based on objective or subjective criteria [21,23].

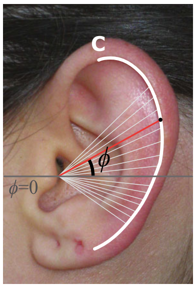

Fig. 1 HRTF selection based on participant's pinna external contour (C), manually extracted within Matlab

In this paper, an image-based HRTF selection technique is exploited (see [19] for details) where relevant individual anthropometric features are extracted from one image of the user's pinna. Specifically, a mismatch function is defined between the main pinna contours and corresponding spectral features (frequency notches) of the HRTFs in the database, according to a ray-tracing interpretation of notch generation [35]. The first notch of the HRTF can be predicted by calculating the distances between a point located approximately at the ear canal entrance and the corresponding reflection point at the border of the helix (the C contour in Fig. 1).

For a given elevation ϕ of the incoming sound, the reflection distance can be computed as $d(\phi) = ct(\phi)$, where $t(\phi)$ is the temporal delay between the direct and reflected rays and c is the speed of sound. The corresponding notch frequencies are estimated as $f_0(\phi) = \frac{c}{2d(\phi)}$, according to the assumption of negative reflection coefficient and one-to-one reflection-notch correspondence [35]. Given a user whose individual HRTFs are not available, the mismatch m between f_0 notch frequencies estimated from the last equation and the notch frequencies F_0 of an arbitrary HRTF set is defined as $m = \frac{1}{|\phi|}\sum_{\phi}\frac{|f_0(\phi)-F_0(\phi)|}{F_0(\phi)}$, where elevation ϕ spans all the available frontal angles for available HRTFs. Finally, the HRTF set that minimizes m is selected as the best-HRTF set in the database for that user.

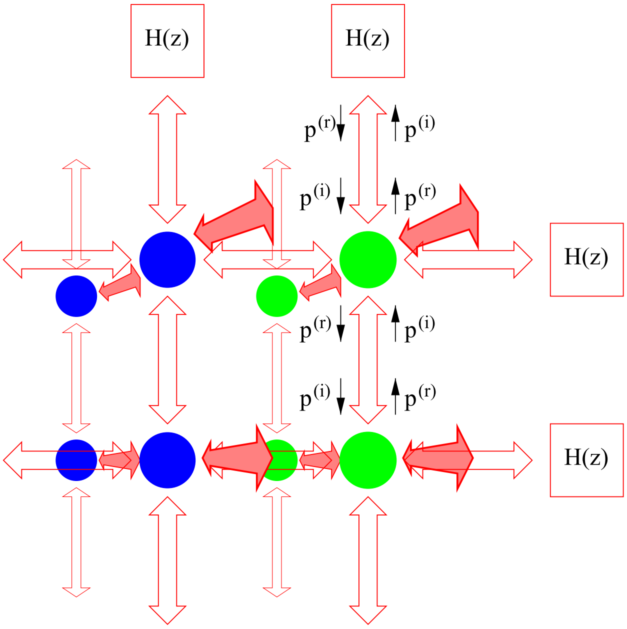

Fig. 2 Detail of the 3D DWM: scattering junctions and boundary filters

3.2 Digital waveguide mesh model

The DWM model used in this work simulates a small 3D tubular cavity with square cross-section. Scattering junctions forming the mesh boundary are coupled with filters modeling frequency-dependent wall absorption [20]. Figure 2 shows a detail of this design, exposing scattering junctions and boundary filters exchanging pressure wave signals each with its adjacent nodes (either junctions or filters). The mesh counts $29 \times 5 \times 5 = 725$ junctions, of which $5 \times 5 = 25$ form either termination of the tube whereas $29 \times 5 = 145$ form each of the four tube surfaces. One termination was modeled as an open end (i.e. $H(z) = -1$) whereas the other one was modeled as a closed end (i.e. $H(z) = 1$). Finally, each surface was modeled as an absorbing wall with larger absorption toward the high frequencies, by realizing $H(z)$ as a 1st-order IIR low-pass filter.

Once running at 44.1 kHz, the DWM simulates a tiny tubular environment measuring about 16 × 3 × 3 cm. The distance rendering effect depends on the relative positions of the source and listening point, i.e. junctions in which the audio signal was injected and picked up. In the current implementation, both are located on the main axis of the tube. The listening point is placed close to the open end, while the source can be moved along the main axis starting from nearby the closed end. Holding the listening point avoids picking up wave discontinuities otherwise caused by its movement at runtime. However, the effects of a similar artifact propagate to any listening position in the DWM if the source point is moved across junctions during the simulation. This artifact was minimized by linearly de-interpolating the sound, at every temporal step, across two junctions neighboring the moving source point [12].

The DWM models time-frequency characteristics as in Fig. 3, displaying spectrograms referenced at 30 dB of the tube during the first 70 ms for five equally-spaced normalized distances. Resonant modes are visible until about 3 kHz. Their moderate time-stretching with increasing distance testify a progressive decrease of the direct-to-reverberant energy ratio—note the enlargement of the bright areas and proportional reduction of the dark regions. The increasing delay with distance of the direct signal in reaching the listening point results in a slight shift rightward of the peak magnitudes (in pale yellow and white) of all modes.

The notch around 11 kHz is caused by spectral mirroring at half the Nyquist frequency. The spectral distortion causing the mirror modes to compress near the Nyquist frequency is due to traveling wave dispersion. Both such artifacts are typical of the DWM and introduced audible color in our stimuli, however in frequencies where headphones cannot be equalized [5]. For this reason, these artifacts distorted directional cues that were already out of control under all experimental conditions. Low-passing the stimuli would have removed the DWM artifacts; on the other hand, this choice would have required to low-pass also distance cues which were free from any distortion, since rendered through loudness changes. By considering also that the majority of artificial reverberation models introduce color in the sound [37], we chose to maintain the high frequency content in the stimuli for preserving the occasionally distorted, however broadband sound of the source.

By varying the direct-to-reverberant energy ratio, alternative reverberation models to DWM would have served the same purpose, improving upon the performance of the proposed method. However, we aimed at a scalable realization of an interactive reverberator having physical meaning. While recognizing that the current tube is too small, we expect to enlarge and shape the virtual listening environment in the near future.

4 Experiments

The overall goal of the experiments was to assess the validity of the proposed rendering metaphors, the "ears in hand" metaphor for direction and the "acoustic tube" metaphor for distance. Secondly, to analyze the differences and complementarity of the resulting auditory information by means of

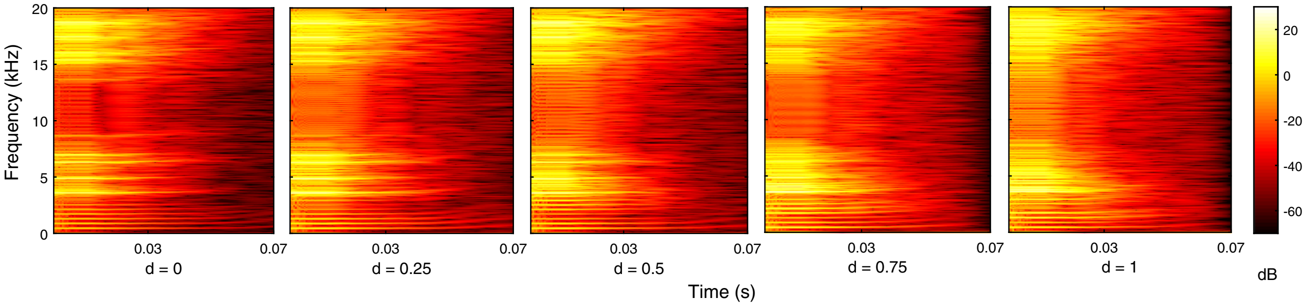

Fig. 3 Spectrograms showing the DWM magnitude response in dB for five distances. Normalized distance values increasing from the *left* to the *right*

behavioral and performance indicators collected from experimental data.

4.1 Rationale

These general goals were obtained through a target-reaching task, in which participants had to hit a virtual sound source, rendered through headphones displaying the target's relative position inside a workspace physically consisting of a pen tablet (Fig. 4). Three experiments were conducted using the same task and setup, but with different auditory feedback conditions.

4.1.1 Experiment #1

This experiment focused on directional cues only (based on different HRTF rendering methods), and no distance rendering. The goal was assess the effects of personalization and dimensionality of directional rendering, and to have a benchmark about the effects of directional rendering, in order to compare the effects of distance rendering in subsequent tests.

Two HRTF-based directional rendering methods were considered: (1) generic, and (2) personalized HRTF rendering. The dimensionality could be set to 3D or downscaled to 2D. The combination of rendering method and dimensionality resulted in four experimental conditions, summarized here along with their acronyms:

1. generic HRTF directional cues in 2D (2Dgen);
2. personalized HRTF directional cues in 2D (2Dpers);
3. generic HRTF directional cues in 3D (3Dgen);
4. personalized HRTF directional cues in 3D (3Dpers).

These conditions are listed in increasing order of auditory information, in terms of dimensionality (2D/3D) and personalization (generic/personalized).

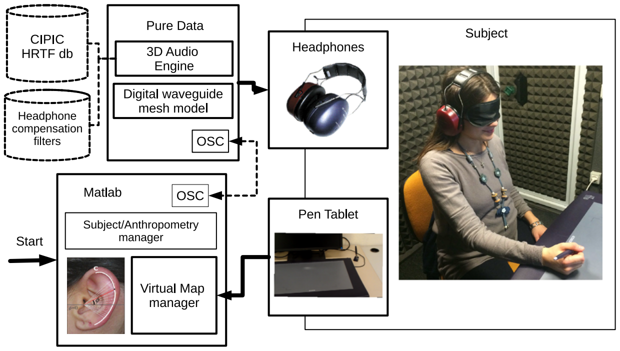

Fig. 4 System architecture for the experimental setup. Pure Data synthesizes HRTFs from the database and DWM reverberation cues at runtime. Matlab interacts with Pure Data via OSC protocol, selecting HRTFs and sending pen coordinates recorded by the tablet

4.1.2 Experiment #2

This experiment dealt with the interaction between tubular acoustics and directional cues. Specifically, three different directional rendering methods were tested: (1) intensity panning, (2) generic HRTF rendering, and (3) personalized HRTF rendering. The dimensionality could be set to 3D or downscaled to 2D, resulting in five conditions, summarized here along with their acronyms:

1. Tube and intensity panning (DWM+2Dpan);
2. Tube and generic 2D HRTFs (DWM+2Dgen);
3. Tube and personalized 2D HRTFs (DWM+2Dpers);
4. Tube and generic 3D HRTFs (DWM+3Dgen);
5. Tube and personalized 3D HRTFs (DWM+3Dpers).

In particular, DWM+2Dpan played the role of a control condition providing simple angular cues of intensity, which do not interact with the spectral cues originating from the tube.

4.1.3 Experiment #3

This final experiment focused on different combinations of distance and directional cues. In this case, the goal was to compare two different approaches for distance rendering: a 6-dB law modeling open-space loudness attenuation with distance, and the tubular model. Directional rendering was enabled by the 3Dpers condition only. The combination of direction and distance rendering resulted in five experimental conditions, summarized here along with their acronyms:

1. personalized HRTFs only (3Dpers);
2. 6-dB law only (L);
3. tube model only (DWM);
4. tube and personalized HRTFs (DWM+3Dpers);
5. 6-dB law and personalized HRTFs (L+3Dpers).

Conditions 3Dpers, L and L+3Dpers were used for control purposes. In particular, 3Dpers provided only directional cues, L provided only intensity cues, and the combination of L+3Dpers played the role of "ground truth", i.e., possibly most robust feedback condition.

4.2 Apparatus

Figure 4 depicts a schematic view of the system architecture. All tests were performed using Matlab, which also recorded the 2D position on the pen tablet, a 12 × 18 in (standard A3 size) Wacom Intuos2 connected via USB to the computer. Spatial audio rendering was realized in Pure Data. Open Sound Control (OSC) protocol managed communi-

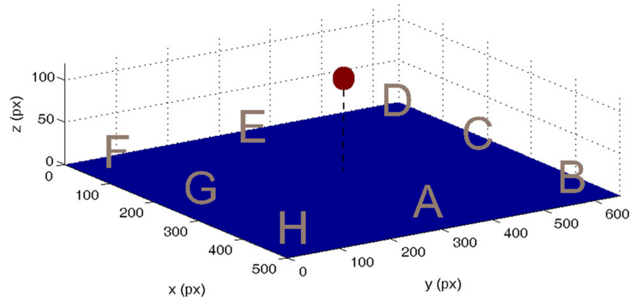

Fig. 5 The virtual map in pixels. The target is the central *red* sphere; relative starting positions for audio exploration are marked in lexicographic order

cation between Matlab and Pure Data. The overall system latency was 27 ms.[1]

Audio output was operated by a Roland Edirol Audio Capture UA-101 board working at 44.1 kHz sampling rate, and delivered to a pair of Sennheiser HDA 200 headphones. These headphones provide effective passive ambient noise attenuation, have an almost flat frequency response in the range 0.1–10 kHz and are largely insensitive to accidental movements around a users' head [14]. Headphone equalization filters were designed based on measurements made with the KEMAR with its pinnae unmounted, and then applied to the auditory stimuli. Although non-individual, this compensation strategy guaranteed no corruption of the localization cues contained in the HRTFs [28], as well as effective equalization of the headphones up to approximately 8–10 kHz. In this type of auditory experiments, in fact, using individualized headphone equalization filters can introduce dependencies on the headphone position around the head, making them less recommended than a generalized equalizer compensating the frequency response of a stable headphone [14].

4.3 Stimuli

The virtual target sound was placed at the center of the 640 × 480 pixels working area. It had the form of a sphere with radius equal to 25 pixels, placed at a height of 120 pixels (see Fig. 5). The 3D-position of the user (pen) was spatially rendered relative to the target. User movements were limited to the horizontal plane (the tablet), whereas the egocentric view had a fixed height of 60 pixels from the ground. If the pen was moved beyond the boundaries of the working area then the system signalled the illegal position, by playing white noise until a correct position was restored.[2]

The source sound consisted of a camera click with 100 ms duration repeated every 300 ms [15], with maximum amplitude level at the entrance of the ear canal amounting to 62 dB(A) for experiment #1, 60 dB(A) for experiment #2 and 65 dB(A) for experiment #3, respectively. The interval between subsequent clicks was large enough to include reverberant tails introduced by the tubular environment.

Regarding directional cues, the CIPIC database [1] was chosen as a source of HRTFs. It contains 45 HRTF sets measured in the far field and free-field compensated, hence free of distance information, with azimuth and elevation angles spanning the ranges $[0^\circ, 360^\circ)$ and $[-45^\circ, 230.625^\circ]$, respectively. Directional cues in 2D were generated by locking the elevation angle to 0° hence forcing the rendering model to span the horizontal plane only.

Generic HRTF directional cues (2Dgen, 3Dgen) were generated using CIPIC subject no. 165 (KEMAR with large pinnae), yielding a template HRTF for all participants. Personalized cues (2Dpers, 3Dpers) were generated using the procedure described in Sect. 3 to select the best-matched HRTF set among 45 CIPIC subjects, for each participant. Accordingly, one pinna image of each participant was required for computing the mismatch between manually traced contours and notch central frequencies. Finally, directional cues based on intensity panning (2Dpan, 3Dpan) were generated using the panning law $G_{l,r} = \frac{1}{2}\left(1 \pm \cos(\theta + 90^\circ)\right)$, $G_{l,r} \in [0, 1]$, in which $\theta\ in\ [0^\circ, 360^\circ)$ is the azimuth between source and listener in the horizontal plane. This law leads in particular to positions: $\theta = 0^\circ/180^\circ$, corresponding to in-axis position with the sound source ($G_{l,r} = 1/2$), and $\theta = \pm 90^\circ$ respectively denoting lateral sources on the left ($G_{l,r} = 1$) and right ($G_{l,r} = 0$) side.

Distance cues (DWM, L) were rendered on top of the directional cues using either the tubular model described in Sect. 3.2 or a 6-dB law decreasing the sound loudness for every distance doubling. Figure 6 depicts, for all conditions of Experiment #3, average amplitudes measured as a function of normalized egocentric distance. To make the comparison easier, all amplitudes were equalized to a 0-dB reference value at null distance. The corresponding amplitude offsets are reported in Table 1. It can be noted that intensity in DWM and DWM+3Dpers conditions changed when the virtual source approached the auditory target, but not when it moved in the far-field due to the progressive stabilization of the direct-to-reverberant energy ratio, which is especially evident from 0.8 normalized distance units on. Furthermore, DWM+3Dpers produced higher intensity val-

[1] System latency was measured by means of two condenser microphones connected to an audio card working at 48 kHz sampling rate; microphones were placed at the headphone pad and on top of the pen tablet, respectively. Latency was estimated as the time difference between these two events: (1) pen impact on the tablet and (2) changing in audio output at the headphones.

[2] The geometrical properties of the virtual map were chosen in ways to ensure detectable elevation cues from the HRTF selection procedure (see Sect. 3.1).

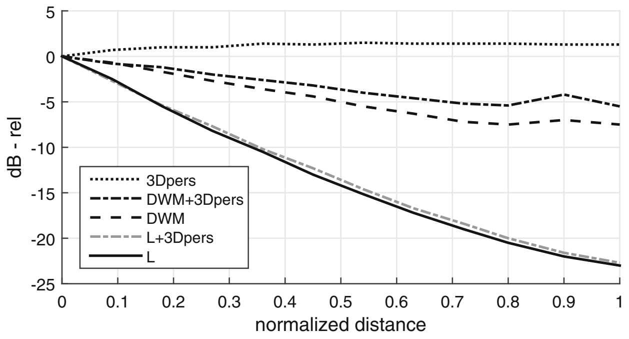

Fig. 6 Average amplitude of the stimuli used in the experimental conditions, as a function of normalized distance. Amplitude values ranging from the smallest to the largest (corresponding to position "A" in Fig. 5) egocentric distance

Table 1 Amplitudes in dB RMS measured at the smallest egocentric distance, for each auditory condition

	3Dpers	L	DWM	DWM+3Dpers	L+3Dpers
Amplitude (dB RMS)	65	60	72	78	65

Measurements for 3Dpers had KEMAR HRTFs [13] as reference

ues than DWM alone, showing an interaction between HRTF resonances and the tubular model. Finally, intensity in condition 3Dpers slightly decreased in the proximity of the target, that is, where the virtual listener position was below the target and, thus, pinna resonance around 12 KHz decreases [3].

4.4 Procedure

A brief tutorial session introduced the experiment. Participants were verbally informed that they had to explore a virtual map using only auditory information, and they had to be blindfolded during the experiment. Participants were then instructed that their goal was to move towards an auditory target as closely and quickly as possible, while only information regarding "ears in hand" exploration metaphor and no information regarding localization cues were provided. Each trial was completed when a participant was able to stand for at least 1.2 s within a 25-pixel neighborhood around the auditory target, similarly to the protocol in [17].

In order to minimize proprioceptive memory coming from the posture of the arm and the hand grasping the pen, the starting position was set to be always different across trials. Before each trial began, the experimenter lifted and moved the pen to a random position within the tablet area, and then helped the participant to grasp it again. Participants were asked to complete the task at several unknown different locations, corresponding to relative starting positions at the boundary of the workspace, depicted in Fig. 5.

Every condition was repeated for each virtual starting position. Starting position and auditory conditions were randomly balanced across trials. Due to the fast-screening nature of experiment #1 only four starting positions, namely B, D, F, and H in Fig. 5, were considered yielding to a total of 16 randomized trials (4 positions × 4 conditions). All positions were used for experiments #2 and #3, yielding 40 trials per participant (8 positions × 5 conditions).

4.5 Data collection and analysis

Three main performance indicators were used:

- **Mt** absolute reaching time: the time spent by the participant to complete the trial;
- **Md** total traveled distance: the length of the trial trajectory;
- **Mdf** final traveled distance: the length of the trial trajectory in the last 240 ms of exploration.

Participants' trajectories had large variability, and **Mt** with **Md** were thus assumed to be appropriate global indicators of performance. Moreover, **Mdf** was added as a third indicator, as it was assumed to be related to one's confidence in being nearby the target [17,39].

Preliminary analysis of gaussianity was performed on each condition by means of a Shapiro-Wilk test for normality, which revealed violations in some distributions. Each distribution exhibited skewness towards a physical constraint, i.e. the minimum possible traveled distance; accordingly, a logarithmic transformation was applied to data distributions which were subsequently subjected to Levene's test for homoscedasticity. Since the proposed experiments followed a one-factor within-subject design, if normality was not violated, within-subject ANOVAs with four/five levels of feedback condition were performed on **Mt**, **Md**, and **Mdf** as dependent variables. Pairwise *post-hoc* t-tests for paired samples with Holm-Bonferroni correction procedure on p-values provided statistical significances in performance between auditory conditions. On the other hand, if normality was violated, Kruskal-Wallis nonparametric one-way ANOVAs with four/five levels of feedback condition were performed to assess the statistical significance of the proposed indicators. Pairwise *post-hoc* Wil-coxon tests for paired samples with Holm-Bonferroni correction procedures on p-values provided statistical significances in performance between conditions. Finally, following a split-plot design, within-subject ANOVAs with four/five levels of feedback condition and a between factor were performed on **Mt**, **Md**, and **Mdf** to assess interaction between conditions and groups.

For the sake of simplicity, in the next sections we report adjusted p-values for *post-hoc* tests and adjusted p-values of within-subject ANOVAs which were corrected with

Greenhouse-Geisser procedure according to Mauchly's test for sphericity.

In addition to quantitative measures, at the end of each experiment participants answered the following four questions in a short post-test questionnaire concerning their self-reported experimental behaviour:

- Q1 : Adopted navigation strategies.
- Q2 : Elements of help in the navigation.
- Q3 : Elements of difficulty in the navigation.
- Q4 : Did you perceive elevation?

In particular, the yes/no answers to Q4 of Experiment #1 were used to create two groups for a split-plot design. Such a split turned useful in Experiment #2, to analyse the interaction between the tubular acoustics and the proposed personalization method.

5 Results

Nine participants (7 males and 2 females, mean age 27.±5.7) took part in the first experiment. Ten (8 males and 2 females, mean age 30.8 ± 8.7) took part in the second experiment. Eleven (8 males and 3 females, age ranging 26 to 41, mean 28.5 ± 5.2) took part in the third experiment. Eight participants took part to all experiments. All participants reported normal hearing.

5.1 Experiment #1

A within-subject ANOVA shows significance of the indicator **Mt** [F(3,24) =10.26, $p \ll 0.01$, $\eta^2 = 0.35$]. Pairwise *post-hoc* t-tests in Fig. 7a, above, report significantly different reaching times between certain 2D and 3D conditions. Similarly, a within-subject ANOVA shows significance of **Md** [F(3,24) = 6.08, $p < 0.05$, $\eta^2 = 0.21$]. Pairwise *post-hoc* t-tests in Fig. 7a, below, reveal significantly longer traveled distances. Together these results suggest that the task was carried out more efficiently in the 2D rather than the 3D domain. One particular difference between generalized and personalized HRTFs emerged from the traveled distance: 3Dgen results in worse navigation than both 2D conditions.

The data for **Mdf** were not normally distributed. A Kruskal-Wallis nonparametric one-way ANOVA did not show any significance of this indicator [$\chi^2(3)$ = 3.41, p = 0.335]; accordingly, no results are reported.

Concerning the answers to Q4 of the final questionnaire, four participants reported they did not perceive elevation cues, conversely five participants said they clearly distinguished the third dimension even if three of them were naïve to psycho-acoustic tests using binaural audio. Based on this questionnaire, two within-subject ANOVAs with a between factor (elevation perception Yes/No) were performed on **Mt** and **Md**. Condition effects were significant for reaching time [F(3,21) = 17.44, $p \ll 0.001$, $\eta^2 = 0.42$], and the group interaction was also significant [F(3,21) = 6.15, $p < 0.001$, $\eta^2 = 0.20$]. Moreover, condition effects were significant for traveled distance [F(3,21) = 8.29, $p < 0.01$, $\eta^2 = 0.22$], and the group interaction was also significant [F(3,21) = 4.23, $p < 0.05$, $\eta^2 = 0.13$]. The group analysis revealed a significant difference in navigation performance between participants. In fact, those who perceived the elevation also traveled faster to the target w.r.t. condition and group—see "condition: 3Dpers" and "elevation perception: Yes" in Fig. 7a.

5.2 Experiment #2

Since Experiment #1 suggested the existence of two groups depending on the subjective perception of elevation, a within-subject ANOVA with four levels of feedback condition and a between factor (elevation perception No/Yes) was performed to assess the significance of **Mt**. The condition was significant for reaching time [F(4,32) = 12.04, $p \ll 0.001$, $\eta^2 = 0.18$], and the group interaction was also significant [F(4,32) = 3.33, $p < 0.05$, $\eta^2 = 0.06$]. Pairwise *post-hoc* t-tests are shown in Fig. 7b, above, overall revealing that DWM+2Dgen was able to provide reliable and sufficient cues compared to personalized auditory conditions in 3D space. No significant statistical effects were found in the remaining pairs ($p > 0.05$ in any case).

Similarly, a within-subject ANOVA with one between factor (elevation perception Yes/No) was performed to asses the significance of **Md**. There was no condition effect for traveled distance [F(4,32) = 2.03, $p = 0.127$, $\eta^2 = 0.06$], however group interaction was significant [F(4,32) = 3.13, $p < 0.05$, $\eta^2 = 0.09$]. Figure 7b (bottom) depicts global statistics for traveled distance among conditions and groups, qualitatively suggesting that participants who were able to perceive elevation performed better under the condition DWM+2Dgen. This result may appear counterintuitive, and will be discussed in Sect. 6.

One further analysis was performed on **Mdf**, to assess the participants' awareness of being in proximity of the target. A within-subject ANOVA with one between factor (elevation perception Yes/No) was performed to assess the statistical significance of **Mdf**. The condition effect was near significance after a Greenhouse-Geisser correction [F(4,32) = 2.67, $p = 0.068$, $\eta^2 = 0.13$] and it became significant using a less conservative sphericity correction method like Huynh-Feldt [F(4,32) = 2.67, $p = 0.049$, $\eta^2 = 0.13$]; the group interaction was not significant [F(4,32) = 0.31, $p = 0.818$, $\eta^2 = 0.02$]. Pairwise *post-hoc* t-tests revealed significant differences between DWM+2Dpan and DWM+3Dgen, and between DWM+3Dpers and DWM+3Dgen (see Fig. 8). This result suggests the low reliability of generic HRTFs;

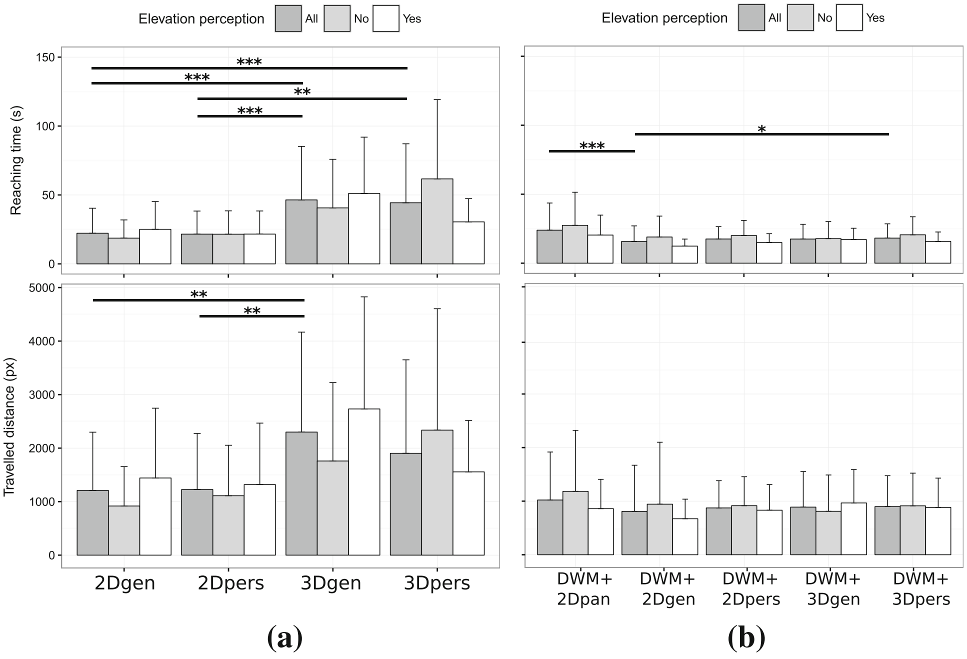

Fig. 7 Global statistics for (**a**) Experiment #1 and (**b**) Experiment #2 on reaching times and total traveled distance, grouped by feedback condition and elevation perception. *Asterisks* and *bars* indicate, where present, a significant difference (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$ at *post-hoc* test)

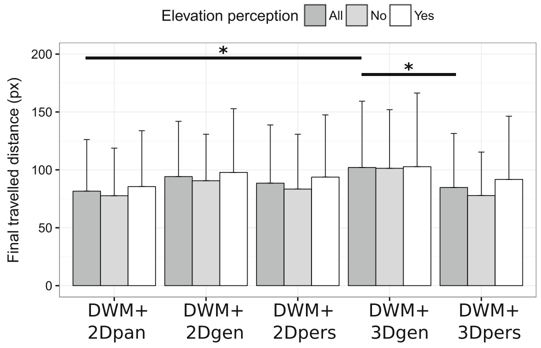

Fig. 8 Global statistics for Experiment #2 on "final" traveled distance grouped by feedback condition and elevation perception. *Asterisks* and *bars* indicate, where present, a significant difference (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$ at *post-hoc* test)

interestingly, DWM+2Dpan had good performance in the target proximity, where the interaction with DWM was highest in terms of intensity cues. No significant statistical effects were found in the remaining pairs ($p > 0.05$ in any case).

5.3 Experiment #3

Due to non-Gaussian data, a Kruskal-Wallis nonparametric one-way ANOVA was performed showing the significance of **Mt** [$\chi^2(4) = 124.43$, $p \ll 0.0001$].[3] Pairwise *post-hoc* Wilcoxon tests (Fig. 9, left) overall suggest that 3Dpers/DWM alone performed worse than all the remaining conditions, and that they also did not differ significantly between each other, while their combination (DWM+3Dpers) provided better performance than all remaining conditions except for the best condition, i.e., L+3Dpers.

Similarly, a Kruskal-Wallis nonparametric one-way A-NOVA reveals significance of **Md** [$\chi^2(4) = 137.62$, $p \ll 0.0001$]. In Fig. 9, middle, pairwise *post-hoc* Wilcoxon tests again suggest that 3Dpers and DWM performed poorly with regard to **Md** when rendered separately, while their auditory information integrated effectively when presented in combination (DWM+3Dpers), leading to similar performance with respect to L+3Dpers.

Similarly to the other indicators, a further analysis was performed on **Mdf**, through a Kruskal-Wallis nonparametric one-way ANOVA [$\chi^2(4) = 11.26$, $p < 0.05$]. Pairwise *post-hoc* Wilcoxon tests (Fig. 9, right) revealed only one significant difference in the final traveled distance, between DWM+3Dpers and L+3Dpers ($p < 0.05$). The impact of directional rendering in **Mdf** suggests a robust integration with the DWM, which was able to overcome performances in L+3Dpers.

6 General discussion

Figure 7a shows that the time to hit a 3D target is larger than the time to hit a 2D target; this can be attributed to the dimensionality of the task, recalling that the 3D task implied subjective processing of elevation cues. However, better performances in **Mt** are exhibited by participants of group

[3] Each distribution exhibited very high skewness towards a physical constraint. After logarithmic and Box-Cox transformations not all conditions passed the Shapiro-Wilk test.

'Elevation perception: Yes' reaching the 3D target using the individualized HRTF rather than the generic HRTF. This trend suggests that the use of 3D individualized HRTFs shows variabilities in accomplishing the task efficiently depending on individual perceptual factors (e.g. listener expertise [33] and sensitivity to spectral shape [2]).

Figures 7b, and 8 show that subjects overall exhibited similar times to reach 2D and 3D targets, and spanned comparable trajectory lengths as well. The only significant difference in performance between DWM+3Dgen and DWM+3Dpers appears in proximity of the target (see Fig. 8). In this situation, listening to personalized rather than template HRTFs is advantageous. More generally Fig. 8 provides values which, in the limits of their significance, show a trend that is coherent with the effort of adopting individual HRTFs as opposed to what appeared from Fig. 7a.

From qualitative data on navigation strategies collected with Q1, Q2, and Q3, participants' responses revealed key elements for the interpretation of the results:

- participants used lateralization cues first, i.e., they first moved the pen until the target was heard to be in the median plane, and then approached it by forward and backward movements;
- the proprioception of the virtual space boundaries, corresponding to the physical limits of the tablet surface, was useful to resolve front/back confusion at the beginning of each trial;
- azimuthal information had rapid changes in the proximity of the target, conversely elevation resulted in smoother spatial transitions which were ecologically more consistent and reported to be more pleasant by many participants;
- since the 2D target had no finite volume, the directional cues became progressively more unstable as the listening point approached the target. Although reliable and powerful, this additional cue was not natural as opposed to the consistency of the auditory information provided by the 3D scene.

Personalization, hence, played a key role in the proximity of a 3D elevated target, where conversely the 2D scenario exposed unrealistic cues.

Experiment #3 also indicates that navigation near the target based on 3D personalized directional cues alone was slightly (but not significantly) more efficient compared to when tubular and directional cues were displayed together—see Fig. 9 (right). One possible explanation of this result has been found by Shinn-Cun-ningham [34], who showed that environmental cues distorted directional cues: since our tubular model introduces strong reverberation cues and spectral artifacts which possibly overwhelm HRTF information, they could make the subjective decision on the direction to choose more problematic especially at greater distances; conversely, when the target gets closer then the tubular cues become progressively less invasive from a spectral point of view, hence making the localization process a joint combination of subjective fit to HRTFs and intensity boosting of DWM model.

Moreover, observed variations grouped by elevation perception for **Mt** and **Mdf** denoted listener-specific differences due to acoustic and non-acoustic factors [2,27,33]. In particular, the adopted personalization procedure enhances vertical discrimination and externalization with individual differences [19] leading to additional spatial information which might be exploited by the majority of the listeners.

From Fig. 9, it appears that the joint adoption of individualized HRTFs and DWM model (DWM+3Dpers) leads to subjective performances that are comparable to using individualized HRTFs and loudness model (L+3Dpers). This result is somewhat surprising, considering that listeners perform much better when using loudness alone (L) as opposed to the tube model (DWM) alone, i.e. once they are deprived of individualized directional cues. This evidence suggests that, while the use of absolute distance cues is of relatively little help for the reaching task compared to the use of loud-

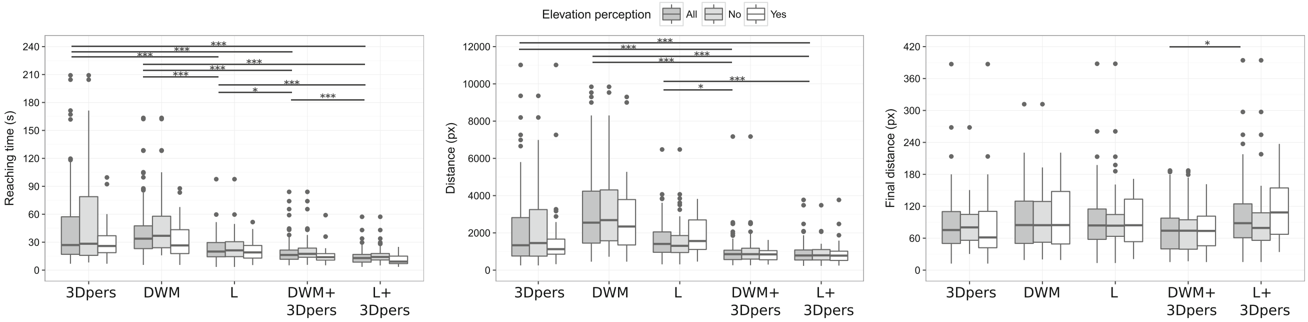

Fig. 9 Global statistics for Experiment #3 on (*left*) reaching time, (*middle*) total traveled distance, and (*right*) "final" traveled distance, grouped by feedback condition. *Asterisks* and *bars* indicate, where present, a significant difference ($*p < 0.05$, $**p < 0.01$, $***p < 0.001$ at *post-hoc* test)

ness, conversely such two types of cues become comparably powerful when used jointly with binaural information. A closer inspection shows significantly lower reaching times in L+3Dpers configuration, that is counterbalanced by significantly shorter final parts of the trajectories in DWM+3Dpers. Finally, trajectory lengths were not significantly different in the two configurations.

As reported in Table 1, reflections of the DWM tube add energy to the received signal, raising its amplitude by about 10 dB RMS. Such an effect may be responsible for the increase of the indicator **Mdf** in the DWM+3Dpers condition against the control condition L+3Dpers. An informal post-experimental questionnaire reported that participants detected the proximity of the target also thanks to the loudness cues [30] provided by the DWM. Accordingly, they tended to decelerate while searching for an increase in the higher intensity range even in mid-range distances: this may be a reason why the L+3Dpers condition performs statistically better in reaching time, **Mt**, than DWM+3Dpers.

In spite of the slightly better overall performance shown by the L+3Dpers over the DWM+3Dpers condition, once more it must be emphasized that the DWM-based approach has potential to result in a distance rendering model independent of loudness and other auditory cues which may be used to label source sounds and parallel sonification blocks. This peculiarity would leave designers free to employ the proposed model in rich auditory displays, although at greater computational cost than the L+3Dpers option.

7 Conclusions and future works

The distance cues from a DWM-based acoustic tube metaphor promise to integrate well with binaural cues though headphones. It is hard to evaluate our experimental results in absolute terms, mainly because there is no benchmark condition to compare with all of the rendering variations considered in this work. Ideally such a benchmark condition should employ real sound sources or individual HRTFs. Nonetheless, we have discovered that the proposed combination of the DWM model and personalized HRTFs performs comparably to auditory techniques employing salient and robust cues, such as panning and loudness.

An important design requirement for the proposed model was to provide distance cues that were independent of the source signal. A confirmation of this independence may come from repeating the tests using different sources, such as vocal and other auditory messages that are typical in these experiments [40]. Artifacts arising from the joint use of the tubular model and individualized HRTFs may be attenuated by employing a larger tube, with resonances falling within frequencies where the dispersion of the DWM is smaller. A larger tube, hence, should bring more realistic distance cues in the far-field. Moreover, novel personalization procedures such as ITD optimization [22] and frequency scaling techniques [29] will be considered in the future under static as well as dynamic scenarios, while looking for correspondences between localization accuracy and navigation performances [38].

Further experimental validation will thus require a larger 3D volume. We expect to achieve realistic volumes by substituting the DWM with equivalent finite-difference time-domain schemes [24] which allow for more efficient realization able to render complex scenarios. Besides involving multiple sound sources displayed together, such scenarios may ultimately provide enough flexibility to make it possible to synthesize distance cues such as those arising in real listening rooms. Multimodal displays [16,17] in mobile devices and web platforms [18] are some among the many applications that may benefit from such rendering techniques.

Acknowledgments The Authors are grateful to the volunteers who participated in this study, and to F. Altieri for his support in data collection. This work was supported by the research project PADVA , University of Padova, under Grant No. CPDA135702.

References

1. Algazi VR, Duda RO, Thompson DM, Avendano C (2001) The CIPIC HRTF database. In: Proc. IEEE Work. Appl. Signal Process., Audio, Acoust., New Paltz, New York, pp 1–4
2. Andéol G, Savel S, Guillaume A (2015) Perceptual factors contribute more than acoustical factors to sound localization abilities with virtual sources. Auditory Cogn Neurosci 8:451
3. Asano F, Suzuki Y, Sone T (1990) Role of spectral cues in median plane localization. J Acoust Soc Am 88(1):159–168
4. Blauert J (1983) Spatial hearing: the psychophysics of human sound localization. MIT Press, Cambridge
5. Boren B, Geronazzo M, Brinkmann F, Choueiri E (2015) Coloration metrics for headphone equalization. In: Proc. of the 21st Int. Conf. on auditory display (ICAD 2015), Graz, pp 29–34
6. Bronkhorst AW, Houtgast T (1999) Auditory distance perception in rooms. Nature 397:517–520
7. Campbell D, Palomaki K, Brown G (2005) A matlab simulation of “shoebox” room acoustics for use in research and teaching. Comput Inf Syst 9(3):48
8. De Sena E, Hacihabiboglu H, Cvetkovic Z (2011) Scattering delay network: an interactive reverberator for computer games. In: Audio engineering society conf.: 41st Int. conf.: audio for games
9. Devallez D, Fontana F, Rocchesso D (2008) Linearizing auditory distance estimates by means of virtual acoustics. Acta Acust United Acust 94(6):813–824
10. Fontana F, Rocchesso D (2003) A physics-based approach to the presentation of acoustic depth. In: Proc. Int. Conf. on Auditory Display, Boston, pp 79–82
11. Fontana F, Rocchesso D (2008) Auditory distance perception in an acoustic pipe. ACM Trans Appl Percept 5(3):16:1–16:15
12. Fontana F, Savioja L, Välimäki, V (2001) A modified rectangular waveguide mesh structure with interpolated input and output points. In: Proc. Int. Computer Music Conf., ICMA, La Habana, pp 87–90

13. Gardner WG, Martin KD (1995) HRTF measurements of a KEMAR. J Acoust Soc Am 97(6):3907–3908
14. Geronazzo M (2014) Mixed structural models for 3D audio in virtual environments. Ph.D. thesis, Information Engineering, Padova
15. Geronazzo M, Avanzini F, Fontana F (2015) Use of personalized binaural audio and interactive distance cues in an auditory goal-reaching task. In: Proc. of the 21st int. conf. on auditory display (ICAD 2015), Graz, pp 73–80
16. Geronazzo M, Bedin A, Brayda L, Avanzini F (2014) Multimodal exploration of virtual objects with a spatialized anchor sound. In: Proc. 55th int. conf. audio eng. society, spatial audio, Helsinki, pp 1–8
17. Geronazzo M, Bedin A, Brayda L, Campus C, Avanzini F (2016) Interactive spatial sonification for non-visual exploration of virtual maps. Int. J. Hum Comput Stud 85:4–15
18. Geronazzo M, Kleimola J, Majdak P (2015) Personalization support for binaural headphone reproduction in web browsers. In: Proc. 1st Web Audio Conference. Paris
19. Geronazzo M, Spagnol S, Bedin A, Avanzini F (2014) Enhancing vertical localization with image-guided selection of non-individual head-related transfer functions. In: IEEE int. conf. on acoustics, speech, and signal processing (ICASSP 2014), Florence, pp 4496–4500
20. Huopaniemi J, Savioja L, Karjalainen M (1997) Modeling of reflections and air absorption in acoustical spaces: a digital filter design approach. In: Proc. IEEE workshop on applications of signal processing to audio and acoustics. IEEE, New Paltz, pp 19–22
21. Iida K, Ishii Y, Nishioka S (2014) Personalization of head-related transfer functions in the median plane based on the anthropometry of the listener's pinnae. J Acoust Soc Am 136(1):317–333
22. Katz BFG, Noisternig M (2014) A comparative study of interaural time delay estimation methods. J Acoust Soc Am 135(6):3530–3540
23. Katz BFG, Parseihian G (2012) Perceptually based head-related transfer function database optimization. J Acoust Soc Am 131(2):EL99–EL105
24. Kowalczyk K, van Walstijn M (2008) Formulation of locally reacting surfaces in FDTD/K-DWM modelling of acoustic spaces. Acta Acust United Acust 94(6):891–906
25. Lu YC, Cooke M, Christensen H (2007) Active binaural distance estimation for dynamic sources. In: Proc. INTERSPEECH, Antwerp, pp 574–577
26. Magnusson C, Danielsson H, Rassmus-Gröhn K (2006) Non visual haptic audio tools for virtual environments. In: McGookin D, Brewster S (eds.) Haptic and audio interaction design, no. 4129 in Lecture Notes in Computer Science. Springer, Berlin, pp 111–120
27. Majdak P, Baumgartner R, Laback B (2014) Acoustic and non-acoustic factors in modeling listener-specific performance of sagittal-plane sound localization. Front Psychol 5:1–10
28. Masiero B, Fels J (2011) Perceptually robust headphone equalization for binaural reproduction. In: 130th AES convention, London, England, pp 1–7
29. Middlebrooks JC (1999) Virtual localization improved by scaling nonindividualized external-ear transfer functions in frequency. J Acoust Soci Am 106(3):1493–1510
30. Moore BC, Glasberg BR, Baer T (1997) A model for the prediction of thresholds, loudness, and partial loudness. J Audio Eng Soc 45(4):224–240
31. Neuhoff JG (2001) An adaptive bias in the perception of looming auditory motion. Ecol Psychol 13(2):87–110
32. Parseihian G, Katz BFG, Conan S (2012) Sound effect metaphors for near field distance sonification. In: Proc. int. conf. on auditory display, Atlanta, pp 6–13
33. Schönstein D, Katz BFG (2010) Variability in Perceptual Evaluation of HRTFs. In: 128th Convention of the Audio Engineering Society, AES London, 11 p
34. Shinn-Cunningham B (2000) Learning reverberation: considerations for spatial auditory displays. In: Proc. int. conf. auditory display (ICAD'00). Atlanta, pp 126–134
35. Spagnol S, Geronazzo M, Avanzini F (2013) On the relation between pinna reflection patterns and head-related transfer function features. IEEE Trans Audio Speech Lang Process 21(3):508–519
36. Speigle J, Loomis J (1993) Auditory distance perception by translating observers. In: Virtual reality, 1993. Proceedings., IEEE 1993 symposium on research frontiers in, pp 92–99
37. Valimaki V, Parker JD, Savioja L, Smith JO, Abel JS (2012) Fifty years of artificial reverberation. Audio Speech Lang Process IEEE Trans 20(5):1421–1448
38. Viaud-Delmon I, Warusfel O (2014) From ear to body: the auditory-motor loop in spatial cognition. Front Neurosci 8:283
39. Wiener JM, Büchner SJ, Hölscher C (2009) Taxonomy of human wayfinding tasks: a knowledge-based approach. Spat Cogn Comput 9(2):152–165
40. Zahorik P (2002) Assessing auditory distance perception using virtual acoustics. J Acoust Soc Am 111(4):1832–1846
41. Zahorik P (2002) Auditory display of sound source distance. In: Proc. int. conf. on auditory display. Kyoto
42. Zahorik P (2002) Direct-to-reverberant energy ratio sensitivity. J Acoust Soc Am 112(5):2110–2117
43. Zahorik P, Brungart DS, Bronkhorst AW (2005) Auditory distance perception in humans: a summary of past and present research. Acta Acust United Acust 91(3):409–420