Where Are We? Evaluating the Current Rendering Fidelity of Mobile Audio Augmented Reality Systems

Florian Heller* Jayan Jevanesan* Pascal Dietrich‡ Jan Borchers*

*RWTH Aachen University
52056 Aachen, Germany
{flo,borchers}@cs.rwth-aachen.de jayan.jevanesan@rwth-aachen.de

‡KLANG:technologies
52066 Aachen, Germany
dietrich@klang.com

ABSTRACT

Mobile audio augmented reality systems (MAARS) simulate virtual audio sources in a physical space via headphones. While 20 years ago, these required expensive sensing and rendering equipment, the necessary technology has become widely available. Smartphones have become capable to run high-fidelity spatial audio rendering algorithms, and modern sensors can provide rich data to the rendering process. Combined, these constitute an inexpensive, powerful platform for audio augmented reality.

We evaluated the practical limitations of currently available off-the-shelf hardware using a voice sample in a lab experiment. State of the art motion sensors provide multiple degrees of freedom, including pitch and roll angles instead of yaw only. Since our rendering algorithm is also capable of including this richer sensor data in terms of source elevation, we also measured its impact on sound localization. Results show that mobile audio augmented reality systems achieve the same horizontal resolution as stationary systems. We found that including pitch and roll angles did not significantly improve the users' localization performance.

ACM Classification Keywords

H.5.1. Information Interfaces and Presentation (e.g. HCI): Multimedia Information Systems

Author Keywords

Virtual Audio Spaces; Spatial Audio; Mobile Devices; Audio Augmented Reality; Navigation.

INTRODUCTION

Spatial audio rendering uses special audio filters to create the impression that a recorded sound emerges from a certain position outside the user's head. In combination with a constant tracking of head position and orientation, mobile audio augmented reality systems (MAARS) use this technology to overlay the physical space with a virtual audio space. When experiencing this audio space through headphones, the virtual audio sources appear to be fixed at certain positions in the physical space. These audio spaces are used, among other, for an engaging presentation of information in museums [3, 18] or as non-visual navigation system [16, 17, 22].

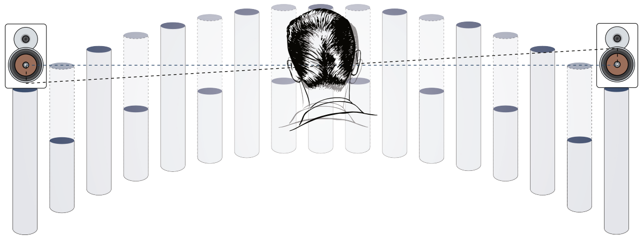

Figure 1. Using an HRTF-based algorithm allows to integrate richer sensor data including head pitch and roll into the rendering. We used this to evaluate the localization accuracy for sources at equal level and at different heights.

A key aspect of audio augmented reality systems is the ability for users to localize the origin of sounds. While simple approaches like stereo panning are robust and work well for outdoor navigation [6], sources are typically much closer in an indoor scenario, making it desirable to increase the spatial resolution of the rendering algorithm. Algorithms based on head-related transfer functions (HRTFs), which can render a realistic impression for several sources on a modern smartphone [13], achieve this precision. An HRTF describes the modification of sound reaching the ear depending on its origin relative to the listener's head. Spatial localization of sound sources using an individually tuned HRTF (which depends, e.g., on the specific shape of that user's outer ear) is as precise as localizing natural sound sources in the real world [24].

To create this realistic experience, the algorithm has to know the user's head orientation. Headphones equipped with the required sensors like the Jabra Intelligent Headset[1] have recently become commercially available, reducing the technological prerequisites to a minimum.

Since measuring individual HRTFs is not feasible for a large user population [11], spatial audio rending is usually done using a set of generalized HRTFs. We used such a state of the art to evaluate the minimal angular distance between two source candidates at which users can successfully localize a recorded voice sample, a common scenario in MAARS.

In contrast to other existing approaches, HRTF-based algorithms can also simulate the elevation of sound sources. This

[1] intelligentheadset.com

MobileHCI '16, September 06 – 09, 2016, Florence, Italy
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-4408-1/16/09…$15.00
DOI: http://dx.doi.org/10.1145/2935334.2935365

can be used in two ways: **1)** in combination with richer sensor data it can react to different tilt angles of the user's head, providing additional cues for sound source localization. Modern inertial measurement units (IMUs), like the InvenSense MPU-9250 used in the Intelligent Headset, already measure 9 degrees of freedom (accelerometer, gyroscope, magnetometer, all in 3D) to provide a tilt-compensated heading output [7]. As a result, when attached to the user's headphones to measure head orientation in the 2D plane, these chips provide the additional information about roll and pitch of the user's head for free. **2)** to simulate sound sources at different vertical elevation which would be useful to increase the realism if sound sources are attached to physical artifacts at different heights.

In the remainder of this paper, we analyze (a) the impact of a an HRTF-based rendering algorithm and (b) simulated source elevation and inclusion of head pitch and roll tracking on the ability to discern between proximate virtual sound sources.

RELATED WORK

The underlying technology of audio augmented reality systems has been studied for over 25 years [8], and while early implementations required complex hardware setups to simulate the basic elements of human spatial hearing [1, 8, 12], modern smartphones are sufficiently powerful to render realistic auditory experiences [13]. Early MAARS, such as AudioGPS [6], showed that successful navigation is possible even with crude orientation data and simple stereo panning to provide a sense of orientation. Since the system did not include a digital compass, heading was determined from two consecutive GPS measurements, which means that this information was only updated after having turned and moved in a new direction. More recent implementations, therefore, use a digital compass mounted onto the headband of over-the-ear headphones to acquire head orientation data and simulate the sources in the horizontal plane. The Roaring Navigator [16] assists the user in navigating through a zoo using recordings of animal sounds as beacons for their respective enclosure. Corona [3] recreates a medieval coronation feast at its original location by augmenting the room with virtual characters discussing different aspects of the ceremony. Finally, the Sound Garden installation by Vazquez-Alvarez et al. [21] uses beacon sounds to help people navigate to points of interest in a municipal garden, and once the user is close enough, an explanatory audio sample is played.

To a certain degree, head orientation can be approximated by device orientation, making modern smartphones a feasible sensing and rendering platform. Heller et al. [5] compared the relative orientation of head, body, and device while navigating through a virtual audio space with a mobile device. While device orientation can only partially substitute head orientation, it still enables successful navigation in a space. This idea was exploited in AudioScope [4] by communicating this functional principle with the metaphor of a directional microphone.

Vazquez-Alvarez et al. [20] evaluated the ability to distinguish between virtual audio sources in the horizontal plane using the HRTF-library integrated to a Nokia N95 smartphone and determined that sources at 45° spacing can safely be distinguished. Sodnik et al. [14] evaluated the resolution of a spatial audio rendering engine and achieved a much higher resolution. However, they consecutively played bursts of white noise from two adjacent sources, which means that participants had a direct comparison between the stimuli. We want to determine the angular distance at which users can confidently assign a single sound to a physical artifact.

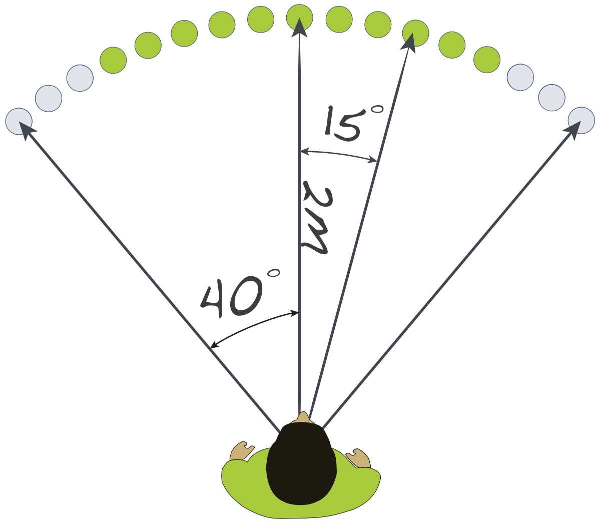

Figure 2. We placed 17 cardboard tubes at 2 m distance to the listener and tested sources spacings of 5°, 10°, 15°, and 20°. No sound sources were rendered at the outer six sound source positions (grey) to not artificially limit the choice of possible candidates towards the border. The active sound sources (green) were thus placed within a range of $[-25^\circ, 25^\circ]$

Mariette [10] studied the effect of head-turn latency, rendering quality, and capture circle on the interaction with virtual audio spaces. His results show that a high-quality rendering algorithm suffers to a higher extent from imprecise measurements, but navigation can still be completed successfully even with a high latency in head-turn measurement.

In summary, even simple rendering approaches can be used to successfully navigate to a certain target or to create engaging interactive experiences. As modern smartphones are a readily available sensing and rendering platform for MAARS, similar to the work by Vazquez-Alvarez et al. [20], we want to gauge the capabilities of such a mobile platform, but in a context where visual representatives of the virtual sound sources are available, as it would be the case, e.g., in a museum.

EXPERIMENT

To measure the minimal distance between two source candidates and the effect of simulated source elevation on the ability to localize virtual sound sources, we conducted the following experiment: We first measured the localization performance for sources that are all at equal height, approximately at the level of the user's ears. As tilting your head has an impact on the relative elevation of the source (cf. Figure 1), we ran the experiment with yaw measurement only (*flat*) and all degrees of freedom (*full*). We placed 17 cardboard tubes of 140 cm height in the range of $[-40^\circ, 40^\circ]$ at 5° intervals and at two meter distance from the listener (cf. Figure 2). To test the different angular distances between two candidate sources, we marked some of the tubes as potential candidates by placing a physical marker (a small loudspeaker) on top of it. We tested 5°, 10°, 15°, and 20° spacings, by marking all, every other, every third, or every fourth tube as active respectively. Participants were instructed to look at the source directly in front of them before every trial to allow the experimenter to check the compass calibration. A sound was then played at the position of one of the markers, and participants had to locate the source and say its number aloud.

Participants were encouraged to perform some trials before the experiment to become acquainted with the system. We only played sounds in the range of $[-25^\circ, 25^\circ]$ since head movement was not restricted and participants would turn their head towards the sound anyway. We also placed markers outside this range (grey positions in Figure 2) to not artificially limit the choices of candidate sources towards the outermost sources. In the third condition (*elevation*) we placed virtual sound sources, along with their physical markers, at two different heights (140 cm or 70 cm), thereby increasing their euclidean distance. We only tested the smallest angle (5° intervals) in this condition to make sure that we can see some effect, as we expected larger angles to be discernible successfully anyways. With 10 trials for all four angles in the *flat* and *full* condition, and 10 trials in the *elevation* condition, this resulted in a total of 90 trials per participant. The experimental sessions took around 20 minutes per participant, including filling out consent forms and questionnaires.

The sound sample we used was a continuous monologue of a male voice in a silent environment. While voice is harder to localize than bursts of white noise or a sonar pulse wich are commonly used as stimulus[19, 14], this is much closer to real applications, e.g., in a museum. The order of conditions was counterbalanced, and locations were randomized using latin squares. We recorded task completion time, head orientation in three degrees of freedom, accuracy, and evaluated the perceived presence in the virtual environment using a questionnaire [25].

TECHNICAL SETUP

We used the KLANG:kern (klang.com) spatial audio rendering platform running on an Apple iPad Air 2, and tracked head orientation using the Jabra Intelligent Headset. The rendering uses a generalized HRTF which has a resolution of 1° in horizontal and 5° in vertical direction. In a small experiment with 5 users, we determined a minimum audible angle of around 6° in horizontal and 16° in vertical direction. The headset reports changes in head orientation at a rate of around 40 Hz and has a specified latency of around 100 ms, which is noticeable [2] but well below the limits of 372 ms defined in [10]. While sensor data was transmitted via Bluetooth, we used a wired connection for audio to minimize latency.

RESULTS

A total of 22 users participated in the study (3 female, average age 28 years, SD=5). None reported having a known problem with spatial hearing. 50% of the participants reported having prior experience with audio augmented reality. While this high number of participants with prior experience is not representative for the general population at this moment, it allows a comparison with novices to detect learning effects.

		Angle			
	Condition	5°	10°	15°	20°
Recognition rate	*full*	29 %	62 %	76 %	86 %
	flat	30 %	64 %	80 %	85 %
	elevation	33 %			
Task compl. time	*full*	9.65s	9.18s	6.24s	5.43s
	flat	9.02s	7.21s	5.62s	5.35s
	elevation	12.64s			

Table 1. Percentage of correctly identified sources and task completion time by angular distance. Including all 3 degrees of freedom of the head into the rendering does not significantly increase the ability to localize the origin of sounds.

The most relevant parameter in practice is the rate of correctly recognized sources, which is reported for the different angular distances in Table 1. If we analyze the *full* and *flat* conditions, there is a significant main effect of SEPARATION ANGLE on RECOGNITION RATE $F_{(3,1751)} = 148.06$, $p < .0001$ The recognition rates rise with increased angular distance. A post-hoc Tukey HSD showed that all differences are significant at a level of $p < .0001$, except for the difference between 15° and 20° with $p = 0.0377$. Overall, the recognition rate for the smallest angle is low (*full*: M=29%, SD=45%, *flat*: M=30%, SD=46%). Placing the sources at different heights does not have a significant effect on the recognition rate (*elevation*: M=33%, SD=47%, $F_{(2,657)} = 0.37$, $p = 0.6911$). Since the 5° spacing is close to the minimal audible angle of the rendering algorithm, this is not surprising. However, the individual recognition rates vary greatly, as some participants achieved 70% correct answers even for the 5° interval. Simulating additional cues, as in the *full* condition, did not significantly improve the recognition rates compared to using head yaw only as in the *flat* condition. Overall, the recognition rate is significantly higher for participants with prior experience ($F_{(1,1758)} = 28.15$, $p < .001$), which indicates that after some time, users accommodate to the auditory experience. For example, at 10° spacing, the average recognition rate jumps from 56% (SD=50%) to 70% (SD=46%) with prior experience. A post-hoc Tukey HSD showed this difference to be significant ($p = .011$).

Participants took much longer to localize the sources on two different levels in the *elevation* condition (M=12.64 *s*, SD=8.5) compared to the other two conditions with a source spacing of 5°, with *full* and *flat* being quite similar (M=9.65 *s*, SD=5.31 vs. M=9.02 *s*, SD=3.69). A post-hoc Student's t-test revealed the differences to be significant ($p < .0001$). Again, prior experience has a significant impact. The task completion time in the *full* condition is significantly shorter for participants with prior experience (M=7.05 *s*, SD=3.9 vs. M=8.2 *s*, SD=8.2), which indicates that after an accommodation phase, localization performance increases [9].

We calculated the root mean squares (RMS) of all three head orientation angles as an indicator of how much participants moved their head along the respective axes (cf. Figure 3). The assumption is that with additional cues, the users would rotate their head less. First we compare the differences across all three conditions in the RMS angles for the 5° source spacing. The amount participants turned their head left and right is very similar in the *elevation* (M=18.96°, SD=5.03) and *full* (M=18.54°, SD=2.64) condition, and only slightly higher in the *flat* condition (M=20.26°, SD=6.38). A repeated measures ANOVA with user as random factor on the log-transformed RMS angles showed a significant effect of the CONDITION on ROLL ($F_{(2,42)} = 7.667$, $p = .0015$) and PITCH ($F_{(2,42)} = 4.91$, $p = .0121$). Post-hoc Tukey HSD tests showed that participants rolled their head significantly more in the *elevation* condition (M=5.2°, SD=2.52) than in the other two (*full* M=3.9°, SD=2.3; *flat* M=3.93°, SD=2.48; $p < .006$). The RMS pitch angles are only significantly different between the *elevation* (M=10.58°, SD=3.84) and *flat* (M=7.97°, SD=4.31) condition, which shows that, although not really noticeable, participants nodded while localizing the sources if all three angles were included in the rendering. The RMS angles for the other spacings do not differ significantly between *flat* and *full* conditions.

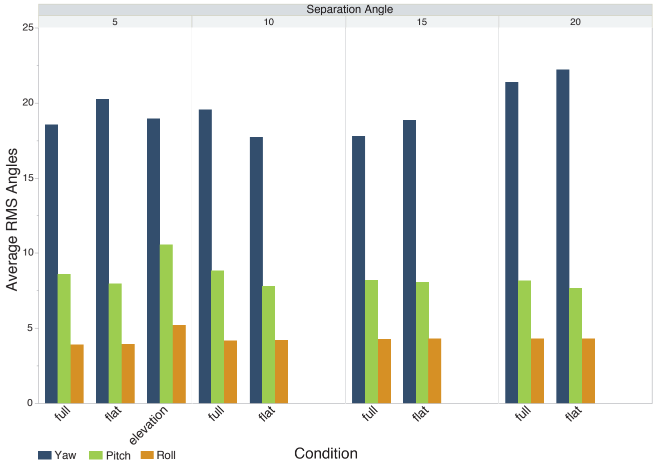

Figure 3. The average RMS angles for yaw (left-right rotation), pitch (looking up or down), and roll (tilting head sideways) by condition and source separation angle. The *elevation* condition was only tested at 5° intervals. RMS Yaw angles are largest as the task was to localize sources in horizontal direction. While the RMS pitch angles are slightly higher in the *elevation* condition, the difference to the *full* condition is not significant.

The median ratings given on a five point Likert scale (1 being the best) only differed marginally. None of the differences in the ratings was statistically significant according to a Wilcoxon signed rank test. Participants felt equally able to localize sounds in both conditions (Mdn=3, IQR=2). When asked how consistent the experience with the virtual environment seemed with the real world, participants gave slightly better ratings for the *elevation* condition (Mdn=2, IQR=2 vs. *flat*: Mdn=3, IQR=2.5). The ratings for the perceived naturalness of the virtual experience were the same for both conditions (Mdn=2, IQR=2). The participants rated the system as very responsive (Mdn=1, IQR=1) and that they could adapt to it quickly (*elevation*: Mdn=1, IQR=1.5; *flat*: Mdn=1, IQR=2).

DISCUSSION

Taking into account that we used a voice sample as beacon, which is harder to localize than white noise or a sonar ping [19], the results are comparable to those from Wenzel et al. [23] who achieved an average angle of error around 21° for the frontal region and at low elevations. Four participants of our experiment reached 100% recognition rate for sources at 10° intervals in the *full* condition, which shows that a very accurate localization is possible.

Contrary to what we expected, using the elevation rendering of the HRTF-based algorithm to place the sources at different heights did not increase horizontal resolution. While the euclidean distance between the sources increases when placing sources at different elevations, the impact of the vertical difference is minimized by the generalized HRTF. We know that the resolution of human sound localization is lower in the vertical than in the horizontal plane and that the generalization of HRTFs mostly affects the features used to the determine the elevation of a source [26]. This is in line with work by Sodnik et al. [15], which showed that localizing sources at arbitrary heights is difficult. The inclusion of head roll and pitch into the rendering algorithm did not further improve horizontal resolution. While users moved their head more in the *full* condition, which indicates that it is noticeable, they also often asked if there was any difference between the *full* and *flat* conditions, indicating that the difference was barely noticeable consciously.

CONCLUSION AND FUTURE WORK

Mobile audio augmented reality systems, running on hardware that consists of a modern smartphone and an augmented headphone achieve a performance similar to stationary implementations running on complex hardware [24]. This substantially minimizes the hardware development effort that, so far, limited the dissemination of mobile audio augmented reality applications. This means that applications such as MAAR navigation systems or audio augmented reality museum guides can easily be deployed to end users. Our implementation used a rendering engine that is capable of simulating sources at different elevations. As the IMU on the headphones reports all necessary information, we evaluated if including this additional cue increases localization accuracy. We measured no significant difference in recognition rate between sources at different heights and sources all on the same level.

Our rendering did not include a simulation of room acoustics, i.e., reverb, which results in a "dry" impression of the sound as it would be in an anechoic chamber. Before deploying such a system in a real world setting, we would include reverb either in form of the real impulse responses or if the room has an unsuitable natural reverb, a more appropriate synthesized type.While the balance between reverberated and direct signal is mostly a distance cue, this might further improve the recognition rate and the immersion of the user.

In our lab experiment, participants were standing at 2 m distance to the sources. A more realistic scenario would include navigation within a virtual audio space and thus include users relative movement to the sources. We plan to conduct further studies with a higher degree of real-world relevance by having users move towards sound sources as we expect to see a bigger impact of richer sensor data on the ability to differentiate different sources.

ACKNOWLEDGMENTS

We thank the participants of our study for their time. This work was funded in part by the German B-IT Foundation.

REFERENCES

1. Jens Blauert. 1996. *Spatial Hearing: Psychophysics of Human Sound Localization* (2 ed.). MIT Press.
2. Douglas S Brungart, Brian D Simpson, and Alexander J Kordik. 2005. The detectability of headtracker latency in virtual audio displays. In *ICAD '05*. http://hdl.handle.net/1853/50185
3. Florian Heller. 2014. Corona: Audio AR for historic sites. *AR[t]* 5 (2014). http://arlab.nl/sites/default/files/ARt5_magazine_webversie.pdf
4. Florian Heller and Jan Borchers. 2015. AudioScope: Smartphones as Directional Microphones in Mobile Audio Augmented Reality Systems. In *CHI '15*. DOI: http://dx.doi.org/10.1145/2702123.2702159
5. Florian Heller, Aaron Krämer, and Jan Borchers. 2014. Simplifying Orientation Measurement for Mobile Audio Augmented Reality Applications. In *CHI '14*. DOI: http://dx.doi.org/10.1145/2556288.2557021
6. Simon Holland, David R Morse, and Henrik Gedenryd. 2002. AudioGPS: Spatial Audio Navigation with a Minimal Attention Interface. *Pers. and Ubiqu. comp.* 6, 4 (2002). DOI:http://dx.doi.org/10.1007/s007790200025
7. Erkka Laulainen, Lauri Koskinen, Marko Kosunen, and Kari Halonen. 2008. Compass tilt compensation algorithm using CORDIC. In *ISCAS '08*. DOI: http://dx.doi.org/10.1109/ISCAS.2008.4541636
8. Jack M Loomis, Chick Hebert, and Joseph G Cicinelli. 1990. Active localization of virtual sounds. *J. Acoust. Soc. Am.* 88 (1990). DOI:http://dx.doi.org/10.1121/1.400250
9. Piotr Majdak, Thomas Walder, and Bernhard Laback. 2013. Effect of long-term training on sound localization performance with spectrally warped and band-limited head-related transfer functions. *J. Acoust. Soc. Am.* 134, 3 (2013). DOI:http://dx.doi.org/10.1121/1.4816543
10. Nicholas Mariette. 2010. Navigation Performance Effects of Render Method and Head-Turn Latency in Mobile Audio Augmented Reality. In *ICAD '09*. DOI: http://dx.doi.org/10.1007/978-3-642-12439-6_13
11. Bruno Sanches Masiero. 2012. *Individualized Binaural Technology. Measurement, Equalization and Subjective Evaluation*. http://d-nb.info/1030407002/34
12. John C Middlebrooks and David M Green. 1991. Sound localization by human listeners. *Annual review of psychology* 42, 1 (1991). DOI: http://dx.doi.org/10.1146/annurev.ps.42.020191.001031
13. Christian Sander, Frank Wefers, and Dieter Leckschat. 2012. Scalable Binaural Synthesis on Mobile Devices. In *Audio Engineering Society Convention 133*. http://www.aes.org/e-lib/browse.cfm?elib=16525
14. Jaka Sodnik, Rudolf Sušnik, Mitja Štular, and Sašo Tomažič. 2005. Spatial sound resolution of an interpolated HRIR library. *Applied Acoustics* 66, 11 (2005). DOI: http://dx.doi.org/10.1016/j.apacoust.2005.04.003
15. Jaka Sodnik, Saso Tomazic, Raphael Grasset, Andreas Duenser, and Mark Billinghurst. 2006. Spatial sound localization in an augmented reality environment. In *OZCHI '06*. DOI: http://dx.doi.org/10.1145/1228175.1228197
16. Christoph Stahl. 2007. The roaring navigator: a group guide for the zoo with shared auditory landmark display. In *MobileHCI '07*. DOI: http://dx.doi.org/10.1145/1377999.1378042
17. Steven Strachan, Parisa Eslambolchilar, Roderick Murray-Smith, Stephen Hughes, and Sile O'Modhrain. 2005. GpsTunes: Controlling Navigation via Audio Feedback. In *MobileHCI '05*. DOI: http://dx.doi.org/10.1145/1085777.1085831
18. Lucia Terrenghi and Andreas Zimmermann. 2004. Tailored audio augmented environments for museums. In *IUI '04*. DOI:http://dx.doi.org/10.1145/964442.964523
19. Tuyen V Tran, Tomasz Letowski, and Kim S Abouchacra. 2000. Evaluation of acoustic beacon characteristics for navigation tasks. *Ergonomics* 43, 6 (2000). DOI: http://dx.doi.org/10.1080/001401300404760
20. Yolanda Vazquez-Alvarez. 2009. Investigating Background & Foreground Interactions Using Spatial Audio Cues. In *CHI EA '09*. DOI: http://dx.doi.org/10.1145/1520340.1520578
21. Yolanda Vazquez-Alvarez, Ian Oakley, and Stephen A Brewster. 2012. Auditory display design for exploration in mobile audio-augmented reality. *Pers. and Ubiqu. comp.* 16, 8 (2012). DOI: http://dx.doi.org/10.1007/s00779-011-0459-0
22. Nigel Warren, Matt Jones, Steve Jones, and David Bainbridge. 2005. Navigation via Continuously Adapted Music. In *CHI EA '05*. DOI: http://dx.doi.org/10.1145/1056808.1057038
23. Elizabeth M Wenzel, Marianne Arruda, Doris J Kistler, and Frederic L Wightman. 1993. Localization using nonindividualized head-related transfer functions. *J. Acoust. Soc. Am.* 94, 1 (1993). DOI: http://dx.doi.org/10.1121/1.407089
24. Elizabeth M Wenzel, Frederic L Wightman, and Doris J Kistler. 1991. Localization with non-individualized virtual acoustic display cues. In *CHI '91*. DOI: http://dx.doi.org/10.1145/108844.108941
25. Bob G Witmer and Michael J Singer. 1998. Measuring Presence in Virtual Environments: A Presence Questionnaire. *Presence: Teleoper. Virtual Environ.* 7, 3 (1998). DOI:http://dx.doi.org/10.1162/105474698565686
26. Dmitry N Zotkin, Ramani Duraiswami, and Larry S Davis. 2004. Rendering localized spatial audio in a virtual auditory space. *IEEE Transactions on Multimedia* 6, 4 (2004). DOI:http://dx.doi.org/10.1109/TMM.2004.827516