Auditory distance presentation in an urban augmented-reality environment

ROBERT ALBRECHT and TAPIO LOKKI, Aalto University

Presenting points of interest in the environment by means of audio-augmented reality offers benefits compared with traditional visual augmented-reality and map-based approaches. However, presentation of distant virtual sound sources is problematic. This study looks at combining well-known auditory distance cues to convey the distance of points of interest. The results indicate that although the provided cues are intuitively mapped to relatively short distances, users can with only little training learn to map these cues to larger distances.

Categories and Subject Descriptors: H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems—*Artificial, augmented, and virtual realities; Audio input/output*; H.5.2 [**Information Interfaces and Presentation**]: User Interfaces—*Auditory (non-speech) feedback*

General Terms: Design, Experimentation, Human Factors

Additional Key Words and Phrases: Auditory display, spatial sound, localization, distance perception

ACM Reference Format:

Robert Albrecht and Tapio Lokki. 2015. Auditory distance presentation in an urban augmented-reality environment. *ACM Trans. Appl. Percept.* 12, 2, Article 5 (March 2015), 19 pages.
DOI:http://dx.doi.org/10.1145/2723568

1. INTRODUCTION

One of the many applications of augmented reality is the presentation of points of interest (POIs) in the environment. A person might walk the streets of an unfamiliar city, and wants to find a place to eat nearby. One option would be to use visual augmented reality browsers on a smartphone [Grubert et al. 2011], which show the location of and information about different restaurants overlayed on the camera image of the phone, as the user points the phone in different directions. The user thus sees a small window of the surrounding world with additional information augmented to it.

Compared with visually augmented reality, an audio-augmented reality approach offers several advantages in this type of application. While the visual field of view is limited and the display and camera of a smartphone introduce further limitations, virtual sound sources may be presented over headphones in any direction at any time. Auditory displays also allow hands-free use and leave the user's vision free for, e.g., navigational tasks.

The research leading to these results has received funding from Nokia Research Center.

Authors' address: Aalto University, Department of Computer Science, P.O. Box 15500, FI-00076 AALTO, FINLAND; email: robert.albrecht@aalto.fi and tapio.lokki@aalto.fi.

DOI:http://dx.doi.org/10.1145/2723568

Both visual and auditory augmented reality solutions can be used to convey various information about surrounding POIs. This information might include names and short descriptions of the POIs together with the directions and distances to the POIs. Both visually and with sound, distance may be presented simply by specifying the distance in, e.g., meters or kilometers. However, as the number of POIs that are presented increases, presenting the distance of every POI in meters will unnecessarily increase visual or auditory clutter.

Ideally, the distance of POIs should therefore be presented in an intuitive way that takes advantage of our natural ability to perceive depth and distances. With vision, cues such as relative size and height in the visual field may be used, and stereoscopic displays offer further cues through binocular disparities [Cutting and Vishton 1995].

Several factors affect the perceived distance of a sound source, including loudness and reverberation [Zahorik et al. 2005]. For speech sources, the type of speech together with the production level also has a large impact [Brungart and Scott 2001]. However, although the perceived distance of virtual sound sources can effectively be altered through modifications of these cues, presenting sources at larger distances is difficult. For example, a human voice can normally not be heard from several hundred meters away. This is especially challenging in an urban environment, where faint, distant sound sources cannot be heard over traffic and other background noise.

This study investigates the presentation of POIs in a real urban environment using sound. We organized a user study where we presented virtual sound sources representing imaginary POIs in the environment over headphones. We modified different features of the sound sources affecting their perceived distance, and asked test participants where they thought the associated POIs would be. This was compared with having a voice saying the distance to the POIs in meters. We hypothesized that the participants intuitively would locate the POIs at relatively short distances when performing the task based on the provided auditory distance cues, as opposed to knowing the distance in meters. Since we wanted to be able to convey the distance of POIs further away, we also included training with visual feedback to see if participants quickly could learn to map the available cues to the intended POI locations.

This article is organized as follows. In Section 2, we look at research related to auditory distance perception, pedestrian navigation and presentation of POIs with audio. In Section 3, the design and implementation of the user study is presented. The results of the study are presented in Section 4, together with comments from the participants in Section 5. The results are further discussed in Section 6, with conclusions in Section 7.

2. RELATED RESEARCH

2.1 Auditory distance perception

Several studies on auditory distance perception have shown that listeners typically underestimate the distance of faraway sound sources while they overestimate the distance of nearby sources (closer than approximately 1 m). A compressive power function of the form $r' = kr^a$ has been suggested to model this behaviour [Zahorik et al. 2005]. Here, r is the physical source distance and r' an estimate of the perceived distance with fit parameters k and a. Zahorik et al. gathered data from 21 different studies and fit them to this power function, producing mean values of $\bar{k} = 1.32$ and $\bar{a} = 0.54$. Using these values as parameters, a physical source distance of 100 m would result in a perceived distance estimate of only 16 m, while a source distance of 1 000 m would result in an estimate of only 55 m. However, the actual data did not include such large distances.

In fact, little research has been done on auditory distance perception at large distances, say hundreds of meters. Fluitt et al. [2014] studied auditory distance estimation in an open grassy field, with physical source distances in the range of 25 to 800 m. In this study, a physical source distance of 100 m gave a median distance estimate of 63 m and a source distance of 800 m a median estimate of 189 m.

The use of a near-field auditory display where farther distances are mapped to the near-field distances representable with binaural distance cues has priorly been suggested [Brungart 2002]. This mapping requires learning by the users, but allows representation of different distances with robust and absolute cues that do not degrade the comprehensibility of the presented message. This method is, however, limited by the fact that binaural distance cues are not available for sources in the median plane.

When the sound source is familiar, listeners can effectively use production and presentation level together as distance cues. For example, with speech sources, the perceived distance can be controlled over a wide range by adjusting these levels appropriately [Brungart and Scott 2001]. The type of speech has a large impact on the effects that these adjustments have. With whispered speech, neither production nor presentation level has a large impact. The sound source is always perceived very close to the listener. With low-level voiced speech, the production level affects the distance, as does the presentation level as long as it exceeds 72 dB. With high-level voiced speech, both production and presentation level have a large impact on the perceived distance.

2.2 Use of audio for pedestrian exploration and navigation

Holland et al. [2002] created a prototype spatial audio navigation interface to help people carrying out location tasks with their eyes, hands or attention otherwise engaged. Simple stereo panning was used to indicate the direction to the next waypoint. To help distinguish between sound sources in the front of and behind the listener, different sound samples were used. When the source was in the front, a sharp tone was played, while a muffled tone was played when the source was in the back. A so called chase tone was also utilized to give feedback about the user's direction of walking compared with the direction to the next waypoint. The larger the deviation from the correct direction, the more the chase tone differed in pitch from the main navigation tone. If the direction was spot on, the pitch was the same for both tones.

To represent distance, Holland et al. used a Geiger counter metaphor, common in car reversing monitors. If the distance to the next waypoint was large, temporally widely spaced sound pulses were emitted. As the distance decreased, the spacing of the pulses became denser. An arrival tone indicated that the user had reached the destination.

Tran et al. [2000] investigated the use of different acoustic beacons for navigational tasks and suggested four characteristics for good beacon sounds: 1) they must be easy to localize and follow, 2) they should be easily distinguished from other sounds in the environment, 3) they should not easily be masked by noise and other sounds, and 4) they should not distract or annoy the user. In the experiments, relatively wide-band sounds were found to provide the best localization. Speech sounds, however, were found not to be suitable as navigation beacons. Good localization performance correlated with good subjective quality ratings by the participants. Approximately 1–2 repetitions per second was suggested as an optimal rate for presenting beacon sounds.

McGookin and Brewster [2012] studied the use of spatial audio to support pedestrian navigation to physical landmarks. Participants were asked to locate physical landmarks in a botanical garden using a map and assisting acoustic navigation beacons. Of the different conditions tested, participants favoured enabling the auditory beacon at 30 m from the landmark, rather than having it enabled at all distances or only at shorter distances. 30 m was seen as a good distance to enable audio, which helped the participants to confirm the correct direction of the landmark. However, the performance of the participants did not depend significantly on these conditions. In their experiments, McGookin and Brewster found that participants in many cases searched for the landmarks in the wrong location, because of misleading audio feedback caused by inaccurate GPS data.

Pedestrian navigation and exploration through audio interfaces is especially useful for the visually impaired. For example, the In Situ Audio Services (ISAS) application presents POIs in the environment to blind users with a regular smartphone [Blum et al. 2013]. POIs are presented in two modes: shockwave and radar. In shockwave mode, nearby POIs are presented in order of distance, with the nearest POI first. In radar mode, POIs are presented based on their direction, in clockwise order. Through this mechanism, the radar mode has the advantage of providing additional cues to the direction of the POIs, while, e.g., front-back confusions may introduce difficulties in estimating the direction of POIs in shockwave mode.

3. USER STUDY

A user study was devised to answer the following main research questions:

—If we present virtual sound sources representing POIs in an urban audio-augmented reality environment, where in the environment do people locate these?

—If the virtual sound sources are localized at distances shorter than those we want to present, can we easily teach people to map these perceived distances to longer distances?

—How does providing auditory distance cues compare with the alternative of saying the distance in meters?

In addition, the following secondary research questions were considered:

—Does localization performance improve with repeated presentation of sounds compared with a single presentation?

—How does localization of completely artificial sounds (noise bursts) compare with speech?

—Can we provide additional cues that certain POIs are obstructed by buildings?

In short, the user study participants sat in an urban outdoor environment, where they listened to sounds representing different POIs over headphones with head tracking, and were asked to specify where they thought the POIs were on a map. The experiment is described in more details in the following sections.

3.1 Location and weather conditions

The tests took place in the centre of Leppävaara, a district in the city of Espoo, Finland. Photographs of the location are shown in Figure 1. The participants sat next to a sidewalk with a wall approximately six meters behind them and a street approximately ten meters in front of them. The speed limit on the street was 40 km/h. The street was normally lightly trafficked, but more traffic was present during rush hours. Due to the small distance to the street, the occasional sound of, e.g., trucks could mask the sounds representing distant POIs quite effectively.

During the tests, the ambient temperature ranged from 0 to +5 °C, with wind speed between 1 and 6 m/s and the sky varying between clear and overcast. The tests took place mostly during daylight. However, some of the tests started or ended during dawn or dusk.

3.2 Apparatus

Participants wore Sennheiser HD 590 open circumaural headphones with a SHAKE SK7 head tracker attached on top. The head tracker provided the orientation of the participant's head through compass, gyroscope, and accelerometer sensors. The virtual sound sources were modified based on this orientation information, so that they maintained their position relative to the surrounding world independent

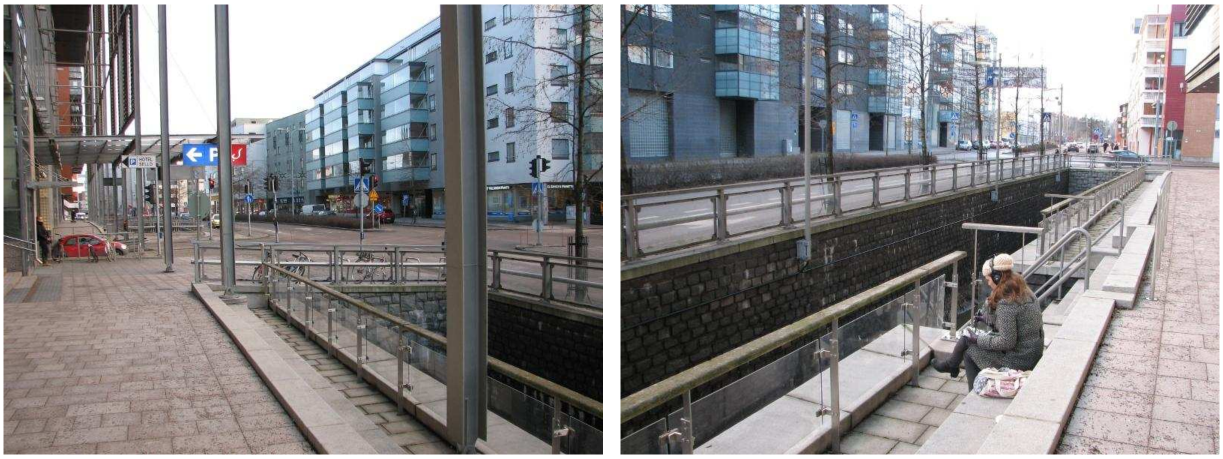


Fig. 1. The location of the user study, with one person participating in the test.

of the participant's head movements. Stimuli were generated by a tablet computer running Pure Data. An aerial photography view (hereafter referred to as a map) of the environment was shown, where participants selected the location where they thought the presented POIs would be. The map was zoomable in 1:2 steps between scales of approximately 1:1150 and 1:18400. The large range of scales was chosen so that it would not give participants any hints about the range of distances over which the POIs may be located.

3.3 Conditions

The test consisted of four main conditions, with two versions of each, resulting in a total of eight different conditions. The four main conditions were:

(1) Speech, single presentation, with distance cues through intensity and early-to-late energy ratio modifications, type of speech as well as obstruction cues.

(2) Speech, repeated presentation, with distance cues through intensity and early-to-late energy ratio modifications, type of speech as well as obstruction cues

(3) Speech, single presentation, with the distance announced in meters as the only distance cue.

(4) Noise bursts, repeated presentation, with distance cues through intensity and early-to-late energy ratio modifications as well as obstruction cues.

The two versions, *a* and *b*, of the main conditions differed in the training phase before the actual test. In version *a*, a set of training POIs was presented before the test to allow the participants to familiarize themselves with the type of POIs that were to be presented in the test. In version *b*, a similar training phase was included, but during the training, the intended location of each POI was shown on the map once the participant had selected the location where he or she thought the POI was located. This type of visual feedback was, however, not given during the actual test.

Conditions 1a–4a were first presented to participants in random order. After this, conditions 1b–4b were presented in random order.

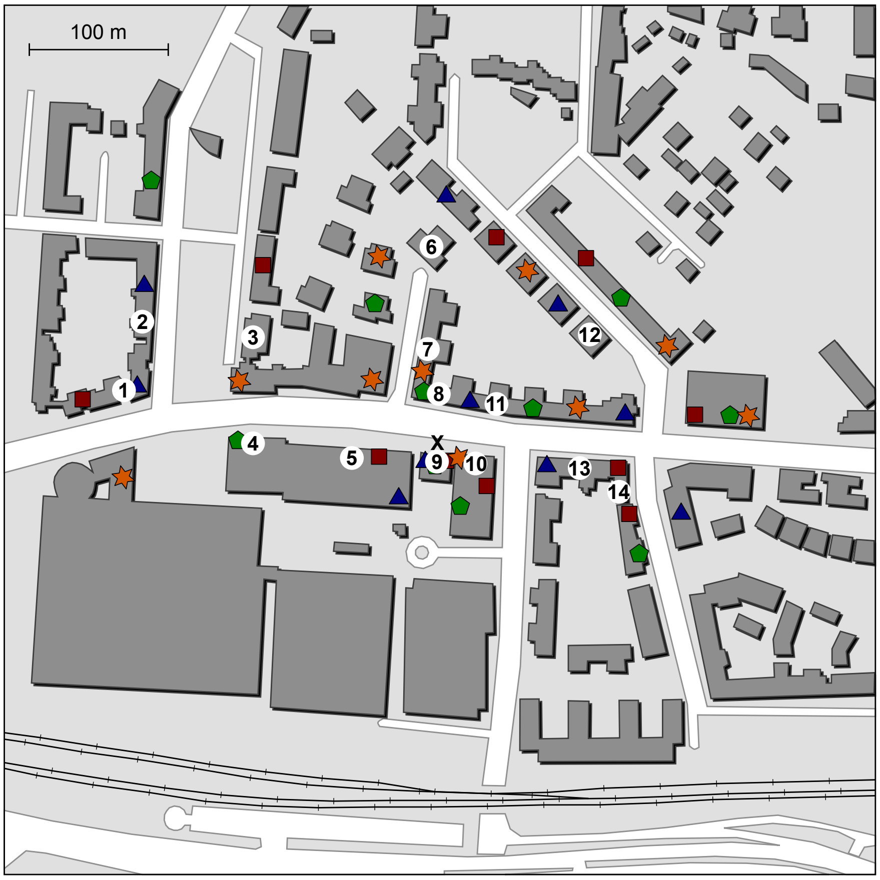


Fig. 2. Locations of the POIs presented in the user study. The POIs are marked with numbers and the location of the participant is marked with a cross. The four sets of training POIs are marked with triangles, rectangles, stars, and pentagons. Note that some training POIs are obscured by POI no. 9.

3.4 POI locations

Under each condition, sounds representing 14 different POIs were presented in random order. The locations of the POIs are shown in Figure 2. The participants sat at the location marked with a cross, facing the street in front of them, but were allowed to move their head and torso. POIs no. 2, 3, 6, 7, 12, and 14 were obstructed by buildings from the point of view of the participants. These obstructed POIs were presented differently than the rest of the POIs under conditions 1, 2, and 4, as described in Section 3.7. The POI locations used in the tests were chosen irrespective of the existence of any real POIs at these locations. The order in which the POIs were presented was varied between participants and conditions. Each POI was presented only once under each condition.

During the training phase before each test, ten different POIs were presented, each POI once. Four sets of ten training POIs, shown in Figure 2, were created and randomly associated with conditions 1a–4a and again with conditions 1b–4b separately for each participant. The training POIs were chosen so that they represented the same types of POI locations and distances as those used in the actual tests, but with differing locations. The order in which the ten POIs were presented was randomized for each training session.

3.5 Stimuli

Speech samples were obtained from Acapela Group's text-to-speech synthesizer[1], which provides convincing synthesized whispering, speaking, and shouting voices. The voices used were *English (USA) WillUpClose, Will,* and *WillFromAfar*. Under conditions 1 and 2, the message "Here's a point of interest" was presented. The duration of the message was 2.0 s for the whispering, 1.5 s for the speaking, and 2.1 s for the shouting voice. Under condition 3, the spoken message "Here's a point of interest at x meters" was presented, where x was the distance to the POI rounded to the nearest ten meters. The duration of this message was 2.6–3.7 s, depending on the distance. Under condition 4, a pattern of four 250-ms long bursts of white noise with 250 ms silence between bursts was presented. Both this pattern of noise bursts under condition 4 and the spoken message under condition 2 were repeated with approximately 1.25 s silence between repetitions until the participant chose the location where he or she thought that the POI was located.

Both to aid in the integration of the virtual sound sources in the surrounding acoustic environment and to provide the means for adjusting their perceived distance through early-to-late energy ratio modifications, binaural room impulse responses (BRIRs) were convolved with the dry speech and noise samples. When preparing the user study, BRIRs were measured in several different outdoor environments. The measurements were done using Philips SHN2500 headphones, which have integrated binaural microphones for noise-cancellation purposes. The sound source used consisted of two flat wooden batons banged against each other. The first author wore the headphones and played the role of the receiver.

Measurements were made in the following different types of environments: in the vicinity of a large building on one side, on a street with buildings on both sides, in an open field, in a moderately dense forest, and with small buildings between the sound source and the receiver. Impulse responses were measured with the receiver facing in four different directions, with the azimuth of the source being 0°, 90°, 180°, and 270°.

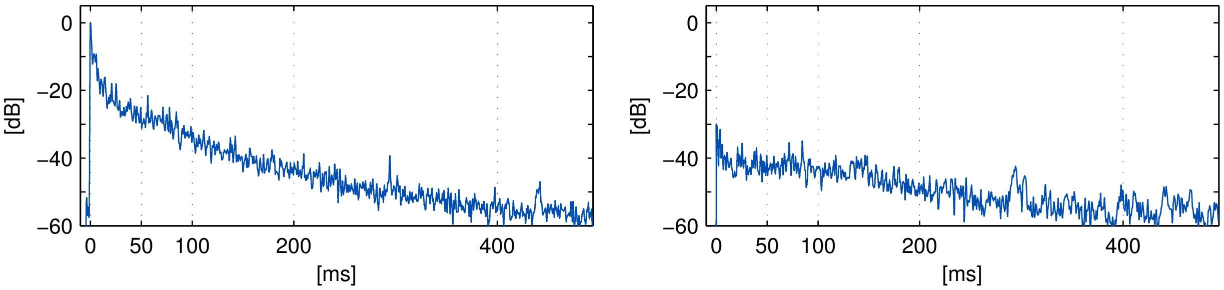
The measured impulse responses contained a substantial amount of background noise (internal mic noise, wind noise, etc.). Therefore, noise suppression was performed by means of spectral subtraction [Boll 1979]. Additionally, the length of the impulse responses was limited to 0.5 s, as later arriving reflections did not exceed the level of the noise floor.

The suitability of the measured impulse responses for presenting virtual speech sources was tested by the first author through informal listening in different environments. The impulse responses measured in a forest with 5 m between the source and the receiver (see Figure 3(a)) proved the best alternative to be used in varying acoustic environments. This is probably due to the fact that the forest setting provided a good amount of diffuse reflections in different directions.

The appropriate BRIR for the presentation of a sound source at a particular azimuth was constructed from the two closest BRIR azimuths by means of interpolation. For example, if the source azimuth ϕ was between 0° and 90°, the gain factors $g_{0°}$ and $g_{90°}$ for the 0° and 90° BRIRs were calculated as follows:

$$g_{0^\circ} = \sqrt{1 - \frac{\phi}{90^\circ}} \tag{1}$$

[1] http://www.acapela-box.com/



(a) Unobstructed sound source.

(b) Obstructed sound source.

Fig. 3. Energy envelope of the binaural room impulse responses used in the study (left ear, 0° source azimuth).

$$g_{90°} = \sqrt{\frac{\phi}{90°}} \tag{2}$$

To provide better directional perception than that achieved by simple interpolation of BRIRs measured in four directions, the direct sound was removed from these BRIRs, and added through head-related impulse responses (HRIRs) instead. The HRIRs of subject no. 12 from the CIPIC HRTF database [Algazi et al. 2001] were used through the CW_binaural˜ Pure Data external [Doukhan and Sédès 2009].

3.6 Distance cues

Intensity and early-to-late energy ratio modifications [Albrecht and Lokki 2013] were used to alter the perceived distance of the virtual sound sources. The direct sound, i.e., the HRIRs, were first attenuated according to the distance to the source. As realistic attenuation of the direct sound would have caused the virtual sound sources to be inaudible even at moderately large distances, a much lesser degree of attenuation was chosen to provide audible virtual sound sources over the desired range of distances. To provide the appropriate attenuation, a gain factor of

$$g_d = \frac{1}{d/30\,m},\ g_d \leq 1 \tag{3}$$

was applied to the direct sound, where d is the distance to the sound source.

Early reflections of the BRIRs were attenuated accordingly, with the attenuation in dB linearly reduced from its full value at the beginning of the impulse response to 0 dB at 100 ms after the beginning of the impulse response. An example room impulse response with 6 dB and 12 dB attenuations of the direct sound is shown in Figure 4. Note, that if the early reflections were not attenuated together with the direct sound, a 12 dB attenuation of the direct sound would result in the earliest reflections being louder than the direct sound.

Additionally, a gain factor of $g_d^{0.5}$ was applied to both the direct and the reflected sound. This provided distance attenuation also for the reverberation, which would have sounded unnaturally loud at large distances otherwise.

For conditions 1 and 2, the type of speech was used as an additional distance cue, as suggested by Brungart and Scott [2001]. As a realistic mapping between distance and speech type would have resulted in all but the very closest POIs being presented by a shouting voice, a different mapping was instead chosen. For sources closer than 20 m, a whispering voice was used. For sources at a distance

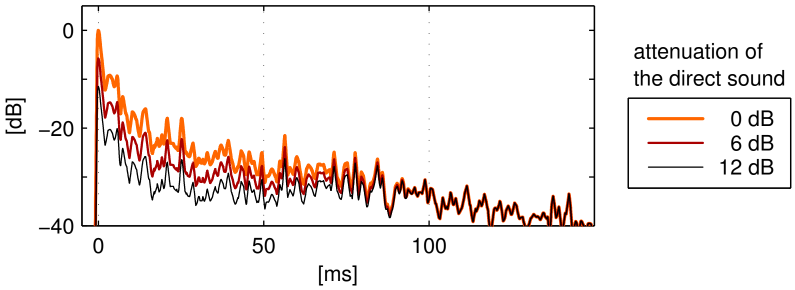


Fig. 4. Modifications of the early-to-late energy ratio of a room impulse response. Here, 6 dB and 12 dB attenuations of the direct sound together with gradually decreasing attenuation of the early reflections are shown together with the unattenuated impulse response.

between 20 and 100 m, a normal conversational voice was used. For sources more than 100 m away, a shouting voice was used.

Under condition 3, no other distance cues were provided than the distance in meters, and this distance was announced with a normal conversational voice. To present this spoken message, BRIRs were used, in addition to HRIRs, to help in providing externalization for the stimulus, and presumably making it more pleasant to listen to, than it would have been using only HRIRs. However, the intensity and the early-to-late energy ratio were not modified according to the POI distance, but kept constant, corresponding to the values used for distances less than or equal to 30 m under conditions 1, 2, and 4.

3.7 Obstruction

Sound sources obstructed by large buildings will, unless the sound power is considerably large, be barely audible or even entirely inaudible. For this reason, realistically simulating these sources in a virtual auditory display is not sensible. In this study, we decided to rather investigate the effects of smaller obstructions and whether these could be applied to provide cues of sound sources obstructed by larger buildings.

The impulse response measurements done in this study included two cases where the sound source was obstructed by a small building. However, these impulse responses had some prominent individual reflections and a poor signal-to-noise ratio, and could for these reasons not be used as such. Instead, the impulse responses that were used for the unobstructed sound sources were modified to include similar cues of obstruction.

The direct sound was attenuated by a factor of 18 (25 dB). Early reflections were attenuated accordingly, with the attenuation factor linearly reduced from its full value at the start of the impulse response to zero at 150 ms from the start of the impulse response. In addition, a first-order Butterworth low-pass filter with a cutoff frequency of 500 Hz was applied. An example of a modified impulse response for presenting obstructed sound sources is shown in Figure 3(b).

Obstructed sound sources were provided with distance cues by applying distance attenuation with the gain factor g_d (Eq. 3). The whole room impulse response was thus equally attenuated based on distance, while the early-to-late energy ratio was kept constant, as the early energy already was heavily attenuated to provide the obstruction cues.

These obstruction cues were applied when presenting the obstructed POIs (no. 2, 3, 6, 7, 12, and 14) under conditions 1, 2, and 4. No obstruction cues were provided under condition 3.

Obstructed sound sources could be presented as heard from the closest unobstructed path (e.g., a street), or simply from the direction of the source. The second case is probably preferable in many applications, since in this case, the actual direction of the source is unambiguous. For this reason, the second approach was chosen in this study.

3.8 Participants and instructions

The age of the participants ranged from 21 to 34 years. Out of the 12 participants, 2 were female and 10 were male. 8 participants were students at Aalto University and 3 of them studying acoustics. The other participants were either working with different technology-related tasks (3) or communications (1). The participants did in general not have experience with this type of experiment. One participant reported having a slight unilateral hearing loss at low frequencies. The other participants were not aware of having any hearing impairments.

Participants were told that they would be presented with different sounds representing POIs in the surroundings, and that they should select the location on the map where they thought the POI corresponding to the presented sound was. They were instructed that they would hear the sounds in the direction of the POIs and that the distance of the POIs would be conveyed either by how distant they sound or by a voice saying the distance in meters, depending on the condition being tested.

Participants were instructed that the POIs might or might not correspond with any real POIs, but that they always are in some part of a building. However, they were also instructed that they may, e.g., select a location in the middle of a street, if they cannot decide on which side of the street they think the POI is. Participants were told to select any location on the map, in case they, e.g., could not hear a stimulus at all because of traffic.

Additionally, typical problems, such as front-back confusion and lack of externalization, associated with presenting spatial audio over headphones were explained. Participants were told that rotating their head often helps to reduce these problems and localize the direction of the sound sources.

3.9 Analysis measures and methods

Where measures could be compared pairwise (with respect to participant and POI) across conditions, the two-sided sign test was utilized, as it makes very few assumptions about the distribution of these measures. A significance level of 0.05 was chosen for judging if an effect is significant. In cases where a pairwise comparison could not be made, the two-sided Wilcoxon rank sum test was used instead.

For the sign test, the Z-statistic is reported when normal approximation was used to calculate the p-value. Otherwise, the number of pairs, S, where the first measure was larger, is reported. For the Wilcoxon rank sum test, the rank sum test statistic, W, is reported together with the Z-statistic.

The interquartile range was chosen as a measure of dispersion for the distance estimates of a single POI under a single condition. To make this range comparable across POIs and conditions, it was divided by the median distance of the POI under the condition in question. This relative interquartile range is thus calculated as

$$IQR = \frac{P_{75} - P_{25}}{P_{50}}, \tag{4}$$

where P_{75} is the 75th percentile, P_{25} is the 25th percentile, and P_{50} is the median of the distance estimates.

To compare the dispersion of the estimated azimuths of POIs, the circular variance is here used as a measure. The circular variance of a sample of n directions is defined by Fisher [1995] as

$$V = 1 - \bar{R}, \tag{5}$$

Table I. The mean IQR and circular variance calculated over all POIs for each condition.

Condition	$\overline{IQR}$		$\bar{V}$	
	a	b	a	b
1	0.82	0.54	0.21	0.18
2	1.18	0.53	0.16	0.11
3	0.52	0.28	0.11	0.06
4	0.98	0.57	0.21	0.24

Table II. Mean answering time for each condition.

Condition	Time [s]	
	a	b
1	5.7	4.7
2	14.1	9.4
3	10.4	7.4
4	12.9	9.1

where $\bar{R}$ is the mean resultant length of the vector resultant of the directions ϕ_i, as specified by the equations

$$C = \sum_{i=1}^{n} cos\ \phi_i,\ S = \sum_{i=1}^{n} sin\ \phi_i, \tag{6}$$

$$R = \sqrt{C^2 + S^2}, \tag{7}$$

and

$$\bar{R} = R/n. \tag{8}$$

4. RESULTS

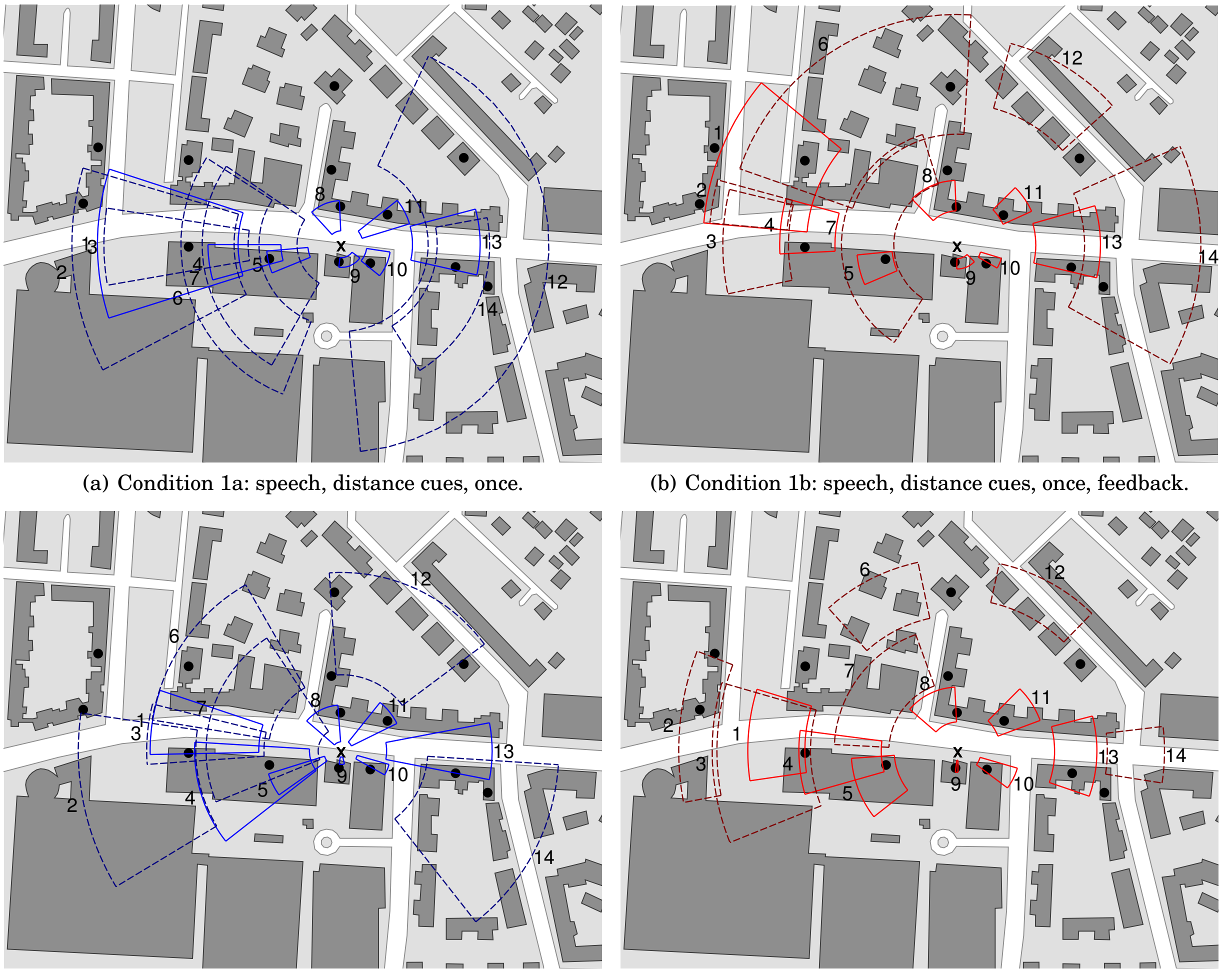
The participants marked the estimated POI locations on a map on a tablet computer using a stylus. Distance and azimuth estimates were extracted from these location estimates. The interquartile ranges of the distance and azimuth estimates are shown in Figure 5 for conditions 1 and 2, and in Figure 6 for conditions 3 and 4. The interquartile ranges are marked with the number of the POI, while the location of the participant is marked with a cross. The mean IQR and mean circular variance for each condition are shown in Table I.

The intended locations of the POIs are shown as black dots in Figures 5 and 6, but the numbers of these are not shown to avoid clutter (see Figure 2 for the corresponding POI numbers). Here, we talk about the intended POI locations (or distances or azimuths) of the POIs, rather than the correct POI locations. For example, the perceived azimuths of the virtual sound sources will not match the intended azimuths, mostly due to head-tracking inaccuracies. Thus, it is the virtual sound sources that should be corrected, so that the intended and perceived azimuths match.

The test itself, without briefing and debriefing, took between 22 and 78 minutes to finish. Most participants spent around 30 minutes doing the test. The answering times for each condition, averaged over all participants and POIs, are presented in Table II. The answering time is the time from the start of the playback of the sound until the participant had chosen the final location of the POI.

4.1 How does localization performance differ when presenting distance using modifications of intensity, early-to-late energy ratio, and type of speech compared with distances indicated in meters?

Participants tended to make shorter distance estimates under condition 1a, where distance cues were given using modifications of intensity, early-to-late energy ratio, and type of speech, than under condition 3a, where distances where expressed in meters ($Z = -4.8$, $p < 0.001$). The median increase in the distance estimates made under condition 3a compared with those made by the same participant under condition 1a was 35%. However, after visual feedback was given, there was no significant difference ($Z = -0.85$, $p = 0.39$) between the corresponding conditions (1b and 3b).



(a) Condition 1a: speech, distance cues, once.

(b) Condition 1b: speech, distance cues, once, feedback.

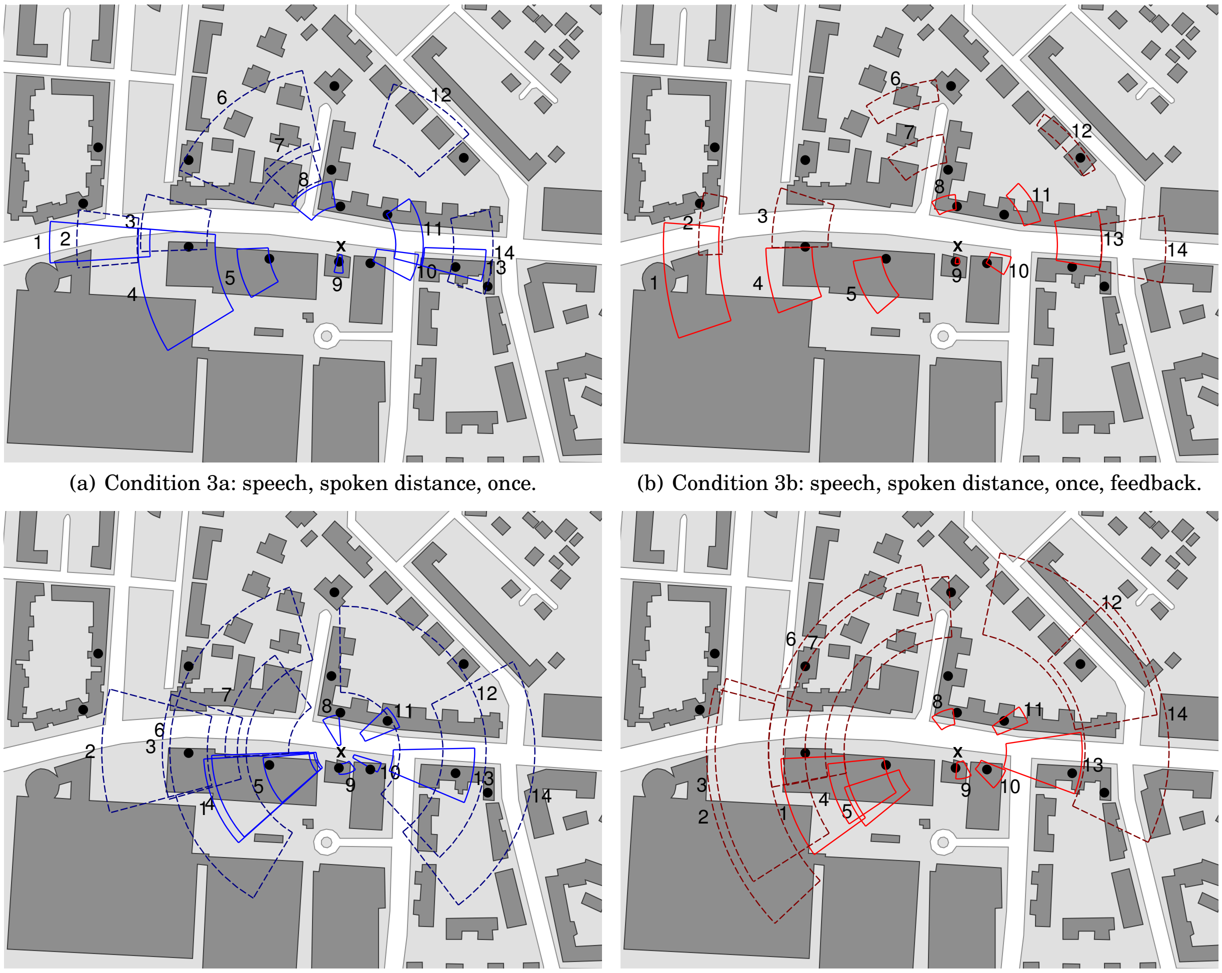
(c) Condition 2a: speech, distance cues, repeated.

(d) Condition 2b: speech, distance cues, repeated, feedback.

Fig. 5. Interquartile ranges of the distance and azimuth estimates under conditions 1 and 2. The black dots represent the intended POI locations. Interquartile ranges of obstructed POIs are shown with dashed lines.

The dispersion in the distance estimates, as measured by the *IQR*, was significantly larger under condition 1 (a and b) than it was under condition 3 (a and b) ($S = 25$ [*of* 28], $p < 0.001$). The circular variance was also significantly larger ($S = 22$ [*of* 28], $p = 0.004$) under condition 1 than it was under condition 3.

Answering times were significantly shorter under condition 1 than under condition 3 ($Z = -11.4$, $p < 0.001$). The increased precision of the distance and azimuth estimates under condition 3 came at the expense of the longer stimulus duration required to present the POI locations. Participants presumably also spent the longer answering times trying to measure the announced distances on the map.



(a) Condition 3a: speech, spoken distance, once.

(b) Condition 3b: speech, spoken distance, once, feedback.

(c) Condition 4a: noise, distance cues, repeated.

(d) Condition 4b: noise, distance cues, repeated, feedback.

Fig. 6. Interquartile ranges of the distance and azimuth estimates under conditions 3 and 4. The black dots represent the intended POI locations. Interquartile ranges of obstructed POIs are shown with dashed lines.

4.2 Does localization performance improve with repeated presentation of a sound compared with a single presentation?

Before feedback, participants tended to judge the distance as being larger ($Z = 2.1$, $p = 0.036$) with a single presentation (condition 1a) than with repeated presentation (condition 2a). However, the median decrease in the distance estimates made under condition 2a compared with those made by the same participant under condition 1a was only 11%. Between conditions 1b and 2b, after visual feedback was given, there was no significant difference ($Z = -0.31$, $p = 0.76$) in the distance estimates. Neither was there any significant difference ($S = 9$ $[of\ 28]$, $p = 0.087$) in the dispersion of the distance estimates between conditions 1 and 2. A clear advantage of repeated presentation can, however, be seen in the azimuth estimates, where it significantly reduced the circular variance ($S = 21$ $[of\ 28]$, $p = 0.013$), presumably by allowing more time for participants to use head movements to pinpoint the direction. This

Table III. The root-mean-square deviation from the intended locations of the POIs. The Z-statistics and p-values for a two-sided sign test comparing the deviations without feedback (a) and after visual feedback (b) are also shown.

	distance				azimuth			
Condition	a	b	Z	p	a	b	Z	p
1	73%	54%	4.2	**< 0.001**	58°	48°	1.9	0.062
2	75%	46%	5.5	**< 0.001**	47°	34°	0.39	0.70
3	65%	36%	3.9	**< 0.001**	37°	26°	1.9	0.054
4	53%	49%	2.1	**0.037**	51°	55°	1.5	0.14

increased precision came at the expense of the participants using significantly more time to perform the task when listening to repeated presentations was possible ($Z = -11.8$, $p < 0.001$).

4.3 How does localization of speech and non-speech signals compare?

There was no significant difference between the distance estimates of repeated speech (condition 2) and repeated noise bursts (condition 4), either before ($Z = 1.0$, $p = 0.31$) or after feedback ($Z = 0.77$, $p = 0.44$). Neither was there any significant difference ($S = 16$ [*of* 28], $p = 0.57$) in the dispersion of the distance estimates.

However, the direction of the noise bursts proved more difficult to localize, with a significant difference ($S = 7$ [*of* 28], $p = 0.013$) in the circular variance. This difference is probably due to the difficulty in estimating the direction of the obstructed POIs presented with noise, as can be observed in Figures 6(c) and 6(d), as compared with Figures 5(c) and 5(d), where the stimulus was speech.

4.4 How does training with visual feedback affect distance and azimuth estimates?

Before visual feedback was given, participants tended to make shorter distance estimates than after feedback was given under conditions 1 ($Z = -3.6$, $p < 0.001$), 2 ($Z = -3.8$, $p < 0.001$), and 4 ($Z = -5.2$, $p < 0.001$). However, under condition 3, where the distance in meters was given, training had no significant effect on the distance estimates ($Z = -0.85$, $p = 0.40$).

Training with feedback significantly reduced the dispersion in the distance estimates ($S = 49$ [*of* 56], $p < 0.001$), with the mean IQR being 0.88 before feedback and 0.48 after feedback was given. On the other hand, training had no significant effect on the circular variance ($S = 29$ [*of* 56], $p = 0.89$).

The root-mean-square deviations from the intended locations of the POIs are presented in Table III. The distances are calculated as percentages of the intended distances. As shown in the table, the distance deviations were significantly reduced after feedback for all conditions, while the azimuth deviations did not decrease significantly for any of the conditions.

Before training with feedback, participants used significantly more time for the task than after training with feedback (condition 1: $Z = 3.8$, $p < 0.001$; condition 2: $Z = 5.9$, $p < 0.001$; condition 3: $Z = 5.2$, $p < 0.001$, condition 4: $Z = 4.1$, $p < 0.001$). This may be due to the fact that training with visual feedback made it easier for the participants to perform the task, but it may also be due to the fact that they otherwise were more familiar with the task at this stage, and therefore presumably performed it faster. In addition, the participants' motivation to spend excess time performing the task had probably decreased at this stage.

4.5 How does localization of obstructed POIs compare with that of unobstructed POIs?

In the case of obstructed vs. unobstructed POIs, we cannot do a paired comparison, as we are comparing not only different presentation of the POIs, but also different POI locations. Therefore, we cannot

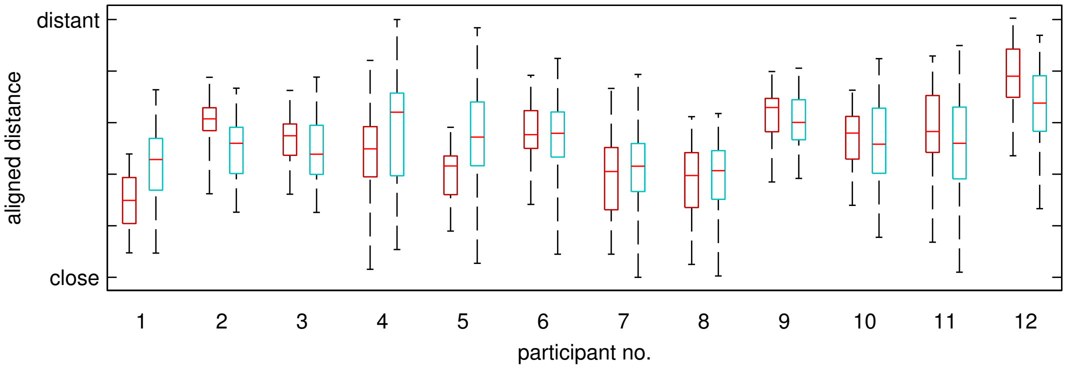


Fig. 7. Box plot of the aligned distance estimates. For each participant, estimates are shown separately for conditions 1a–4a (before feedback), on the left, and conditions 1b–4b (after feedback), on the right.

determine exactly which part of the observed effects is due to the differing presentation and which part is due to the differing locations.

The direction of the obstructed POIs proved significantly more difficult to estimate than that of the unobstructed POIs ($W = 2091$, $Z = 5.07$, $p < 0.001$ with a two-sided Wilcoxon rank sum test comparing the circular variances). The mean circular variance of the azimuth estimates for the unobstructed POIs under conditions 1, 2, and 4 was 0.10, while the circular variance for the obstructed POIs was 0.29. This difference is easily explained by the fact that the obstructed POIs were presented by interpolation between BRIRs in only four directions, with the direct sound heavily attenuated.

The difficulty in estimating the direction of the obstructed POIs probably explains most of the difference in the circular variance between conditions 1 and 3 (see Section 4.1). Another factor explaining the difference is the fact that all sounds under condition 3 were presented at a relatively high level, while sounds representing further-away POIs were heavily attenuated under condition 1.

On the other hand, the dispersion in the distance estimates tended to be smaller for the obstructed than for the unobstructed POIs ($W = 1187$, $Z = -3.10$, $p = 0.002$ with a two-sided Wilcoxon rank sum test). The mean IQR was 0.64 for the obstructed and 0.87 for the unobstructed POIs.

4.6 How do the distance estimates differ across participants?

A box plot of the distance estimates made by each participant, separately for conditions before and after feedback was given, is shown in Figure 7. These distance estimates were aligned by first taking the logarithm, to give equal weight to large and small distance estimates, and then calculating the z-score (by subtracting the mean and dividing by the sample standard deviation) for each POI under each condition separately. Before feedback was given, some participants consistently estimated the distances as being further away than others did. After feedback, the same tendency existed, but to a lesser degree.

4.7 Order effects

Although no visual feedback was given during conditions 1a, 2a, 3a, and 4a, the order in which participants were presented with these conditions might have had an effect on the distance estimates. In particular, the distances in meters were presented under condition 3a, so participants might have assumed that the distances of the POIs were in the same range during subsequent conditions.

Table IV. Comparison of distance estimates made before and after a certain condition was presented by means of a two-sided Wilcoxon rank sum test. Conditions 1a–4a.

Condition	p, before vs. after condition				Z, before vs. after condition			
	1a	2a	3a	4a	1a	2a	3a	4a
1a	—	**0.002**	0.075	0.069	—	3.04	−1.78	−1.82
2a	**< 0.001**	—	**< 0.001**	**< 0.001**	−4.11	—	−7.86	−4.73
3a	0.573	**< 0.001**	—	0.634	−0.56	5.30	—	−0.48
4a	0.618	**< 0.001**	**0.001**	—	−0.50	5.73	−3.27	—

Table V. Comparison of distance estimates made before and after a certain condition was presented by means of a two-sided Wilcoxon rank sum test. Conditions 1b–4b.

Condition	p, before vs. after condition				Z, before vs. after condition			
	1b	2b	3b	4b	1b	2b	3b	4b
1b	—	0.435	0.828	0.241	—	0.78	0.22	−1.17
2b	0.948	—	0.336	0.266	−0.07	—	0.96	1.11
3b	**0.027**	**0.023**	—	0.896	−2.21	−2.27	—	0.13
4b	0.298	0.882	0.487	—	−1.04	−0.15	−0.69	—

The two-sided Wilcoxon rank sum test was used to compare the distance estimates made before a certain condition was presented with those made after it. For each case, the rank sum test statistic was calculated separately for all POIs and the sum of these was used as a combined test statistic, as suggested by Lehmann [1975]. The p-values and Z-statistics are shown in Tables IV and V.

Strong order effects can be seen for conditions 2a and 4a, when comparing distance estimates made before and after condition 3a was presented. These might partly be due to the fact that distances were given in meters during condition 3a. However, strong order effects can also be seen in many other cases. These are presumably rather due to an overall learning effect as the test progressed. They might also result from differing answering tendencies (as mentioned in Section 4.6) between the specific group of participants that was presented with one condition before another, compared with the group that was presented with the same condition after the other condition. The order effects related to condition 3a might thus also be due to these reasons.

5. COMMENTS FROM THE PARTICIPANTS

In general, participants commented that the task was more difficult for noise than for speech samples. Especially far-away POIs presented with noise bursts were experienced as difficult to localize. Participants also thought that the task was more interesting with speech samples.

Some participants said that the task was easier when the distances were specified in meters. However, other participants said that they had difficulties associating these distances with distances in the environment or on the map. Participants said that they both looked at the map and at the environment when trying to localize POIs. However, some relied more on the map while others relied more on the environment.

With speech samples, the whispering and shouting voices were mentioned as good distance cues. Speech samples were also naturally associated with the environment. They were experienced, e.g., as announcements from the nearby shopping mall or from a nearby lamp post.

Several participants commented that the task was difficult with samples that were located in front of or behind them, while samples heard from the left or right were considered much easier to localize. Many participants also said that quiet, far-away samples were difficult to localize. This was the case also for the samples that were perceived as more reverberant, thus presumably those representing obstructed POIs.

During the debriefing, no questions were explicitly asked about how the participants interpreted the cues related to obstructed POIs. However, many participants did comment on the sounds representing these points of interest, as they sounded much more reverberant than other sounds, and thus stood out from them. These comments made it clear that these cues were difficult to interpret, as only one participant said she thought it sounded clearly like these POIs were behind buildings. Another participant said that these cues were easy to understand once visual feedback was given, but not before that. Other participants said that they intuitively associated these cues with, e.g., the POI being somewhere behind them, or being inside the nearby library. Two participants said that it only became clear to them that POIs might be behind buildings, i.e., not visible, once feedback was given.

6. DISCUSSION

For producing good distance estimates without any prior training or feedback to the user, presenting the distance in meters is a good approach. This allows for longer distance estimates than what would intuitively be produced by providing auditory distance cues. However, with feedback given, users are able to map these distance cues, which produce relatively short perceived distances, to longer actual POI distances. The resolution of such mapping is of course limited, and more precise estimates will be made when the distance in meters is presented, especially as the range of POI distances presented grows larger.

Using auditory distance cues to provide hints about the distance of a POI is a good approach when precise distance estimates are not important and sound duration should be kept at a minimum. When giving the user a rough idea about the distance is enough, such distance cues will suffice, and presumably require less attention and cognitive load from the user than presenting the distance in meters. However, the user needs to have some experience with the system to know the exact mapping between these cues, the perceived distance of the virtual sound source that they produce, and the actual POI distance.

Training not only reduced the deviation of the distance estimates from the intended distances of the POIs, but reduced the dispersion in the distance estimates among participants. This was the case both when providing auditory distance cues and when presenting the distance in meters. Training did not, however, significantly improve either the precision or the accuracy, with respect to the intended POI locations, of the azimuth estimates.

When precision in judging the direction of a POI is desired, repeated or otherwise longer sounds allowing more time to be localized should be preferred. For distance estimation, no advantage can be seen from the longer localization time allowed by repeated presentation of the sound. In practical applications, allowing users to repeat previous sounds would be a good option, in case they cannot localize the POI or comprehend the presented message, e.g., due to background noise.

The methods for presenting obstructed POIs in this study were not successful. No questions were explicitly asked about these, but only one participant said that the given cues and what they meant was clear, while another said it became clear once feedback about the POI locations was given. Many participants said that the more reverberant sounds, which presumably were those presenting obstructed POIs, were difficult to localize. The methods utilized resulted in poor azimuth estimates of obstructed POIs. This seems to be the case especially when presenting POIs with noise bursts.

Noise bursts perform equally good as speech for conveying distance information. However, in this study, more precise azimuth estimates were made with speech than noise. The poor azimuth estimates for noise stimuli may presumably be attributed to the apparent difficulty in estimating the direction of noise bursts presented with obstruction cues.

In an urban environment, background noise typically poses limitations for presenting virtual sound sources using distance attenuation. The most quiet sounds that may be presented while still being audible and understandable must be considerably louder than in a quiet environment. However, as the level of background noise increases, so does the maximum level at which sounds may be presented comfortably. Still, the range of comfortable listening levels is reduced with increasing noise [Pollack 1952].

In applications where the background noise may vary substantially, the noise levels should ideally be monitored and the presentation levels adjusted accordingly, to provide comfortable listening levels and take advantage of the dynamics available. Naturally, the range of perceived distances will vary together with the range of presentation levels, but the masking effect of the noise will at the same time presumably affect both reverberation and loudness cues to distance. Studies on the effect of background noise on perceived distances have given contradictory results. A study where the main sound source was noise concluded that background noise masked the reverberation of the main source and thereby reduced its perceived distance [Mershon et al. 1989]. On the other hand, an effect in the opposite direction, i.e., increased perceived distance with increased background noise, was observed in a study where speech was the main sound source [Cabrera and Gilfillan 2002].

7. CONCLUSIONS

Presenting virtual sound sources in an augmented reality environment is challenging when it comes to presenting the sources at large distances. For example, a human voice can usually not realistically be heard from several hundred meters away. In this study, we presented sounds representing different points of interest in an urban environment. We modified features affecting the perceived distance of the virtual sound sources and organized a user study to see where in the environment participants thought the corresponding POIs would be. This was compared with simply having a voice saying the distance to the POIs in meters. We hypothesized that the participants would locate the POIs relatively close to themselves when utilizing the provided auditory distance cues. Therefore, we also included short training with visual feedback to see if participants quickly could learn to map the available cues to the, presumably larger, intended POI distances.

The results indicate, that to provide fairly accurate distance estimates without training, distances should be given in meters. Distance estimates made based on the provided auditory distance cues were generally, as expected, relatively short with large dispersion among the estimates. However, with only little training, the test participants learned to map these cues to the intended POI distances fairly accurately and precisely.

Repeated presentation of the POI sounds improved azimuth estimation, by allowing more time to estimate the direction of the POIs by means of, e.g., head movements, but provided no advantage for distance estimation. The type of signal also had an impact on the localization performance. Noise bursts provided equally good distance estimates as speech, but the direction of noise bursts was more difficult to establish.

In this study, we also attempted to simulate sound sources obstructed by buildings to provide an additional cue as to their locations. However, because the direct sound was heavily attenuated in this simulation, much of the directional information was lost, resulting in relatively poor azimuth estimation. The obstruction cues used also proved ambiguous and difficult to interpret.

This study shows, that although virtual sound sources are difficult to present so that they are perceived at very large distances, people can with only little training learn to use the provided auditory distance cues and map these to large distances in the real world. Virtual auditory displays can thus be used as means to present the locations of points of interest together with other information in a natural and effective way.

REFERENCES

Robert Albrecht and Tapio Lokki. 2013. Adjusting the perceived distance of virtual speech sources by modifying binaural room impulse responses. In *Proceedings of the 19th International Conference on Auditory Display (ICAD 2013)*. Lodz, Poland, 233–241.

V. Ralph Algazi, Richard O. Duda, Dennis M. Thompson, and Carlos Avendano. 2001. The CIPIC HRTF database. In *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (WASPAA01)*. New Paltz, New York, USA, 99–102. DOI:http://dx.doi.org/10.1109/ASPAA.2001.969552

Jeffrey R. Blum, Mathieu Bouchard, and Jeremy R. Cooperstock. 2013. Spatialized audio environmental awareness for blind users with a smartphone. *Mobile Networks and Applications* 18, 3 (2013), 295–309. DOI:http://dx.doi.org/10.1007/s11036-012-0425-8

Steven F. Boll. 1979. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech and Signal Processing* 27, 2 (1979), 113–120. DOI:http://dx.doi.org/10.1109/TASSP.1979.1163209

Douglas S. Brungart. 2002. Near-field virtual audio displays. *Presence: Teleoperators and Virtual Environments* 11, 1 (2002), 93–106. DOI:http://dx.doi.org/10.1162/105474602317343686

Douglas S. Brungart and Kimberly R. Scott. 2001. The effects of production and presentation level on the auditory distance perception of speech. *The Journal of the Acoustical Society of America* 110, 1 (2001), 425–440. DOI:http://dx.doi.org/10.1121/1.1379730

Densil Cabrera and David Gilfillan. 2002. Auditory distance perception of speech in the presence of noise. In *Proceedings of the 8th International Conference on Auditory Display (ICAD 2002)*. Kyoto, Japan.

James E. Cutting and Peter M. Vishton. 1995. Perceiving layout and knowing distances: The integration, relative potency, and contextual use of different information about depth. In *Perception of Space and Motion*, William Epstein and Sheena Rogers (Eds.). Academic Press, Chapter 3, 69–117. DOI:http://dx.doi.org/10.1016/B978-012240530-3/50005-5

David Doukhan and Anne Sédès. 2009. CW_binaural~: A binaural synthesis external for Pure Data. In *Proceedings of the 3rd International Pure Data Convention (PdCon09)*. São Paulo, Brazil.

Nicholas I. Fisher. 1995. *Statistical analysis of circular data*. Cambridge University Press.

Kim Fluitt, Timothy Mermagen, and Tomasz Letowski. 2014. Auditory distance estimation in an open space. In *Soundscape Semiotics - Localization and Categorization*, Hervé Glotin (Ed.). InTech, Chapter 7, 135–165. DOI:http://dx.doi.org/10.5772/56137

Jens Grubert, Tobias Langlotz, and Raphaël Grasset. 2011. *Augmented reality browser survey*. Technical Report ICG-TR-1101. Institute for Computer Graphics and Vision, Graz University of Technology, Austria.

Simon Holland, David R. Morse, and Henrik Gedenryd. 2002. AudioGPS: Spatial audio navigation with a minimal attention interface. *Personal and Ubiquitous Computing* 6, 4 (2002), 253–259. DOI:http://dx.doi.org/10.1007/s007790200025

Erich L. Lehmann. 1975. *Nonparametrics: Statistical methods based on ranks*. Holden-Day.

David McGookin and Stephen A. Brewster. 2012. Understanding auditory navigation to physical landmarks. In *Haptic and Audio Interaction Design*, Charlotte Magnusson, Delphine Szymczak, and Stephen Brewster (Eds.). Lecture Notes in Computer Science, Vol. 7468. Springer, 1–10. DOI:http://dx.doi.org/10.1007/978-3-642-32796-4_1

Donald H. Mershon, William L. Ballenger, Alex D. Little, Patrick L. McMurtry, and Judith L. Buchanan. 1989. Effects of room reflectance and background noise on perceived auditory distance. *Perception* 18, 3 (1989), 403–416. DOI:http://dx.doi.org/10.1068/p180403

Irwin Pollack. 1952. Comfortable listening levels for pure tones in quiet and noise. *The Journal of the Acoustical Society of America* 24, 2 (1952), 158–162. DOI:http://dx.doi.org/10.1121/1.1906871

Tuyen V. Tran, Tomasz Letowski, and Kim S. Abouchacra. 2000. Evaluation of acoustic beacon characteristics for navigation tasks. *Ergonomics* 43, 6 (2000), 807–827. DOI:http://dx.doi.org/10.1080/001401300404760

Pavel Zahorik, Douglas S. Brungart, and Adelbert W. Bronkhorst. 2005. Auditory distance perception in humans: A summary of past and present research. *Acta Acustica united with Acustica* 91, 3 (2005), 409–420.

Received August 2014; revised December 2014; accepted December 2014