# Auditory navigation with a tubular acoustic model for interactive distance cues and personalized head-related transfer functions: an auditory target-reaching

**J. on Multimodal User Interfaces manuscript No.**
(will be inserted by the editor)

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

## An auditory target-reaching task

**Michele Geronazzo · Federico Avanzini · Federico Fontana**

Received: date / Accepted: date

**Abstract** While the angular spatialization of sounds through individualized Head-related transfer functions (HRTFs) has been extensively investigated in auditory display research, also leading to effective real-time rendering of these functions, conversely the interactive simulation of egocentric distance information has received less attention. The latter, in fact, suffers from lack of real-time rendering solutions also due to a too sparse literature on the perception of dynamic distance cues. By adding a virtual environment based on a Digital waveguide mesh (DWM) model simulating a small tubular shape to a binaural rendering system through selection techniques of HRTF, we have come up with an auditory display affording interactive selection of absolute 3D spatial cues of direction as well as egocentric distance. The tube metaphor in particular minimized loudness changes with distance, hence providing mainly direct-to-reverberant and spectral cues. A target-reaching task assessed the proposed display: participants were asked to explore a 2D virtual map with a pen tablet and hit a sound source (the target) using only auditory information; then, subjective time to hit and traveled distance were analyzed. In a series of tests conducted using different directional cue rendering methods we showed that subjects performed similarly, either they had to reach an elevated (3D) or vertically unbounded (2D) target in spite of the lower complexity of the latter task. In another series of tests we showed that subjects using absolute distance cues from the tube model performed comparably to when they could rely on more robust, although relative, intensity cues. These results suggest that participants made proficient use of both elevation and reverberation cues during the task once they were displayed as part of a coherent 3D sound model, in spite of the known complexity of use of both such cues. Further work is needed to add full physical consistency to the proposed auditory display.

**Keywords:** head-related transfer function, distance rendering, digital waveguide mesh model, individualization, target-reaching task, navigation.

M. Geronazzo, and F. Avanzini
Dept. of Information Engineering, University of Padova
via Gradenigo 6B, 35131-Padova, Italy
Tel.: +39-049-827-6976
E-mail: {geronazzo,avanzini}@dei.unipd.it

F. Fontana
Dept. of Mathematics and Computer Science,
University of Udine
via delle Scienze 206, 33100-Udine, Italy
E-mail: federico.fontana@uniud.it

## 1 Introduction

The accurate acoustic rendering of sound source distance is an uncertain task; in fact, the auditory cues of egocentric distance have been shown to be essentially unreliable since they depend on several factors, which can be hardly kept under control in the experimental setup. Researchers along the years 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 that is at the origin of the stimulus: the more unfamiliar an original sound is, the more difficult for a subject to disaggregate acoustic source information from the environmental cues that shape the sound on its way to the listener.

The ambiguity about the origin (either source- or environment-based) of the auditory cues that confer distance attributes to a sound makes the perception of a moving sound source especially interesting to investigate: by listening to dynamic cues humans in fact receive a range of psychophysical information about the source sound in relation with its continuous modifications due to the environment: by progressively isolating the former out of these modifications, listeners in theory should learn about both and hence be able to improve the source localization. On the other hand, the robust control of a distance recognition experiment involving moving sound sources has proven inherently difficult to achieve. So far, the literature on the topic is sparse and limited to virtual acoustic setups; furthermore, due to some unavoidable complexity of the dynamic rendering models this literature merges psychological issues with arguments of sound processing: Lu *et al.* describe a model capable of rendering motion parallax and acoustic $\tau$, already noted by Spiegle and Loomis as salient cues for the positional recognition in a moving listener and source scenario [25,37]. Perhaps more importantly, moving sound sources evoke so-called "looming" effects causing localization bias especially if they elicit emotional cues, such as when the sound of a rapidly approaching wild animal is displayed [32].

In spite of its unreliability and subjective dependency, the egocentric distance remains highly interesting for auditory display purposes as an informative dimension having immediate physical interpretation and, hence, strong ecological meaning. Inaccuracies in its quantitative interpretation deriving from the uncertainty of the psychophysical maps are counterbalanced by the importance that distance has in auditory scene description. Zahorik suggested design guidelines that are of great help for realizing accurate auditory displays provided specific technological constraints [41]. Such guidelines would probably become even more challenging if moving sources were accounted for. To date, the mentioned scarcity of experimental results makes the design of dynamic, especially interactive distance rendering models still a matter of discussion.

Near-field distance has been sonified using auditory metaphors, too [33]: by rendering robust effects (such as the repetition rate of a beep) that are essentially disjoint with the sound source properties, clearly this approach has a good chance to translate in reliable distance estimations as soon as listeners get used with the proposed sonification. As well, in our research we put the focus on *absolute* cues, i.e., those which are not a function of the source sound. Specifically, we made an effort to select absolute references among the standard auditory distance cues: loudness, direct-to-reverberant energy ratio, spectrum, and binaural differences when the source is nearby the listener's head. This effort had a threefold aim: i) to preserve the sonic signature of the sound source, particularly its loudness, ii) to avoid cannibalization of otherwise informative additional cues, and iii) to maintain sufficient ecological consistency of 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.

If the rendering is not limited to nearby sources then 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 in general, it is known that the presence of these environmental cues impact spatial auditory perception in two respects: while a listener's ability in perceiving sound source distance is enhanced, his/her ability in perceiving sound source direction is degraded in a complementary fashion [35]. 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 have been proven to provide effective distance cues even in uncommon/unrealistic environments. In an experiment where a loudspeaker could be moved inside a long, narrow pipe, listeners were in fact 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 quite compressed and non-linear. Later experiments made use of virtual rather than real environments, 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: all such shapes were realized through the use of distributed computational models, and at least have demonstrated that the search for a virtual environment capable of shaping the auditory cues until defining a linear map is a hard task [9].

Despite their psychophysical limitations, these computational models provide high versatility. For instance, simple Digital Waveguide Mesh (DWM) models and similar computational schemes have been employed offline to render auditory distance cues [10,7]; in practice they allow for moving source and listener positions everywhere inside the 3D shape. Interactivity, however, requires to make a leap forward: the model, in fact, 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-based model has been used to enable interactive reverberation for computer game applications [8].

In this work we propose a spatial sound rendering architecture that combines binaural (individualized HRTF based) rendering with a virtual (non-individualized DWM based) environment simulating a tubular shape. Partial support for this choice comes from an experiment making use of HRTFs containing also distance cues [41]: by stimulating subjects with such functions, directional cues were shown to be highly individual whereas distance evaluations were robust against non-individualization of the HRTFs. The motivations for the proposed architecture hence are twofold. First, it allows to decouple to some extent the rendering of directional and distance cues: in this way, we expect that environmental effects simulated through the DWM model can improve listeners' performance in sound distance estimation, while preserving their ability to estimate sound direction, as HRTF-related cues are not degraded or distorted by this simplified environment. Second, the proposed architecture allows real-time rendering.

The technical features of both binaural rendering and the DWM model are illustrated in Section 2. Section 3 describes the design and the results of an experimental task aimed at assessing the validity of the proposed approach using different rendering strategies: the experiment consists of 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 in the map (the target) 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 analyzed and discussed in Section 4. They show that participants using a 3D, HRTF-personalized display enabling absolute distance cues achieved a first level of spatial knowledge [39] by performing comparably to i) when they reached a 2D (i.e., vertically unbounded) instead of 3D (i.e., bounded and vertically offset) target, and ii) when they relied on relative (i.e., intensity) instead of absolute (i.e., direct-to-reverberant energy and spectral) cues of distance.

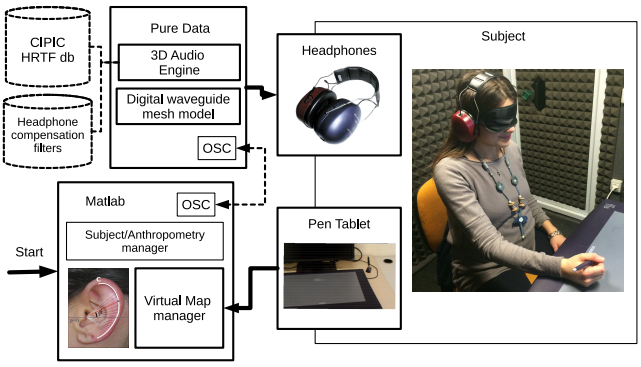



**Fig. 1** A schematic view of the system architecture.

These two results are particularly interesting, considered the known unreliability of the monaural cues of elevation 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 ecological sense.

## 2 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 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.

### 2.1 HRTF-based spatialization

The recording of individual HRIRs/HRTFs is both time- and resource-consuming, and technologies for binaural audio usually employ non optimal choice of 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 of the human body heavily affect the perception and the quality of the rendering [30]. 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].

In this paper, an image-based HRTF selection technique is briefly summarized (see [**?**] for details) where relevant individual anthropometric features are extracted from one image of the user's pinna. Specifically, a mismatch function between the main pinna contours and corresponding spectral features (frequency notches) of the HRTFs in the database is defined according to a ray-tracing interpretation of notch generation [36]. The first notch of HRTF responsible for the first pinna reflection 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 Figure 1).

For a given elevation $\phi$ of the incoming sound, the reflection distance can be computed as follow

$$d(\phi) = ct(\phi), \tag{1}$$

where $t(\phi)$ is the temporal delay between the direct and reflected rays and $c$ is the speed of sound. The corresponding notch frequency, $f_0(\phi)$, is estimated by the following equation

$$f_0(\phi) = \frac{c}{2d_c(\phi)}, \tag{2}$$

according to the assumption of negative reflection coefficient and one-to-one correspondence between reflection and generated notch [36]. Given a user whose individual HRTFs are not available, the mismatch $m$ between $f_0$ notch frequencies estimated from Eq. (2) and the notch frequencies $F_0$ of an arbitrary HRTF set is defined as:

$$m = \frac{1}{|\phi|} \sum_{\phi} \frac{|f_0(\phi) - F_0(\phi)|}{F_0(\phi)}, \tag{3}$$

where elevation $\phi$ 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.

### 2.2 Digital waveguide mesh model

The DWM we use in our experiment was obtained by translating existing MATLAB code from the authors into a C++ external program for the Pure Data real-time environment. [1]

[1] As its optimization would have required different simulation schemes, such as finite-difference time-domain, that was outside the scope of this work, we chose to go on with the experimental plan as soon as a reliable interactive distance rendering tool was obtained in the form of an object for Pure Data (http://puredata.info).

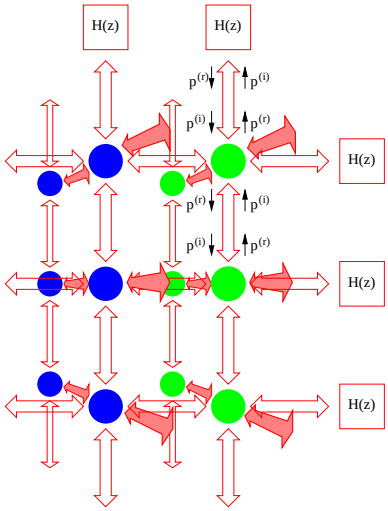
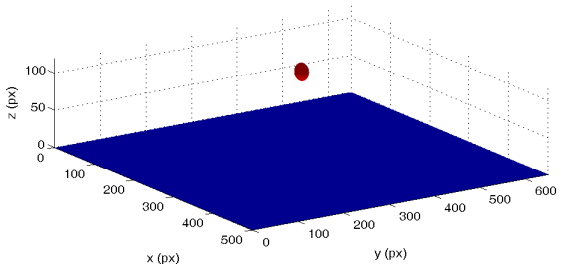


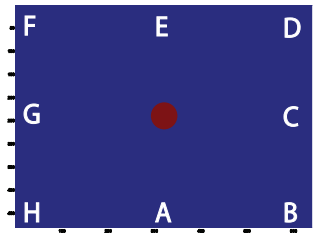

**Fig. 2** Detail of the 3D DWM: scattering junctions and boundary filters.

The DWM model follows a straightforward design, in which the scattering junctions forming the mesh boundary are coupled with filters modeling frequency-dependent air absorption [20]. Figure 2 shows a particular of this design, exposing scattering junctions and boundary filters exchanging pressure wave signals each with its adjacent nodes (either junctions or filters). The mesh has the shape of a square tube counting $29 \times 5 \times 5 = 725$ junctions. Of these junctions, $5 \times 5 = 25$ form either termination of the tube whereas $29 \times 5 = 145$ form each of the four tube surfaces. One termination was modeled like an open end (i.e. $H(z) = -1$) whereas the other termination was modeled like a closed end (i.e. $H(z) = 1$). Finally, each surface was modeled like an absorbing wall with larger absorption toward the high frequencies: this model is made by realizing the transfer function $H(z)$ of each boundary filter in the form of a simple first-order low-pass characteristic.

Once running at 44.1 kHz, the proposed DWM simulates sound wave propagation along a tiny tubular environment. The distance rendering effect depends on the relative positions of the source and listening point, respectively corresponding to junctions in which the audio signal was injected and picked up. We simulated an acoustic scenario in which both the source and the listening point laid in the center of the square section, and the listening point was close to the open end. Conversely the source could be moved back and forth along the main axis of the tube starting from nearby the closed end, in this way varying its relative distance from the listening point. Moving the source point alone was sufficient for our purposes, as it has the advantage of avoiding sound discontinuities caused by dynamically varying the junction where the signal is picked up. Be-




(a)




(b)

**Fig. 3** The virtual map in pixels. (a) The target is the central red sphere. (b) Virtual starting positions for audio exploration are marked in lexicographic order.

sides these discontinuities, a similar artifact arises at the listening point supposed stationary also if the moving source signal is injected in the DWM with occasional jumps from one junction to another, even if these junctions are adjacent each to the other. This artifact can be minimized by distributing the signal, for instance by linearly de-interpolating each sample value across such junctions as we did in our model when the source point position laid in between two pick-up points [12].

## 3 Experiments: target reaching

The main goal of this experimental evaluation 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 behavioral and performance indicators collected from experimental data. These data were obtained through a target-reaching task, in which participants had to hit a virtual sound source under different auditory feedback conditions, rendered through headphones displaying the target's relative position inside a workspace physically consisting of a pen tablet (Figure 1). Experiment #1 dealt with the interaction between tubular acoustics and different rendering methods for directional cues. Experiment #2 focused on auditory navigation using different combinations of distance and directional cues.

Eight participants (6 male and 2 female, age ranging 23 to 51, mean $30.8 \pm 8.7$) took part in the first experiment. Six participants (4 male and 2 female, age ranging 26 to 41, mean $30.8 \pm 5.9$) took part in the second experiment. Four of these participants took part to both experiments. All participants reported normal hearing and had previous experience in psychoacoustic experiments with binaural audio reproduction through headphones.

### 3.1 Apparatus

Figure 1 depicts a schematic view of the overall system architecture. All tests were performed using Matlab, that controlled the entire setup by also recording the 2D position on the pen tablet, a $12 \times 18$ 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 communication between Matlab and Pure Data.

Audio output was operated by a Roland Edirol AudioCapture 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 a frequency response with no pronounced peaks or notches in between 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 made upon regular and stable frequency responses guaranteed no corruption of the localization cues contained in the HRTFs [28], as well as an effective equalization of the headphones up to approximately $8 - 10$ kHz. In this type of auditory experiments, in fact, the design of individualized headphone equalization filters can introduce dependencies on the headphone position around the head, making the subjective design less recommended than the use of a generalized equalizer compensating the frequency response of a stable headphone [14].

### 3.2 Stimuli

The virtual target sound was placed at the center of the $640 \times 480$ pixels working area. It had the form of a sphere with radius equal to 25 pixels. The sphere was placed at a height of 120 pixels from the virtual ground

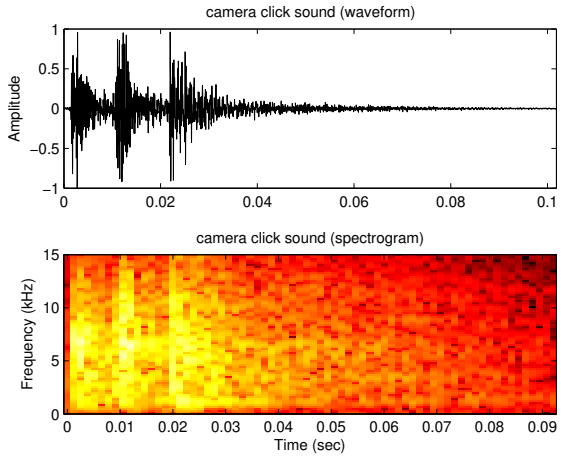



**Fig. 4** Waveform and spectrogram of the camera click.

level (see Figure 3). 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.[2]

The source sound consisted of a camera click with 100 ms duration (see Figure 4) repeated every 300 ms, with maximum amplitude level at the entrance of the ear canal amounting to 60 dB(A) for experiment #1 and 65 dB(A) for experiment #2, respectively. The period between subsequent clicks was large enough to include possible reverberant tails due to reverberation cues being introduced by the tubular environment. If the pen was moved beyond the boundaries of the working area then the system signalled the illegal position of the pen by playing white noise until a correct position was restored.

### *3.2.1 #1: Directional cues and tubular acoustics*

Experiment #1 tested three different directional rendering methods: (i) intensity panning, (ii) dummy-head HRTF rendering, and (iii) personalized HRTF rendering. The CIPIC database [1] was chosen as source of HRTFs. This database contains 45 HRTF sets measured in the far field, hence free of distance information, with azimuth and elevation angles spanning the ranges $[0°, 360°)$ and $[-45°, 230.625°]$, respectively. Details about the three directional rendering methods are as follows.

(i) The intensity panning was computed as:

$$G_{l,r} = \frac{1}{2}\left(1 \pm \cos(\theta + 90°)\right), \quad G_{l,r} \in [0, 1] \tag{4}$$

where $\theta$ corresponds to the azimuthal angle between source and listener in the horizontal plane, varying in the range $[0°, 360°)$. Eq. (4) leads in particular to these positions: $\theta = 0°/180°$, corresponding to in-axis position with the sound source ($G_{l,r} = 1/2$), and $\theta = \pm 90°$ respectively denoting lateral sources on the left ($G_{l,r} = 1$) and right ($G_{l,r} = 0$) side.

(ii) CIPIC subject no. 165, that is a KEMAR with large pinnae, was chosen, yielding a template HRTF for all participants.

(iii) The personalization procedure described in Section 2 was used to select 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 his/her manually traced contours and notch central frequencies.

The dimensionality could be set to 3D or downscaled to 2D, by locking the elevation angle to 0 degrees hence forcing the rendering model to span the sole horizontal plane. Distance, conversely, was always rendered on top of the directional cues through the tubular model described in Section 2.2.

The combination of directional and distance rendering resulted in five experimental conditions, which are summarized here along with their acronyms:

1. tube model and intensity panning (DWM+2Dpan);
2. tube model and generic HRTF directional cues in 2D (DWM+2Dgen);
3. tube model and personalized HRTF directional cues in 2D (DWM+2Dpers);
4. tube model and generic HRTF directional cues in 3D (DWM+3Dgen);
5. tube model and personalized HRTF directional cues in 3D (DWM+3Dpers).

These conditions are listed in increasing order of auditory information, in terms of dimensionality (2D/3D) and personalization (generic/personalized). In particular, DWM+2Dpan acted as a control condition since panning provides the simplest angular cues of intensity in headphone reproduction.

### *3.2.2 #2: Complementarity of auditory information*

In Experiment #2, the rendering of angular position (azimuth and elevation) was enabled by the 3Dpers condition only, whereas distance was rendered through two different approaches: a 6-dB law modeling ideal loudness attenuation in open air with distance, and the tubular model described in Section 2.2. The combination of direction and distance rendering resulted in five experimental conditions, which are summarized here along with their acronyms:

[2] The geometrical properties of the virtual map were chosen in order to ensure detectable elevation cues from the HRTF selection procedure (see Sec. 2.1).

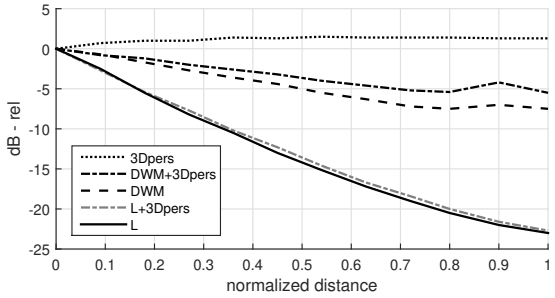



**Fig. 5** Average amplitude of the stimuli used in the respective experimental conditions as a function of normalized distance. Amplitude values ranging from the smallest (normalized value equal to 0) to the largest (normalized value equal to 1, corresponding to position "A" in Figure 3.b) egocentric distance.

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

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

Figure 5 depicts, for all conditions, average amplitudes measured as a function of egocentric distance. The relative values were computed by subtracting the dB RMS values measured at the smallest distance, reported in Table 1 below.

| | 3Dpers | L | DWM | DWM+ 3Dpers | L+ 3Dpers |
|---|---|---|---|---|---|
| amplitude (dB RMS) | 65 | 60 | 72 | 78 | 65 |

**Table 1** Amplitudes in dB RMS of stimuli at the smallest egocentric distance for each auditory condition. Measurements for 3Dpers had HRTFs from KEMAR [13] as reference.

From these measurements it can be noted that intensity in DWM and DWM+3Dpers conditions changed when the virtual source was moved nearby the auditory target, but not when it was kept moving in the far-field. Moreover DWM+3Dpers produced higher intensity values than DWM alone, showing an interaction between HRFT 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 resonances were no longer present.

### 3.3 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 were 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 far from 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. Participants were asked to complete the task starting from eight different positions at the boundary of the workspace, as depicted in Figure 3(b). Before each trial began, the experimenter lifted and moved the pen to random positions of the tablet area as it can be made with any relative pointing device such as the mouse, and then helped the participant to grasp it again.

Every condition was repeated 8 times (one for each virtual starting position), for a total of 40 trials per participant. Starting position and auditory conditions were randomly balanced across trials.

### 3.4 Measures and data analysis

Each trial was evaluated in terms of three main performance indicators:

- **M1** absolute reaching time: the time spent by the participant to complete the trial;
- **M2** total traveled distance: the length of the trial trajectory;
- **M3** final traveled distance: the length of the trial trajectory in the last 240 ms of exploration.

In these experiments participants trajectories had large variability, and **M1** with **M2** were thus assumed to be appropriate global indicators of performance. Moreover, **M3** was added as a third indicator, as it was assumed to be related to participants' confidence in being nearby the target [17].

Preliminary analysis of gaussianity was performed on each condition by means of Shapiro-Wilk test for

normality, which revealed violations in sample distributions.[3] Accordingly, a Kruskal Wallis nonparametric one-way ANOVA with five levels of feedback condition was performed to assess the statistical significance of **M1**, **M2**, and **M3** in both experiments. Pairwise *post-hoc* Wilcoxon tests for paired samples with false discovery rate (FDR) correction procedures on p-values provided statistical significances in performance between auditory conditions. For the sake of simplicity, in the next section we report the adjusted p-values.

## 3.5 Results

### *3.5.1 Experiment #1*

A Kruskal Wallis nonparametric one-way ANOVA with five levels of feedback condition was performed to assess the statistical significance of **M1** [$\chi^2(4)$=15.13, $p < 0.01$]. Pairwise *post-hoc* Wilcoxon tests (Figure 6(a)) revealed statistically significant improvements in performance (decreasing reaching times) between conditions DWM+2Dgen and DWM+2Dpan ($p < 0.001$), DWM+2Dgen and DWM+3Dpers ($p < 0.05$), and between DWM+3Dgen and DWM+3Dpan ($p < 0.05$). These results suggest that DWM+2Dpan performed worse than conditions with generic HRTFs but did not differentiate from conditions with personalized HRTFs. Moreover, DWM+2Dgen was able to provide reliable and sufficient cues compared to personalized auditory conditions in 3D space. It has to be noted that the degree of statistical significance for the pair DWM+2Dpan and DWM+2Dgen is very high ($p < 0.001$), denoting an outperformance of binaural rendering in **M1**. On the other hand, no significant statistical effects were found in pairs

- DWM+2Dgen and DWM+3Dgen ($p = 0.404$),
- DWM+2Dgen and DWM+2Dpers ($p = 0.106$),
- DWM+3Dgen and DWM+2Dgen ($p = 0.257$),
- DWM+3Dgen and DWM+3Dpers ($p = 0.133$),
- DWM+2Dpers and DWM+3Dpers ($p = 0.581$),
- DWM+2Dpers and DWM+2Dpan ($p = 0.132$),
- DWM+3Dpers and DWM+2Dpan ($p = 0.417$).

Similarly, a Kruskal Wallis nonparametric one-way ANOVA with five levels of feedback condition was performed to asses the significance of **M2** [$\chi^2(4)$=5.42, $p = 0.247$]. This metric did not exhibit any significant result. Figure 6(b) depicts global statistics for traveled distance among conditions, suggesting that DWM+2Dgen provided the best performance in average. DWM+3Dgen had similar average performances but many outliers in the distribution, suggesting that the accessibility to 3D information was highly variable among participants.

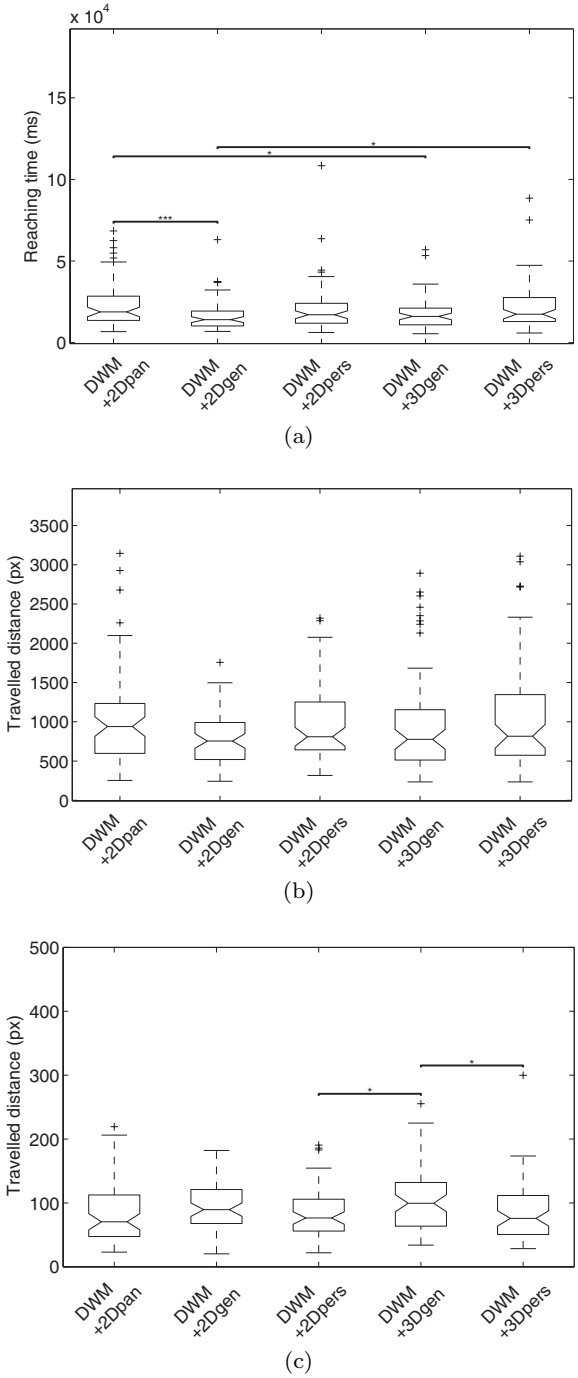



**Fig. 6** Global statistics for Experiment #1 on (a) reaching times, (b) total traveled distance, and (c) "final" traveled distance, grouped by feedback condition. Asterisks and bars indicate, where present, a significant difference (* $p < 0.05$, ** $p < 0.01$ , *** $p < 0.001$ at *post-hoc* test).

[3] Each distribution exhibited high skewness towards a physical constraint, i.e. the minimum possible traveled distance. After logarithmic and Box-Cox transformations not all conditions passed the Shapiro-Wilk test.

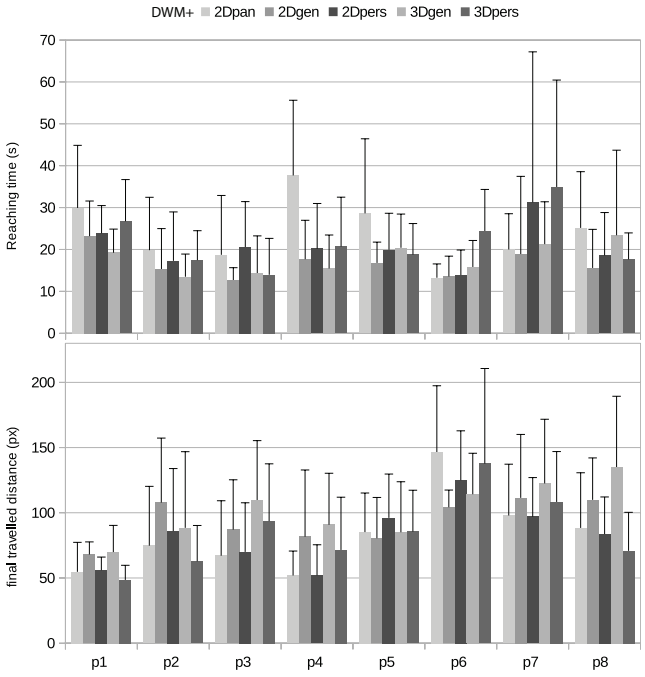



**Fig. 7** Results for Experiment#1; (top) average and standard deviation (across all trials for each condition) of reaching times for each participant, and (bottom) average and standard deviation (across all trials for each condition) of final traveled distance for each participant.

A further analysis was performed on **M3**, in order to assess participants' awareness of being in proximity of the target through auditory spatial information [39]. A Kruskal Wallis nonparametric one-way ANOVA with five levels of feedback condition was performed to asses the statistical significance of **M3** [$\chi^2(4)$=11.64, $p < 0.05$]. Pairwise *post-hoc* Wilcoxon tests revealed the following decreases in the final traveled distance: DWM+ 2Dpers and DWM+3Dgen ($p < 0.05$), and DWM+ 3Dpers and DWM+3Dgen ($p < 0.05$). This result provided a confirmation of the previous ones, pointing at the low reliability of generic HRTF rendering in 3D space navigation, and supporting the need for personalization. DWM+2Dgen condition exhibited a similar average trend, but did not statistically differentiate from other conditions in this experiment. Finally, panning had similar average performances in **M3** than personalized auditory conditions, but higher standard error. No significant statistical effects were found in pairs

- DMW+2Dgen and DMW+3Dgen ($p = 0.237$),
- DMW+2Dgen and DMW+2Dpers ($p = 0.194$),
- DMW+2Dgen and DWM+3Dpers ($p = 0.194$),
- DMW+2Dgen and DWM+2Dpan ($p = 0.194$),
- DMW+3Dgen and DWM+3Dpers ($p = 0.052$),
- DMW+3Dgen and DWM+2Dpan ($p = 0.128$),
- DMW+2Dpers and DWM+3Dpers ($p = 0.828$),
- DMW+2Dpers and DWM+2Dpan ($p = 0.828$),
- DMW+3Dpers and DWM+2Dpan ($p = 0.828$).

Figure 7 illustrates the performances **M1** and **M3** for each of the eight participants. It can be seen that p1, p4, and p5 totalized the worst time to hit with DWM+2Dpan; conversely, personalization had negative impact in the same performance figure for p7, and, limitedly to the 2D and 3D space, respectively for p3 and p6. Moreover, 3D generic cues were an issue for p8. On the other hand, personalization played a determinant role for p1, p2, and p8 in **M3**. Again, personalization cues were not able to provide reliable information to p6 in the final traveled distance.

### *3.5.2 Experiment #2*

A Kruskal Wallis nonparametric one-way ANOVA with five levels of feedback condition was performed to asses the statistical significance of **M1** [$\chi^2(4)$=78.23, $p << 0.0001$]. Pairwise *post-hoc* Wilcoxon tests (Figure 8(a)) revealed statistically significant improvements in performance (decreases in reaching times) between 3Dpers and L, DWM+3Dpers, L+3Dpers (all with $p << 0.001$), between 3Dpers and DWM ($p < 0.05$), between L and L+3Dpers ($p < 0.001$), between DWM and DWM+ 3Dpers ($p < 0.001$), between DWM and L, L+3Dpers (all with $p << 0.001$), between DWM+3Dpers and L+3Dpers ($p << 0.001$). These results suggest that 3Dpers/DWM alone performed worse than all the remaining conditions, and that they also differed significantly between each other, while their combination (DWM+3Dpers) provided better performance than all remaining conditions except that L+3Dpers (the best condition). It has to be noticed that degree of statistical significance is very high with the exception of L and L+3Dpers, DWM and DWM+3Dpers, and DWM and 3Dpers comparisons. On the other hand no statistical significance was found between L and DWM+3Dpers ($p = 0.359$).

Similarly, a Kruskal Wallis nonparametric one-way ANOVA with five levels of feedback condition was performed to asses the significance of **M2** [$\chi^2(4)$=77.95, $p << 0.0001$]. In Figure 8(b), statistical significances are computed using pairwise *post-hoc* Wilcoxon test. Decreases in total traveled distance were reported for following condition pairs: 3Dpers and L ($p < 0.001$), 3Dpers and DWM+ 3Dpers ($p << 0.001$), 3Dpers and L+3Dpers ($p << 0.001$), L and L+3Dpers ($p < 0.001$), L and DWM+3Dpers ($p < 0.05$), DWM and L ($p << 0.001$), DWM and DWM+3Dpers ($p << 0.001$), DWM and L+3Dpers ($p << 0.001$). On the other hand, no statistical differences were found between 3Dpers and DWM ($p = 0.874$), and DWM+3Dpers and L+3Dpers ($p = 0.190$). Again, 3Dpers and DWM performed poorly

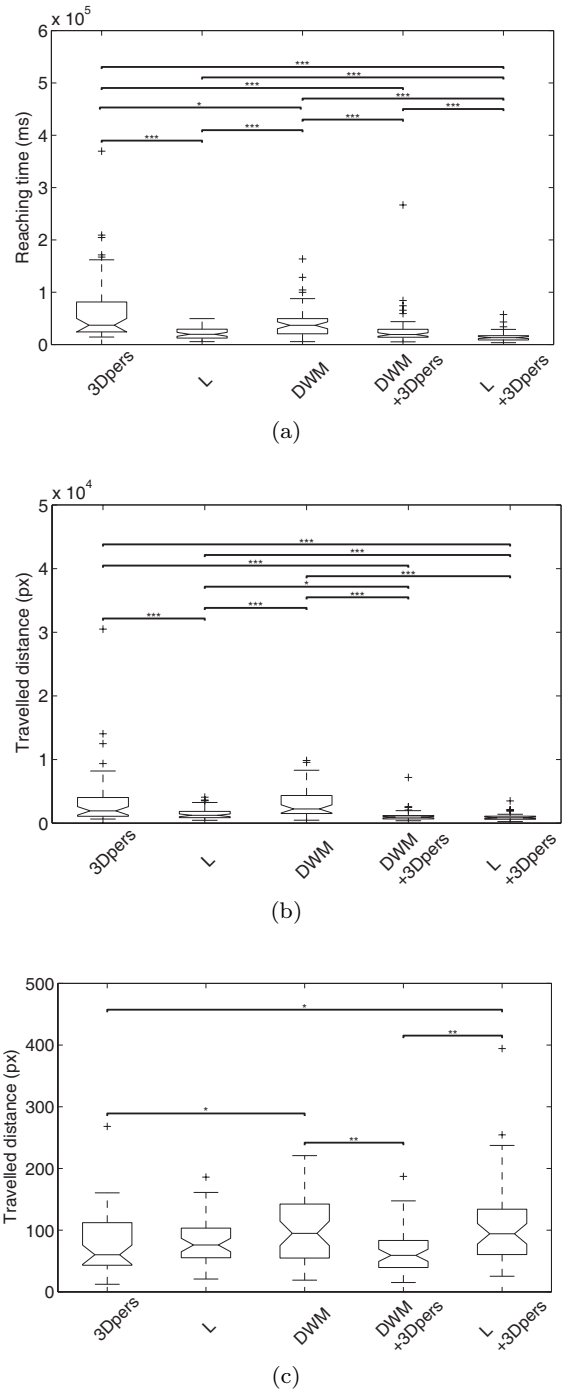



**Fig. 8** Global statistics for Experiment #2 on (a) reaching time, (b) total traveled distance, and (c) "final" traveled distance, grouped by feedback condition. Asterisks and bars indicate, where present, a significant difference (* $p < 0.05$, ** $p < 0.01$ , *** $p < 0.001$ at *post-hoc* test).

with regard to **M2** when rendered separately, while these results suggest that their auditory information integrated very effectively when rendered in combination (DWM+3Dpers), leading to similar performance with respect to L+3Dpers.

A further analysis was performed on **M3**, through a Kruskal Wallis nonparametric one-way ANOVA with five levels of feedback condition [$\chi^2(4)$=17.76, $p < 0.01$]. Pairwise *post-hoc* Wilcoxon tests revealed the following decreases in the final traveled distance: DWM and 3Dpers ($p < 0.05$), DWM and DWM+3Dpers ($p < 0.05$), L+3Dpers and 3Dpers, DWM+3Dpers (both $p < 0.05$). No significant statistical effects were found in pairs

- 3Dpers and L ($p = 0.745$),
- 3Dpers and DMW+3Dpers ($p = 0.271$),
- L and DWM ($p = 0.079$),
- L and DWM+3Dpers ($p = 0.119$),
- L and L+3Dpers ($p = 0.095$),
- DWM and L+3Dpers ($p = 0.971$).

The impact of directional rendering in **M3** suggested a robust integration with DWM which will be discussed in the following section.

## 4 General discussion

Figures 6(a), 6(b) and 6(c) show that subjects overall exhibited similar times to reach both the 2D and 3D targets, and that they spanned comparable trajectory lengths as well. In order to discuss in more depth the differences between tasks in two vs. three dimensions, it must be emphasized once more that reaching the 3D target implied for each subject to elaborate also elevation cues. Now, Figure 6(a) shows that the time to hit a 3D target using personalized cues is larger than the time to hit a 2D target using the template HRTF; this can be explained on the light of the dimensionality of the task. However, better performances in **M1** are exhibited by participants reaching the 3D target using the template HRTF rather than the individualized HRTF. Although not significant, this difference in the performance suggests that the use of 3D individualized HRTFs may be not of real help for the subjects in accomplishing the task more efficiently.

Similar performances are figured out by observing the trajectory lengths, illustrated by Figure 6(b) this time with no significant differences. Indeed, the only significant difference in performance between DWM+3Dgen and DWM+3Dpers appears towards the completion of the task (see Figure 6(c)), when subjects are in proximity of the target. In this situation, listening to personalized rather than template HRTFs is advantageous. More in general Figure 6(c) 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 Figures 6(a) and 6(b).

An informal post-test questionnaire on navigation strategies was conducted. Partcipants' responses revealed key elements for the interpretation of the results:

- participants usually tried to minimize lateralization, i.e. centered the target, and then disambiguated front to back in order to reach the target;
- the virtual space boundaries and physical limits of the tablet surface gave strong cues to resolve front/back confusion at the very beginning of each trial;
- azimuthal information had rapid changes in the proximity of the target and elevation cues allowed smooth spatial transitions which are ecologically consistent and reported as pleasant from many participants;
- in 2D conditions, elevation was always set to 0 and participants experienced an unstable spatial panning in the proximity of the target that was used as the dominant reliable cue for target detection.

The latter element appeared to be another additional cue which was powerful but not natural in the fruition of 2D compared to 3D information for navigation. Accordingly to aforementioned strategies, personalization played a critical role especially in 3D scenes in target proximity. It is worthwhile to notice that experiment #2 provided some insight about navigation performance with 3D personalized directional cues alone and their dominance nearby target compared to DWM alone (see Fig. 8(c)). One possible explanation of this evidence may be given by recalling the conclusions drawn by Shinn-Cunningham [35], who found that environmental cues distorted directional cues: since our tubular model introduces strong absolute distance cues which possibly overwhelm HRTF information at large source-listener distances, they could in fact make the subjective decision on the direction to choose relatively more problematic for such large distances; conversely, when the target gets closer then the tubular cues become progressively less invasive, hence making the localization process more strongly dependent on the HRTFs and, consequently, on their subjective fit to the listener.

Moreover, observed variations in $M1$ and $M3$ denoted listener-specific differences due to acoustic and non-acoustic factors [34, 2, 27]. 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 (see Fig. 7 for subjective characterization).

From Figures 8(a), 8(b) and 8(c) 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 loudness cues, these two cues have instead comparable salience when they are used in conjunction with binaural information. A closer inspection to experimental result shows significantly lower reaching times in the (L+3Dpers) configuration, that is counterbalanced by significantly shorter final parts of the trajectories in the (DWM+3Dpers) configuration. Finally, the entire trajectories have lengths that are not significantly different in the two configurations.

Table 1 shows a maximum amplitude difference among auditory conditions, reporting higher values for conditions with DWM. The reflectivity properties of both terminations of the acoustic tube act as an additive resonance for the source signal, by raising the average amplitude of the stimulus to about 10 dB RMS. Such an effect may be responsible of the increase of the indicator **M3** in the DWM+3Dpers condition against the control condition L+3Dpers. An informal post-experimental questionnaire reported that participants exploited the higher loudness cues [31] to gain self-awareness of being in the proximity of the target. Accordingly, they tended to decelerate while listening to increases in the higher intensity range: this may be a reason why the L+3Dpers condition performs statistically better in reaching time than DWM+3Dpers.

In spite of the slightly better performance overall 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 if choosing the L+3Dpers option.

## 5 Conclusions & future works

In this paper, sonification of distance with an acoustic tube metaphor based in DWM was proven to be well integrated with binaural audio rendering though headphones without apparent cross-interferences among different types of auditory information. In the proposed tests, the combination of such technologies achieved time and traveled distance performances comparable to

sonification techniques which employ panning and loudness cues. As we said in Section 2, a fundamental design requirement for the distance rendering model consisted of being independent of the source signal. A further proof of this independence may come from repeating the test using different sources, such as vocal and other auditory messages that are typical in these experiments [40]. Possible artifacts arising from the joint use of the tubular model and individualized HRTFs, leading to the performance distortions observed in Figures 6(a) and 6(b), may be solved by employing a larger tube providing more realistic distance cues in the far-field. Moreover, novel personalization procedures, such as ITD optimization [22] and frequency scaling techniques [29] will be taken into account for sonification and they will be evaluated in both static and dynamic scenarios in order to find correspondences between localization accuracy and navigation performances [38].

This, and other experimental activities being necessary to further validate the proposed virtual scenario, are left to future research, particularly when a bigger 3D volume will be available for the experiment. To this regard, we expect to be able to expand the size of the tubular 3D space to realistic volumes, by substituting the DWM with equivalent finite-difference time-domain schemes [24]; the latter in fact allow for more intensive use of efficient data structures, requiring less memory and movement of large signal arrays. Another substantial computational saving and consequent volume increase can be realized by reducing the sampling frequency of the distance rendering model, to levels yet providing acceptable acoustic quality of the interactive stimuli.

Furthermore, once the DWM model implementation will be more computationally efficient, the consequently improved spatial sound rendering architecture will be tested in more complex scenarios involving multiple sound sources displayed together in the auditory scenario. Multimodal virtual environments for spatial data sonification and exploration [17,16], as well as audio rendering in mobile devices and web platforms [18] are expected to substantially benefit from such interactive spatial audio sonification.

**Acknowledgements** The Authors are also grateful to the volunteers who participated to the experiments and to Federico Altieri for his support in data collection. This work was supported by the research project Personal Auditory Displays for Virtual Acoustics, University of Padova, under grant no. CPDA135702.

## References

1. Algazi, V.R., Duda, R.O., Thompson, D.M., Avendano, C.: The CIPIC HRTF database. In: Proc. IEEE Work. Appl. Signal Process., Audio, Acoust., p. 1–4. New Paltz, New York, USA (2001)
2. Andéol, G., Savel, S., Guillaume, A.: Perceptual factors contribute more than acoustical factors to sound localization abilities with virtual sources. Auditory Cognitive Neuroscience **8**, 451 (2015)
3. Asano, F., Suzuki, Y., Sone, T.: Role of spectral cues in median plane localization. The Journal of the Acoustical Society of America **88**(1), 159–168 (1990)
4. Blauert, J.: Spatial Hearing: The Psychophysics of Human Sound Localization. MIT Press, Cambridge, MA, USA (1983)
5. Boren, B., Geronazzo, M., Brinkmann, F., Choueiri, E.: Coloration Metrics for Headphone Equalization. In: Proc. of the 21st Int. Conf. on Auditory Display (ICAD 2015), pp. 29–34. Graz, Austria (2015)
6. Bronkhorst, A.W., Houtgast, T.: Auditory distance perception in rooms. Nature **397**, 517–520 (1999)
7. Campbell, D., Palomaki, K., Brown, G.: A matlab simulation of "shoebox" room acoustics for use in research and teaching. Computing and Information Systems **9**(3), 48 (2005)
8. De Sena, E., Hacihabiboglu, H., Cvetkovic, Z.: Scattering delay network: An interactive reverberator for computer games. In: Audio Engineering Society Conference: 41st International Conference: Audio for Games (2011)
9. Devallez, D., Fontana, F., Rocchesso, D.: Linearizing auditory distance estimates by means of virtual acoustics. Acta Acustica united with Acustica **94**(6), 813–824 (2008)
10. Fontana, F., Rocchesso, D.: A physics-based approach to the presentation of acoustic depth. In: Proc. Int. Conf. on Auditory Display, pp. 79–82. Boston (MA) (2003)
11. Fontana, F., Rocchesso, D.: Auditory distance perception in an acoustic pipe. ACM Trans. Applied Perception **5**(3), 16:1–16:15 (2008)
12. Fontana, F., Savioja, L., Välimäki, V.: A modified rectangular waveguide mesh structure with interpolated input and output points. In: Proc. Int. Computer Music Conf., pp. 87–90. ICMA, La Habana, Cuba (2001)
13. Gardner, W.G., Martin, K.D.: HRTF measurements of a KEMAR. J. of the Acoustical Society of America **97**(6), 3907–3908 (1995)
14. Geronazzo, M.: Mixed structural models for 3D audio in virtual environments. Ph.D. thesis, Information Engineering, Padova (2014)
15. Geronazzo, M., Avanzini, F., Fontana, F.: 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), pp. 73–80. Graz, Austria (2015)
16. Geronazzo, M., Bedin, A., Brayda, L., Avanzini, F.: Multimodal exploration of virtual objects with a spatialized anchor sound. In: Proc. 55th Int. Conf. Audio Eng. Society, Spatial Audio, pp. 1–8. Helsinki, Finland (2014)
17. Geronazzo, M., Bedin, A., Brayda, L., Campus, C., Avanzini, F.: Interactive spatial sonification for non-visual exploration of virtual maps. Int. Journal of Human-Computer Studies, in press (2015)
18. Geronazzo, M., Kleimola, J., Majdak, P.: Personalization support for binaural headphone reproduction in web browsers. In: Proc. 1st Web Audio Conference. Paris, France (2015)

19. Geronazzo, M., Spagnol, S., Bedin, A., Avanzini, F.: 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), pp. 4496–4500. Florence, Italy (2014)
20. Huopaniemi, J., Savioja, L., Karjalainen, M.: 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, pp. 19–22. IEEE, New Paltz (NY) (1997)
21. Iida, K., Ishii, Y., Nishioka, S.: Personalization of head-related transfer functions in the median plane based on the anthropometry of the listener's pinnae. J. of the Acoustical Society of America **136**(1), 317–333 (2014)
22. Katz, B.F., Noisternig, M.: A comparative study of interaural time delay estimation methods. The Journal of the Acoustical Society of America **135**(6), 3530–3540 (2014)
23. Katz, B.F.G., Parseihian, G.: Perceptually based head-related transfer function database optimization. J. of the Acoustical Society of America **131**(2), EL99–EL105 (2012)
24. Kowalczyk, K., van Walstijn, M.: Formulation of Locally Reacting Surfaces in FDTD/K-DWM Modelling of Acoustic Spaces. Acta Acustica united with Acustica **94**(6), 891–906 (2008)
25. Lu, Y.C., Cooke, M., Christensen, H.: Active binaural distance estimation for dynamic sources. In: Proc. INTERSPEECH, pp. 574–577. Antwerp, Belgium (2007)
26. Magnusson, C., Danielsson, H., Rassmus-Gröhn, K.: Non visual haptic audio tools for virtual environments. In: D. McGookin, S. Brewster (eds.) Haptic and Audio Interaction Design, no. 4129 in Lecture Notes in Computer Science, pp. 111–120. Springer Berlin Heidelberg (2006)
27. Majdak, P., Baumgartner, R., Laback, B.: Acoustic and non-acoustic factors in modeling listener-specific performance of sagittal-plane sound localization. Frontiers in Psychology **5** (2014)
28. Masiero, B., Fels, J.: Perceptually robust headphone equalization for binaural reproduction. In: Audio Engineering Society Convention 130 (2011)
29. Middlebrooks, J.C.: Virtual localization improved by scaling nonindividualized external-ear transfer functions in frequency. The Journal of the Acoustical Society of America **106**(3), 1493–1510 (1999)
30. Møller, H., Sørensen, M., Friis, J., Clemen, B., Hammershøi, D.: Binaural technique: Do we need individual recordings? J. Audio Eng. Soc **44**(6), 451–469 (1996)
31. Moore, B.C., Glasberg, B.R., Baer, T.: A model for the prediction of thresholds, loudness, and partial loudness. J. of the Audio Engineering Society **45**(4), 224–240 (1997)
32. Neuhoff, J.G.: An adaptive bias in the perception of looming auditory motion. Ecological Psychology **13**(2), 87–110 (2001)
33. Parseihian, G., Katz, B., Conan, S.: Sound effect metaphors for near field distance sonification. In: Proc. Int. Conf. on Auditory Display, pp. 6–13. Atlanta, GE (2012)
34. Schönstein David; Katz, B.F.G.: Variability in perceptual evaluation of HRTFs. In: Audio Engineering Society Convention 128 (2010)
35. Shinn-Cunningham, B.: Learning reverberation: Considerations for spatial auditory displays. In: Proc. Int. Conf. Auditory Display(ICAD'00). Atlanta (2000)
36. Spagnol, S., Geronazzo, M., Avanzini, F.: On the relation between pinna reflection patterns and head-related transfer function features. IEEE Trans. Audio Speech Lang. Process. **21**(3), 508–519 (2013)
37. Speigle, J., Loomis, J.: Auditory distance perception by translating observers. In: Virtual Reality, 1993. Proceedings., IEEE 1993 Symposium on Research Frontiers in, pp. 92–99 (1993)
38. Viaud-Delmon, I., Warusfel, O.: From ear to body: the auditory-motor loop in spatial cognition. Front. Neurosci **8**, 283 (2014)
39. Wiener, J.M., Büchner, S.J., Hölscher, C.: Taxonomy of human wayfinding tasks: A knowledge-based approach. Spatial Cognition & Computation **9**(2), 152–165 (2009). DOI 10.1080/13875860902906496
40. Zahorik, P.: Assessing auditory distance perception using virtual acoustics. J. of the Acoustical Society of America **111**(4), 1832–1846 (2002)
41. Zahorik, P.: Auditory display of sound source distance. In: Proc. Int. Conf. on Auditory Display. Kyoto, Japan (2002)
42. Zahorik, P.: Direct-to-reverberant energy ratio sensitivity. J. of the Acoustical Society of America **112**(5), 2110–2117 (2002)
43. Zahorik, P., Brungart, D.S., Bronkhorst, A.W.: Auditory distance perception in humans: A summary of past and present research. Acta Acustica united with Acustica **91**(3), 409–420 (2005)