USE OF PERSONALIZED BINAURAL AUDIO AND INTERACTIVE DISTANCE CUES IN AN AUDITORY GOAL-REACHING TASK

Michele Geronazzo, Federico Avanzini

Department of Information Engineering
University of Padova
via Gradenigo 6/B, 35131-Padova, Italy
{geronazzo, avanzini}@dei.unipd.it

Federico Fontana

Department of Mathematics and Computer Science
University of Udine
via delle Scienze 206, 33100-Udine, Italy
federico.fontana@uniud.it

ABSTRACT

While the angular spatialization of source 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 the 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 angular spatialization 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 goal-reaching experiment assessed the proposed display: participants were asked to explore a virtual map with a pen tablet and reach a sound source (the target) using only auditory information; then, subjective time to reach and traveled distance were analyzed. Results suggest that participants achieved a first level of spatial knowledge, i.e., knowledge about a point in space, by performing comparably to when they relied on more robust, although relative, loudness cues. Further work is needed to add fully physical consistency to the proposed auditory display.

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 [1]. Besides this dependence, a broad variability of the distance evaluations across subjects has been observed in most of the tests [2]; 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 τ, already noted by Spiegle and Loomis as salient cues for the positional recognition in a moving listener and source scenario [3, 4]. Perhaps more importantly, moving sound sources evoke so-called "looming" effects which bias their distance perception even if their auditory recognition is not ecological, such as that elicited by the sound of an approaching wild animal and so on [5].

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 [6]. Such guidelines would probably become even more challenging if moving sources were accounted for. Though, the mentioned scarcity of experimental results makes the design of dynamic, especially interactive distance rendering models still a matter of craft.

Near-field distance has been sonified using auditory metaphors, too [7]: 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 those cues which characterize

auditory distance: 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 take relative care about potential interferences with 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 [8]; 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 [9]. 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 monoural 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 [10]; this map was in good accordance with the prediction model proposed by Bronkhorst and Houtgast [11], 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 [12].

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 [13, 14]; 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 [15].

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 [6]: 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.

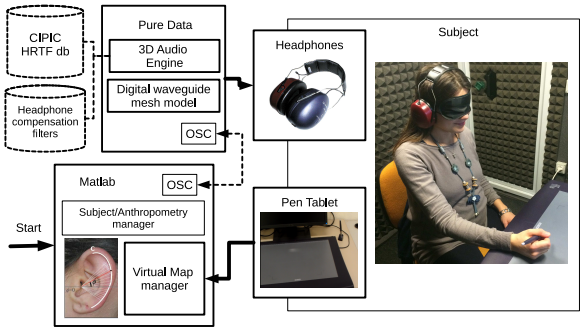


Figure 1: A schematic view of the system architecture.

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 experiment aimed at assessing the validity of the proposed approach: the experiment consists of a goal-reaching task, in which subjects have to explore a virtual map through a stylus on a tablet, and to reach a target point (a sound source in the map) using auditory information in order to reach a first level of spatial knowledge, i.e. knowledge about a point in space [16]. The adopted rendering approach corresponds to an egocentric view of the virtual map in which the pointer corresponds to the listener's head following the "ears in hand" metaphor (ecological rendering) [17]. Experimental results are analyzed and discussed in Section 4, and show that participants using this display achieved a first level of spatial knowledge by performing comparably to when they relied on individualized directional plus loudness cues. This result is particularly interesting, considered the greater robustness of loudness compared to absolute cues of distance such as direct-to-reverberant energy ratio and spectrum.

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 prop-

erties 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 [18]) for any possible listener. However, individual anthropometric features of the human body heavily affect the perception and the quality of the rendering [19]. 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 [20, 21].

In this paper, an image-based HRTF selection technique is briefly summarized (see [22] 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 [23]. 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 ϕ 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 [23]. 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 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.

2.2. DWM

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]. As its optimization would have required labor that was not available at the time when this research was made, 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.

[1]http://puredata.info

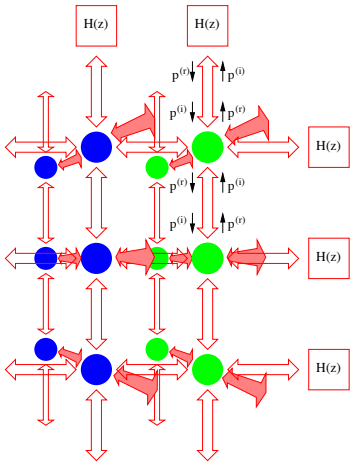
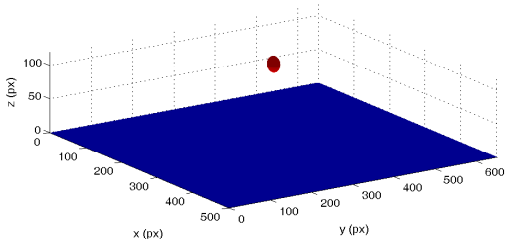


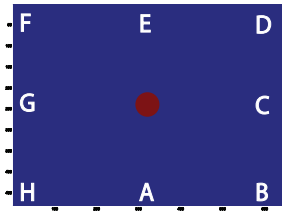
Figure 2: Particular of the 3D DWM: scattering junctions and boundary filters.

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 [24]. 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. Besides 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 [25].



(a)



(b)

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

3. EXPERIMENT: GOAL REACHING

The main goal of this experiment was to assess the validity of the proposed rendering metaphors, the "ears in hand" metaphor for direction and the "acoustic tube" metaphor for distance. One second goal was to analyze the differences and the complementarity of these auditory information, by means of behavioral and performance indicators collected from experimental data. Such assessment were obtained through a goal-reaching task, in which participants had to reach a virtual sound source under different auditory feedback conditions spatially rendered via headphones according to user position in the workspace of a pen tablet.

Six participants (4 males and 2 females whose age varied from 26 to 41 with mean 30.8, SD 5.9) took part at the experiment. 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 12×18 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 Sennheiser HDA 200 headphones. These headphones provide

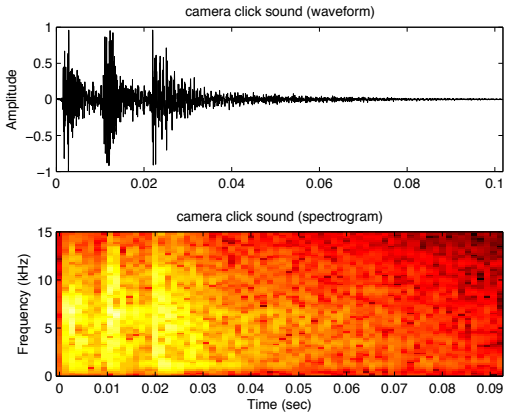


Figure 4: Waveform and spectrogram of the camera click.

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 almost independent of re-positionings on the users' head [26]. Equalization filters based on measurements with KEMAR without pinnae were applied to the auditory stimuli. This non-individualized compensation on regular and stable frequency responses guaranteed no corruption of localization cues in HRTFs [27], as well as an effective equalization of the headphones up to $8 - 10$ kHz on average, simulating a realistic application scenario where it is not always feasible to design individualized headphone compensation filters [26].

3.2. 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 equals to 25 pixels. The sphere was elevated by 120 pixels from the virtual ground 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 65 dB(A). The period between subsequent clicks was large enough to contain 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.

The procedure described in Section 2 drove the selection of best-matched HRTF set. Accordingly, one pinna image for each participant was required in order to compute the mismatch between his/her manually traced contours and notch central frequencies. The source HRTF database was the CIPIC [28], which contains HRTF sets measured in the far field (i.e., no distance infor-

[2] Topological properties of the virtual map were chosen in order to ensure detectable elevation cues from the HRTF selection procedure (see Sec. 2.1). Whereas sphere radius guaranteed a wide dynamic range for loudness control.

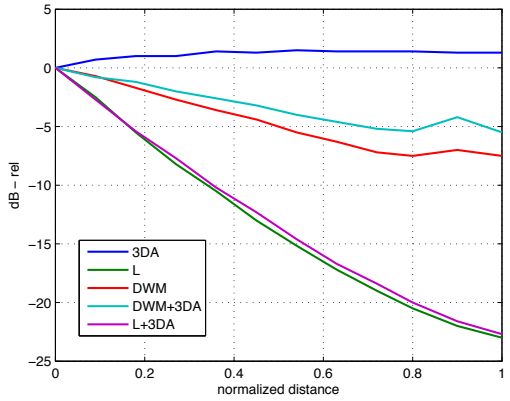


Figure 5: Average amplitude of the stimuli used in the respective experimental conditions as a function of normalized distance. Amplitude values ranging from the smallest (normalized value equal to 0) to the largest (normalized value equal to 1, corresponding to position "A" in Figure 3.b) egocentric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

mation is available in these far-field compensated HRTFs) for 45 subjects, with azimuth angles spanning the range $[0°, 360°)$ and elevation $[-45°, 230.625°]$.

On top of the HRTFs, rendering the angular position (azimuth and elevation) inside the stimuli, 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:

1. HRTF directional cues only (3DA);
2. 6-dB law only (L);
3. tubular shape only (DWM);
4. tubular shape and HRTF directional cues (DWM+3DA);
5. 6-dB law and HRTF directional cues (L+3DA).

Auditory conditions 3DA, L and L+3DA were used for control purposes. In particular, 3DA provided only directional cues, L provided only intensity cue, and the combination of L+3DA 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.

	3DA	L	DWM	DWM+3DA	L+3DA
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. HRTFs from KEMAR [18] were taken as reference for 3DA rendering.

From these measurements it can be noted that loudness under conditions DWM and DWM+3DA changed when the virtual source was moved nearby the auditory target, but not when it was kept moving in the far-field. Moreover DWM+3DA produced higher loudness values than DWM alone, showing an interaction between HRFT resonances and the tubular model. Finally, loudness in condition 3DA 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 [29].

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 subject 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. Results

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

- **M1** absolute reaching time: the time spent by the subject 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 the present experiment trajectories had greater variability, and **M1** with **M2** are assumed to be more appropriate global indicator.

A Kruskal Wallis nonparametric one-way ANOVA with five levels of feedback condition was performed to asses the statistical significance of **M1** [$\chi^2(4, 94262.04)$=78.23, $p << 0.0001$]. Pairwise *post-hoc* Wilcoxon tests (Figure 6(a)) revealed statistically significant improvements in performance (decreases in reaching times) between 3DA and L, DWM+3DA, L+3DA (all with $p << 0.001$), between L and L+3DA ($p < 0.05$), between DWM and DWM+3DA ($p < 0.001$), between DWM and L, L+3DA (all with $p << 0.001$), between DWM+3DA and L+3DA ($p < 0.001$). These results suggest that 3DA/DWM alone performed worse than all the other auditory conditions except in DWM/3DA alone, while their combination (DWM+3DA) had worse performance than L+3DA (the best condition), only. It has to be noticed that degree of statistical significance is very high with the exception of

L and L+3DA comparison. On the other hand no statistical significance was found between 3DA and DWM ($p = 0.163$), L and DWM+3DA ($p = 0.706$).

Again, a Kruskal Wallis nonparametric one-way ANOVA with five levels of feedback condition was performed to asses the statistical significance of **M2** [$\chi^2(4, 93924.4)$=77.95, $p << 0.0001$]. In Figure 6(b), statistical significances are computed using pairwise *post-hoc* Wilcoxon test. Decreases in total traveled distance were reported for following condition pairs: 3DA and L ($p < 0.05$), 3DA and DWM+3DA ($p << 0.001$), 3DA and L+3DA ($p << 0.001$), L and L+3DA ($p < 0.001$), L and DWM+3DA ($p < 0.05$), DWM and L ($p << 0.001$), DWM and DWM+3DA ($p << 0.001$), DWM and L+3DA ($p << 0.001$). On the other hand, no statistical differences were found between 3DA and DWM ($p = 0.181$), and DWM+3DA and L+3DA ($p = 0.320$). Conditions 3DA and DWM poorly performed in terms of **M2** if they were rendered individually, while results suggest their strong integration leading to similar performance with respect to L+3DA.

A further analysis was performed on **M3**, i.e. final traveled distance, in order to assess auditory spatial awareness of the user near the target [16]. A Kruskal Wallis nonparametric one-way ANOVA with five levels of feedback condition was performed to asses the statistical significance of **M3** [$\chi^2(4, 21396.7)$=17.76, $p < 0.01$]. Pairwise *post-hoc* Wilcoxon tests revealed the following decreases in the final traveled distance: DWM and 3DA ($p < 0.05$), DWM and DWM+3DA ($p < 0.05$), L+3DA and L,3DA (both $p < 0.05$), and L+3DA and DWM+3DA ($p < 0.001$). No statistical significant effects were found in pairs: 3DA and L ($p = 0.418$), 3DA and DMW+3DA ($p = 0.439$), L and DWM ($p = 0.087$), L and DWM+3DA ($p = 0.076$), and DWM and L+3DA ($p = 0.904$). The impact of directional rendering in **M3** suggested a robust integration with DWM which will be discussed in the following section.

4. DISCUSSION

From Figures 6(a), 6(b) and 6(c) it appears that the joint adoption of individualized HRTFs and DWM model (DWM+3DA) leads to subjective performances that are comparable to when the individualized HRTFs and loudness model (L+3DA) work in synergy. This result is surprising once one notices that, as expected, listeners perform much better if using loudness (L) as opposed to the tube model (DWM) alone once they are deprived of individualized directional cues. This evidence suggests that while the addition of absolute distance cues in our source sound is of relatively little help for the reaching task compared to adding loudness cues, conversely these two cues have similar strength once used in connection with binaural information. A deeper inspection shows significantly lower reaching times in the (L+3DA) configuration, that is counterbalanced by significantly shorter final parts of the trajectories in the (DWM+3DA) 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+3DA condition against the control condition L+3DA. An informal post-experimental questionnaire reported that participants exploited the higher loudness cues [30] to gain self-awareness of being in the proximity of the target. Accordingly, they tended to decelerate while listening to increases in the higher loudness range: this may be a reason why the L+3DA condition performs statistically better in reaching time than DWM+3DA.

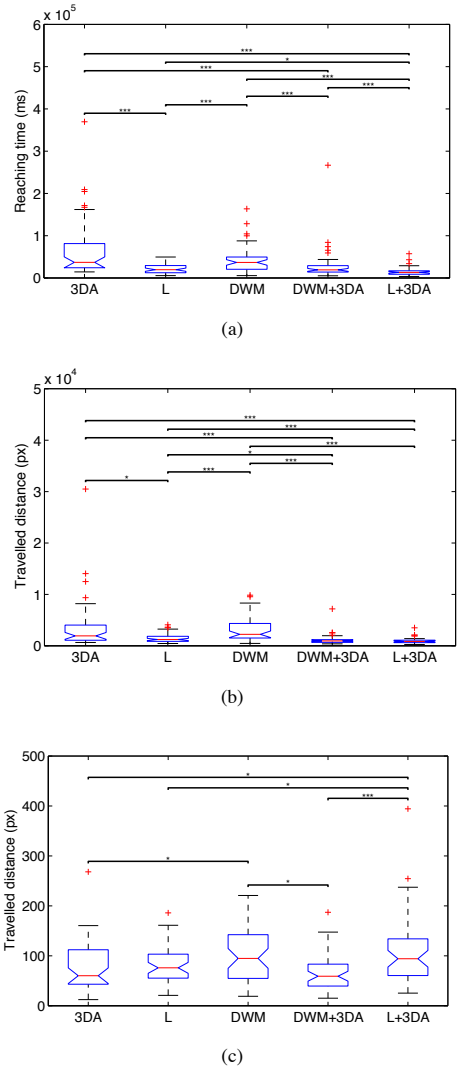


Figure 6: Global statistics on (a) reaching times, (b) total traveled distance, and (c) "final" traveled distance, grouped by feedback condition.

In spite of the slightly better performance overall shown by the L+3DA over the DWM+3DA 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, however at greater computational cost than if choosing the L+3DA 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 noticeable cross-interferences among different types of auditory information. In the proposed experiment, the combination of such technologies achieved time and traveled distance performances comparable to sonification techniques which employ 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 for these experiments [2].

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 through additional software programming activity 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; 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 in order to validate interactions among multiple virtual acoustic tubes. Multimodal virtual environments for spatial data sonification and exploration [29, 31], as well as audio rendering in mobile devices and web platforms [32] are expected to substantially benefit from such interactive spatial audio sonification.

6. ACKNOWLEDGMENT

This work was supported by the research project Personal Auditory Displays for Virtual Acoustics, University of Padova, under grant no. CPDA135702. The Authors are also grateful to the volunteers who participated to the listening test.

7. REFERENCES

[1] P. Zahorik, D. S. Brungart, and A. W. Bronkhorst, "Auditory distance perception in humans: A summary of past and present research," *Acta Acustica united with Acustica*, vol. 91, no. 3, pp. 409–420, 2005.

[2] P. Zahorik, "Assessing auditory distance perception using virtual acoustics," *J. of the Acoustical Society of America*, vol. 111, no. 4, pp. 1832–1846, Apr. 2002.

[3] Y.-C. Lu, M. Cooke, and H. Christensen, "Active binaural distance estimation for dynamic sources." in *Proc. INTERSPEECH*, Antwerp, Belgium, Aug. 27-31 2007, pp. 574–577.

[4] J. Speigle and J. Loomis, "Auditory distance perception by translating observers," in *Virtual Reality, 1993. Proceedings., IEEE 1993 Symposium on Research Frontiers in*, Oct 1993, pp. 92–99.

[5] J. G. Neuhoff, "An adaptive bias in the perception of looming auditory motion," *Ecological Psychology*, vol. 13, no. 2, pp. 87–110, 2001.

[6] P. Zahorik, "Auditory display of sound source distance," in *Proc. Int. Conf. on Auditory Display*, Kyoto, Japan, July 2002.

[7] G. Parseihian, B. Katz, and S. Conan, "Sound effect metaphors for near field distance sonification," in *Proc. Int. Conf. on Auditory Display*, Atlanta, GE, Jun. 18-21 2012, pp. 6–13.

[8] P. Zahorik, "Direct-to-reverberant energy ratio sensitivity," *J. of the Acoustical Society of America*, vol. 112, no. 5, pp. 2110–2117, Nov. 2002.

[9] B. Shinn-Cunningham, "Learning reverberation: Considerations for spatial auditory displays," in *Proc. Int. Conf. Auditory Display(ICAD'00)*, Atlanta, 2000.

[10] F. Fontana and D. Rocchesso, "Auditory distance perception in an acoustic pipe," *ACM Trans. Applied Perception*, vol. 5, no. 3, pp. 16:1–16:15, 2008.

[11] A. W. Bronkhorst and T. Houtgast, "Auditory distance perception in rooms," *Nature*, vol. 397, pp. 517–520, Feb. 1999.

[12] D. Devallez, F. Fontana, and D. Rocchesso, "Linearizing auditory distance estimates by means of virtual acoustics," *Acta Acustica united with Acustica*, vol. 94, no. 6, pp. 813–824, 2008.

[13] F. Fontana and D. Rocchesso, "A physics-based approach to the presentation of acoustic depth," in *Proc. Int. Conf. on Auditory Display*, Boston (MA), June 2003, pp. 79–82.

[14] D. Campbell, K. Palomaki, and G. Brown, "A matlab simulation of "shoebox" room acoustics for use in research and teaching," *Computing and Information Systems*, vol. 9, no. 3, p. 48, 2005.

[15] E. De Sena, H. Hacihabiboglu, and Z. Cvetkovic, "Scattering delay network: An interactive reverberator for computer games," in *Audio Engineering Society Conference: 41st International Conference: Audio for Games*, Feb 2011.

[16] J. M. Wiener, S. J. Büchner, and C. Hölscher, "Taxonomy of human wayfinding tasks: A knowledge-based approach," *Spatial Cognition & Computation*, vol. 9, no. 2, pp. 152–165, May 2009.

[17] C. Magnusson, H. Danielsson, and K. Rassmus-Gröhn, "Non visual haptic audio tools for virtual environments," in *Haptic and Audio Interaction Design*, ser. Lecture Notes in Computer Science, D. McGookin and S. Brewster, Eds. Springer Berlin Heidelberg, Jan. 2006, no. 4129, pp. 111–120.

[18] W. G. Gardner and K. D. Martin, "HRTF measurements of a KEMAR," *J. of the Acoustical Society of America*, vol. 97, no. 6, p. 3907–3908, June 1995.

[19] H. Møller, M. Sørensen, J. Friis, B. Clemen, and D. Hammershøi, "Binaural technique: Do we need individual recordings?" *J. Audio Eng. Soc*, vol. 44, no. 6, p. 451–469, 1996.

[20] K. Iida, Y. Ishii, and S. Nishioka, "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*, vol. 136, no. 1, pp. 317–333, July 2014.

[21] B. F. G. Katz and G. Parseihian, "Perceptually based head-related transfer function database optimization," *J. of the Acoustical Society of America*, vol. 131, no. 2, p. EL99–EL105, Feb. 2012.

[22] M. Geronazzo, S. Spagnol, A. Bedin, and F. Avanzini, "Enhancing vertical localization with image-guided selection of non-individual head-related transfer functions." in *Proc. IEEE Int. Conf. Acoust. Speech and Signal Process. (ICASSP 2014)*, Firenze, May 2014, pp. pages 4496–4500.

[23] S. Spagnol, M. Geronazzo, and F. Avanzini, "On the relation between pinna reflection patterns and head-related transfer function features," *IEEE Trans. Audio Speech Lang. Process.*, vol. 21, no. 3, pp. 508–519, Mar. 2013.

[24] J. Huopaniemi, L. Savioja, and M. Karjalainen, "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*. New Paltz (NY): IEEE, Oct. 1997, pp. 19–22.

[25] F. Fontana, L. Savioja, and V. Välimäki, "A modified rectangular waveguide mesh structure with interpolated input and output points," in *Proc. Int. Computer Music Conf.* La Habana, Cuba: ICMA, Sept. 2001, pp. 87–90.

[26] M. Geronazzo, "Mixed structural models for 3D audio in virtual environments," Ph.D. dissertation, Information Engineering, Padova, Apr. 2014.

[27] B. Masiero and J. Fels, "Perceptually robust headphone equalization for binaural reproduction," in *Audio Engineering Society Convention 130*, 2011.

[28] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database," in *Proc. IEEE Work. Appl. Signal Process., Audio, Acoust.*, New Paltz, New York, USA, Oct. 2001, p. 1–4.

[29] M. Geronazzo, A. Bedin, L. Brayda, C. Campus, and F. Avanzini, "Interactive spatial sonification for non-visual exploration of virtual maps," *Int. Journal of Human-Computer Studies, in press*, 2015.

[30] B. C. Moore, B. R. Glasberg, and T. Baer, "A model for the prediction of thresholds, loudness, and partial loudness," *J. of the Audio Engineering Society*, vol. 45, no. 4, p. 224–240, 1997.

[31] M. Geronazzo, A. Bedin, L. Brayda, and F. Avanzini, "Multimodal exploration of virtual objects with a spatialized anchor sound," in *Proc. 55th Int. Conf. Audio Eng. Society, Spatial Audio*, Helsinki, Finland, Aug. 2014, pp. 1–8.

[32] M. Geronazzo, J. Kleimola, and P. Majdak, "Personalization support for binaural headphone reproduction in web browsers," in *Proc. 1st Web Audio Conference*, Paris, France, Jan. 2015.