AudioTorch: Using a Smartphone as Directional Microphone in Virtual Audio Spaces

Florian Heller
RWTH Aachen University
52056 Aachen
flo@cs.rwth-aachen.de

Jan Borchers
RWTH Aachen University
52056 Aachen
borchers@cs.rwth-aachen.de

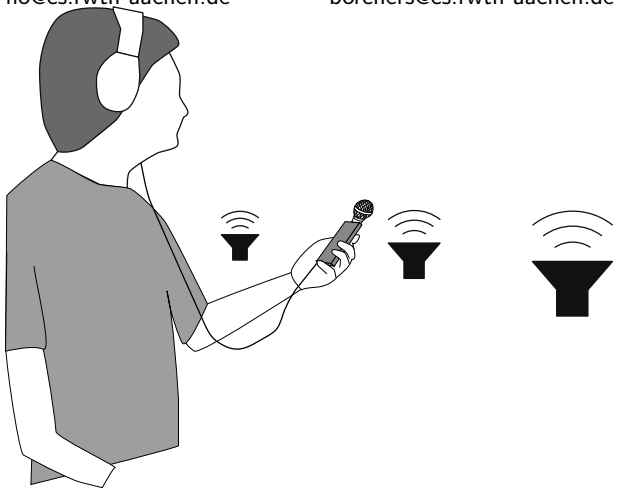

Figure 1: AudioTorch converts your mobile phone into a virtual directional microphone to experience audio augmented reality scenes.

MobileHCI'14, September 23–26, 2014, Toronto, ON, Canada.
ACM 978-1-4503-3004-6/14/09.
http://dx.doi.org/10.1145/2628363.2634220

Abstract
Mobile audio augmented reality systems can be used in a series of applications, e.g., as a navigational aid for visually impaired or as audio guide in museums. The implementation of such systems usually relies on head orientation data, requiring additional hardware in form of a digital compass in the headphones. As an alternative we propose AudioTorch, a system that turns a smartphone into a virtual directional microphone. This metaphor, where users move the device to detect virtual sound sources, allows quick orientation and easy discrimination between proximate sources, even with simplified rendering algorithms. We compare the navigation performance of head orientation measurement to AudioTorch. A lab study with 18 users showed the rate of correctly recognized sources to be significantly higher with AudioTorch than with head-tracking, while task completion times did not differ significantly. The presence in the virtual environment received similar ratings for both conditions.

Author Keywords
Virtual Audio Spaces; Mobile Devices; Audio Augmented Reality; Navigation.

ACM Classification Keywords
H.5.1. [Information Interfaces and Presentation (e.g. HCI)]: Multimedia Information Systems

Introduction

Audio augmented reality systems overlay a physical space with a virtual audio space that the users experience via headphones. This audio space contains several virtual sound sources which, through special filtering, are perceived as emerging from the physical space. These virtual sources can be part of a game [11], contain information about certain assets [1, 2, 15], or be used as a navigational aid that does not rely on the visual sense [4, 6, 10, 13]. To create a realistic experience with sources that appear to be fixed in the physical space independent of the orientation of the listener's head, this orientation has to be measured. This requires a digital compass mounted to the headphones or some other tracking technology. This additional hardware has to be available, powered, and connected to the device rendering the spatial audio. While early implementations used external hardware to render the audio [7, 14], current implementations rely on smartphones for a personal and decentralized rendering [2, 12, 13]. Since current smartphones are equipped with an internal compass, all necessary components are available in the user's hand, and by omitting the need for additional hardware, the distribution of mobile audio augmented reality systems (MAARS) is reduced to a simple download. We propose to use the smartphone as a virtual directional microphone, a metaphor that is not realistic but still results in an engaging experience. We evaluated this metaphor and compared it against the same rendering using head orientation data in a navigation task. A study with 18 participants revealed that the number of correctly detected sound sources is significantly higher with AudioTorch while the task completion time does not differ significantly. The perceived presence is rated similarly high for both conditions.

Related Work

Different aspects of interaction with virtual audio spaces have been analyzed over the last two decades, but most projects concentrate on aspects of the implementation, e.g., rendering quality or tracking latency.

Loomis [7] and Mariette [9] analyzed the paths and head orientation of people walking towards virtual sound sources. Results show that in the case of a large space, users do an initial, large head turn to get an estimate of the direction, followed by smaller movements to stay on the path towards the source.

Heller et al. [3] conducted a navigation experiment comparing head tracking to device tracking. Results indicate that using device orientation does not dramatically influence the navigation performance nor the perceived presence. During the trials, participants were not told which source of orientation was used. Measurements show that the device was aligned to the body most of the time, meaning that the participants did not take advantage of the device's mobility. By telling the users which tracking is active, we want to encourage them to move the device around to simplify navigation.

Marentakis et al. [8] evaluated pointing as an interaction technique in virtual audio spaces. While walking, participants experienced a sound coming from somewhere around them and had to point to that source. Whenever the heading of the pointing gesture was within a certain angle from the actual position of the source, a feedback sound was presented to facilitate the task. Results show that this technique is feasible to interact with virtual audio spaces, e.g., auditory menus where the items are arranged spatially around the head [5]. While our interaction is similar, we do not focus on targeting a certain sound source, but to create an auditory image of

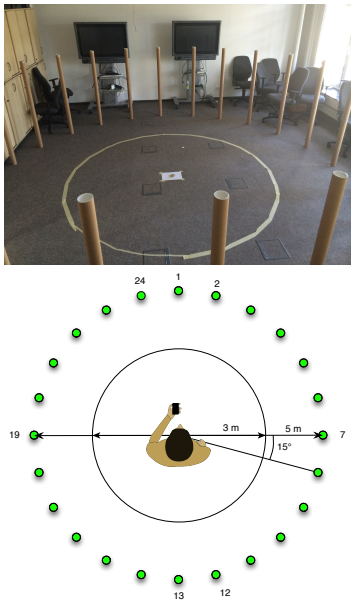

Figure 2: We placed 24 virtual sound sources, spaced by 15° in a circle of 5 m diameter. Participants had to start every trial standing in the center, facing source no. 1. They could move freely within the inner circle of 3 m diameter.

the source positions for navigation. The work presented in this paper fits in between systems focussing on realism on the one hand [9, 14], and systems that do not integrate heading information into the rendering [4, 15, 10].

Implementation

We measured the head position 34 times per second using a Ubisense[1] location tracking system with an accuracy of 5-10 cm. The head orientation was measured using an HMC6343 tilt-compensated digital compass with an update rate of 10 Hz, device orientation was measured with the on-board chip of the iPhone 4S which has an update rate of approximately 15 Hz. Both sensors have similar characteristics and the heading information was adjusted to be the same for both. This data was fed into the spatial audio rendering algorithm integrated to the OpenAL framework in iOS 5.1.1. It bases on the spherical head model and includes the following filter factors: interaural level difference, interaural time difference, head filtering, and a frequency-dependent distance filtering. To reduce the impact of front-backward confusion, we applied a low-pass filter to the sources that were behind the user's head. The filter intensity was interpolated linearly from 0dB to 36dB for sources between 90° and 180° azimuth angle. This algorithm, although less realistic than high-end systems based on head-related transfer functions (HRTF), is a good representative of spatial rendering algorithms for mobile devices.

Evaluation

We placed 24 carton tubes of 150 cm height in a circle with 5 m diameter and spaced by 15° (cf. Fig. 2). The tubes were used as physical representations of the virtual sound sources. The audio rendering algorithm used the orientation either from the head or from the device.

[1] http://www.ubisense.net

Participants started every trial standing in the center of the circle facing source no. 1 and were instructed to find the source currently audible as quickly as possible. In a real scenario, e.g., a museum or a public place, users might not be able to get close to the sources. To account for this factor, and to make sure that the experiment reveals the impact of orientation measurement, we restricted the movement to an inner circle of 3 m diameter, which forced the participants to determine the exact sound source from a distance of approximately 1 m. We used an audio sample of a male voice reciting colors at a fast pace, such that it can be considered a continuous audio source.

We used a within-subjects design with a balanced order of conditions and the order of active sound sources was randomized using Latin squares. Every participant had to navigate to all 24 sources in the *head-* and *device*-measurement condition and had to complete a 10-trial training before each condition. We measured the time between the users starting the trial by pressing the start button on the smartphone until they confirmed standing in front of the audible source by pressing a second button. Participants had to name the source that they assumed was playing. We recorded the paths the users took to walk to the sources, along with the orientation fed into the rendering algorithm. After each condition, participants had to fill out a questionnaire asking about their perceived presence in the virtual environment [16]. Answers had to be reported on a 5-point Likert scale.

Results

We recruited 18 participants (5 female) to complete the study. Their age ranged from 22 to 38 (M=28) and most had a computer science background. The rate of correctly

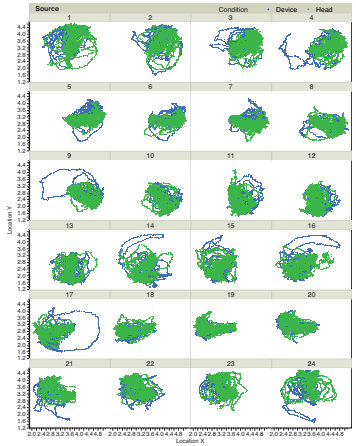

Figure 3: The recorded paths of participants localizing the audible source. There is no substantial difference between both conditions (except for one outlier in the device condition).

localized sources was higher in the *device* condition (M=0.76, SD=0.21) than in the *head* condition (M=0.5, SD=0.17). A pairwise student's-t test showed this difference to be significant ($t(18)$=-4.09, p=0.0003). However, nearly all errors are only off by one source from the correct result. If we consider a source to the left or to the right of the actual source as correct, then we achieve recognition rates of 98% for device tracking and 96% for head tracking. The task completion time was nearly the same in both conditions (*Head*: M=9.92, SD=4.01; *Device*: M=10.11, SD=4.11). Plotting the paths taken to the sources did not reveal substantial differences (cf. Fig 3). The cumulative distance during head-tracking trials (M=8.28 m, SD=3.28) was nearly identical to the device-tracking trials (M=8.32 m, SD=2.63). An analysis of variance showed the difference to be non-significant ($F(1)$=0.0438, p=0.8342).

None of the ratings of the questionnaire showed to be significantly different between the two conditions, which indicates that the perceived presence is equal. The ability to localize sounds nearly received the same overall rating for head-(M=4.0, *95% CI*=[3.7, 4.3]) and device-tracking (M=4.1, *CI*=[3.6, 4.5]).

Not surprisingly, the interaction with the virtual environment was perceived to be more natural using the compass on the head (M=4.2, [3.7, 4.7]) than with device tracking (M=3.9, [3.3, 4.5]), and also less distracting (head: M=4.2, [3.5, 4.9]; device: M=4.0, [3.3, 4.5]). Users felt very proficient in moving and interacting with the virtual environment at the end of the experiment in both conditions (head: M=4.3, [4.0, 4.6]; device: M=4.5, [4.2, 4.8]) and they could concentrate very well on the task rather than on the controls used to perform the task (head: M=4.7, [4.4, 5.0]; device: M=4.2, [3.8, 4.7]).

During the experiments we observed that the orientation was not necessarily used as principal cue to discriminate between two possible sources. Instead, participants used lateral movements and evaluated the changes in the audio signal. Some participants reported the head-tracking condition to be easier to adapt to, although their recognition rate was lower than in the device-tracking condition. As analyzed in [3], most of the participants held the device and head aligned with their body in both conditions.

Discussion

The recognition rate in the head tracking condition is surprisingly low. In contrast to device-tracking, using head orientation tries to reproduce an experience that we know from everyday life. We assume that this makes us very sensitive to anything unnatural, such as lag or error in the tracking data. Using the device as virtual directional microphone requires the users to develop some new technique that possibly includes compensation for the measurement error. The localization process is thus more analytical than with head-tracking. Several users reported that they stopped in front of a candidate source and then checked if the audio representation fit to that source by rotating the device or their arm. We also observed that several participants used lateral movement instead of rotation for this final check. Using optical tracking and high-end rendering hardware would probably lead to different results, however, these are not feasible for MAARS used in everyday life. Since with our setup the errors are mostly off by one source, placing the sources at approximately 1.3 m would be sufficient to ensure a correct recognition.

A technical difficulty we faced during the experiments was the drift of the magnetometers. The reported heading can

differ by as much as 15° from one day to the other. We ensured that both sensors reported the same heading by recalibrating them every day before the experiments. However, this cannot be guaranteed for devices in the wild. Some participants held the device slightly rotated in front of their body, which results in the measurement being off-axis from the users perspective. Both observations support our recommendation of placing the sources further apart from each other than in our experiment.

Conclusion

From our experiment we can conclude that the AudioTorch metaphor is a suitable replacement for head tracking as it reduces the technical complexity of a mobile audio augmented reality system without negatively influencing the presence in the virtual environment nor the source recognition rate. If the exact localization of the audible source is of importance, the sources should be placed farther away from each other than the 65 cm in our experiment. Taking into account the inherent error in orientation and location measurement with wireless sensors, the distance should be more in the order of 1.3 m.

The AudioTorch metaphor is ideal for museums because it allows to distribute a next generation audio guide as a simple app. Indoor location tracking could be realized using bluetooth beacons, while outdoor location tracking can be realized using GPS. Another field is a navigation aid for visually impaired. The device could either detect certain assets through its camera or have pre-programmed locations that are augmented with an audio-source. For example, this would allow to easily differentiate between ATMs and vending machines from a distance.

Future Work

Further evaluation should analyze how the inaccuracies of GPS measurements influence the use of AudioTorch in outdoor navigation. The distance two sources should have from each other to be distinguished reliably is another interesting question. Furthermore, the effect of physical landmarks for the sound sources should be evaluated. In our study, it was clear which objects were augmented with a virtual sound source, however, in a real world application, this might not be the case. Additionally, without physical representation, the effect of tracking inaccuracies is mitigated, as long as the impression remains consistent.

This experiment was performed with the participants holding a smartphone in their hand. While this is part of the AudioTorch metaphor, further experiments could be conducted without a device in the head-tracking condition to see whether this has an influence on the recognition rate.

Since the vertical resolution of spatial hearing and spatial audio rendering is not as high as the lateral resolution, it is mostly ignored in audio augmented reality. However, with AudioTorch, searching for sound sources at different heights is possible through simple volume changes, and does not require individual HRTFs.

From an interaction perspective, it would be interesting to analyze in depth when users do notice if the orientation measurement comes from the device or the head.

Acknowledgments

We would like to thank all the participants of our study. This work was funded in part by the German B-IT Foundation.

References

[1] Bederson, B. B. Audio augmented reality: a prototype automated tour guide. In *Conference Companion of CHI '95*, ACM (1995).

[2] Heller, F., Knott, T., Weiss, M., and Borchers, J. Multi-user interaction in virtual audio spaces. In *Proc. CHI EA '09*, ACM (2009).

[3] Heller, F., Krämer, A., and Borchers, J. Simplifying Orientation Measurement for Mobile Audio Augmented Reality Applications. In *Proc. CHI '14*, ACM (2014).

[4] Holland, S., Morse, D. R., and Gedenryd, H. AudioGPS: Spatial Audio Navigation with a Minimal Attention Interface. *Personal and Ubiquitous computing 6*, 4 (Jan. 2002).

[5] Kajastila, R., and Lokki, T. Eyes-free methods for accessing large auditory menus. In *Proc. ICAD '10* (2010).

[6] Lasorsa, Y., and Lemordant, J. An Interactive Audio System for Mobiles. In *127th Convention of the Audio Engineering Society* (2009), 1–9.

[7] Loomis, J. M., Hebert, C., and Cicinelli, J. G. Active localization of virtual sounds. *The Journal of the Acoustical Society of America 88* (1990), 1757.

[8] Marentakis, G. N., and Brewster, S. A. Effects of feedback, mobility and index of difficulty on deictic spatial audio target acquisition in the horizontal plane. In *Proc. CHI '06*, ACM (2006).

[9] Mariette, N. Navigation performance effects of render method and head-turn latency in mobile audio augmented reality. In *Auditory Display*. Springer Berlin Heidelberg, 2010, 239–265.

[10] McGookin, D., Brewster, S., and Priego, P. Audio Bubbles: Employing Non-speech Audio to Support Tourist Wayfinding. In *Haptic and Audio Interaction Design*. Springer Berlin Heidelberg, 2009.

[11] Paterson, N., Naliuka, K., Jensen, S. K., Carrigy, T., Haahr, M., and Conway, F. Design, implementation and evaluation of audio for a location aware augmented reality game. In *Proc. Fun and Games '10*, ACM (2010).

[12] Sander, C., Wefers, F., and Leckschat, D. Scalable Binaural Synthesis on Mobile Devices. In *133rd Conference of the Audio Engineering Society* (2012).

[13] Stahl, C. The roaring navigator: a group guide for the zoo with shared auditory landmark display. In *Proc. MobileHCI '07*, ACM (2007).

[14] Terrenghi, L., and Zimmermann, A. Tailored audio augmented environments for museums. In *Proc. IUI '04*, ACM (2004).

[15] Wakkary, R., and Hatala, M. ec(h)o: situated play in a tangible and audio museum guide. In *Proc. DIS '06*, ACM (2006).

[16] Witmer, B. G., and Singer, M. J. Measuring Presence in Virtual Environments: A Presence Questionnaire. *Presence: Teleoper. Virtual Environ. 7*, 3 (1998), 225–240.