AudioScope: Smartphones as Directional Microphones in Mobile Audio Augmented Reality Systems

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

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

ABSTRACT

Mobile audio augmented reality systems (MAARS) provide a new and engaging modality to present information or to create playful experiences. Using special filters, spatial audio rendering creates the impression that the sound of a virtual source emanates from a certain position in the physical space. So far, most of the implementations of such systems rely on head tracking to create a realistic effect, which requires additional hardware. Recent results indicate that the built-in sensors of a smartphone can be used as source for orientation measurement, reducing deployment to a simple app download. AudioScope presents an alternative interaction technique to create such an experience, using the metaphor of pointing a directional microphone at the environment. In an experiment with 20 users, we compared the time to locate a proximate audio source and the perceived presence in the virtual environment. Results show that there is no significant difference between head-orientation measurement and AudioScope regarding accuracy and perceived presence. This means that MAARS, such as audio guides for museums, do not require special hardware but can run on the visitor's smartphones with standard headphones.

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 the physical space with a virtual audio layer that is perceived through headphones. Spatial audio rendering creates the impression that the sound of the virtual sources emerges from a certain location in the physical space. Such systems are used, e.g., to augment points of interest in museums, or to provide a navigational aid that does not put additional load on the visual sense [13, 15, 16]. While early audio augmented reality systems used external, dedicated hardware to render the spatial audio [15], today's smartphones provide enough processing power for individual, decentralized rendering [12].

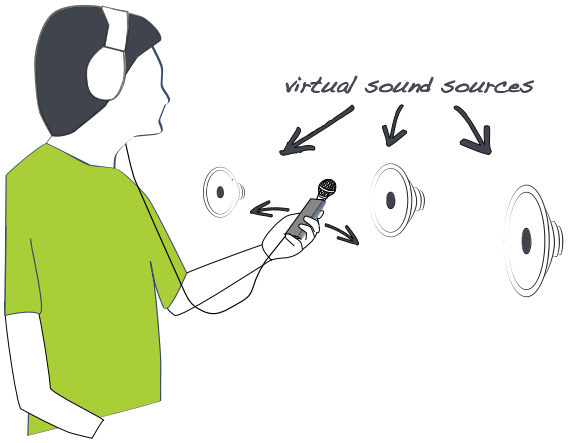

Figure 1. AudioScope uses the built-in sensors of a modern smartphone to create the illusion of a virtual directional microphone to explore mobile audio augmented reality environments.

To create a realistic experience in which the sources appear to be located in the physical space independent of the user's orientation, this orientation has to be measured. Most implementations rely on head tracking as the source of orientation for the spatial audio rendering algorithm, which requires additional hardware, such as a digital compass or intertial measurement unit (IMU). This hardware needs to be attached to the headphones, powered, and connected to the device rendering the spatial audio. The additional costs and handling slow down the distribution of mobile audio augmented reality systems (MAARS).

Recent results indicate that using the device compass is a feasible option [6], however, the resulting experience might not correlate with the users' expectations. We propose AudioScope, a metaphor that turns your smartphone into a virtual directional microphone. You can probe the audio space by simply pointing the device in different directions. If the sound source is to the left of the device, the sound on the left audio channel is louder and vice versa. By providing a well-defined mental model, we avoid disappointing users through a possible lack of realism as they discover that turning their head does not influence the audio output. We compared AudioScope with an implementation using a commercial head tracking system in a source localization experiment with 20 participants. While on average being 1.5 s slower in task completion time, AudioScope is on par with the headphone based measurement regarding accuracy and perceived presence in the virtual environment. This allows designers to deploy fu-

CHI 2015, April 18–23, 2015, Republic of Korea.

ACM 978-1-4503-3145-6/15/04...$15.00.
http://dx.doi.org/10.1145/2702123.2702159

ture MAARS as standard apps on unmodified smartphones, without the need for additional special sensor hardware.

RELATED WORK

Spatial audio rendering and the interaction with virtual audio spaces have been of interest to the research community for over three decades [2]. Most of the navigation tasks, however, included walking over a longer distance, which is reasonable for outdoor navigation, but not directly applicable to settings where the sources are close, e.g., in a museum.

Loomis et al. [8] examined the effect of spatial audio rendering algorithms on paths users take when walking towards virtual sound sources, and concluded that even simple algorithms that do not simulate pinnae effects can externalize sounds well enough to allow successful navigation. Marietta [10] later replicated this experiment with different rendering algorithms and head-tracker latencies and observed significant degradation of source stability ratings and effects in the recorded paths for an added latency of 800 ms and higher. At the same time, this only affected the better of the two tested rendering algorithms, while the simpler one covered the effects due to its lower angular resolution.

Heller et al. [6] analyzed differences in orientation between the head, the chest, and a smartphone while navigating in a virtual audio space. They found that, except for a large initial head-turn, the measures of head and device do not differ much. This suggests device orientation as a promising point of measurement for mobile audio augmented reality applications. However, even if results indicate that navigation performance is not affected, users might still be confused if the measurements are taken at a different point than they expect. During their trials, participants were not informed which source of orientation was used, and the results show that the device was aligned to the body most of the time, thus the participants did not take advantage of the device's mobility.

Marentakis et al. [9] successfully evaluated pointing as an interaction technique in virtual audio spaces. Participants had to point to an audible source with their arm while walking. A feedback sound was played whenever the arm was in certain range around the source to facilitate the task. Deo et al. compared pointing with a mobile phone to head tracking for multichannel audio conferencing [3]. Building on that, a gesture interface with a mobile phone was also tested to navigate trough a two-level spatial auditory menu [5] and the pointing metaphor was applied to this type of menu [7]. Results showed this technique to be feasible to interact with auditory menus where the items are arranged spatially around the head. While our interaction is similar, we do not focus on targeting a certain sound source, but on creating an auditory image of the source positions to navigate in an audio space.

IMPLEMENTATION

We measured head position at 34 Hz using a Ubisense (ubisense.net) location tracking system with an accuracy of around 5 cm. Head orientation was measured with the inertial measurement unit (IMU) of an Intelligent Headset (intelligentheadset.com), while device orientation was measured using the IMU of an iPhone 5S. We compared both IMUs,

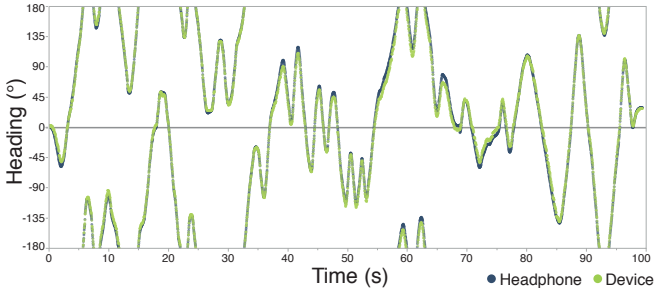

Figure 2. We compared the IMU of the Intelligent Headset (blue) and the iPhone 5S (green). Both report heading with a similar characteristic and update rate.

which report changes in heading with an update rate of around 40Hz, and found that they have a similar characteristic, with an average difference of only 4.8° (SD=3.4°) (cf. Fig. 2). The absolute average orientation error of the iPhone IMU is 4.25° (SD=3.05°). The specified overall latency of the headphone orientation measurement is around 100 ms, which is noticeable [4] but well below the limits of 372 ms defined in [10].

Spatial audio rendering was implemented using the OpenAL framework in iOS 7.1, with the `ALC_EXT_MAC_OSX` extension enabled. It uses a spatialization based on the spherical head model and includes interaural level and time difference, head filtering, and a frequency-dependent distance model as filters. To enhance front-back separation of sources, we added a low-pass filter to sources when they are behind the listener. The intensity increases linearly from 0dB to -36dB for azimuth angles between 90° (side) and 180° (back). The minimum audible angle of the rendering is around 4°. This method, although less realistic than algorithms using individual, natural body cues in form of head-related transfer functions (HRTF), is a good representative of spatial rendering on mobile devices.

EVALUATION

In our experiment, participants were instructed to navigate to single proximate sources with either head or device tracking enabled. We placed 24 loudspeakers spaced by 15° at a height of 150 cm forming a circle of 5 m diameter (Fig. 3). As in this experiment we only wanted to compare the impact of the orientation measurement using the same rendering in both conditions, the loudspeakers did not play any sound but were mere physical representations of the virtual sound sources.

In the two conditions of our experiment, the audio rendering algorithm used the orientation either from the *head* or from the *device*. Participants started every trial standing in the center of the circle facing source no. 1 and were instructed to identify the currently active source as quickly as possible. Correct alignment of the heading information with the physical setup was verified before each trial. 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 revealed the impact of orientation measurement, participants were instructed to move only in an inner circle of 3 m diameter, such that they had to determine the exact sound source from a distance of approximately

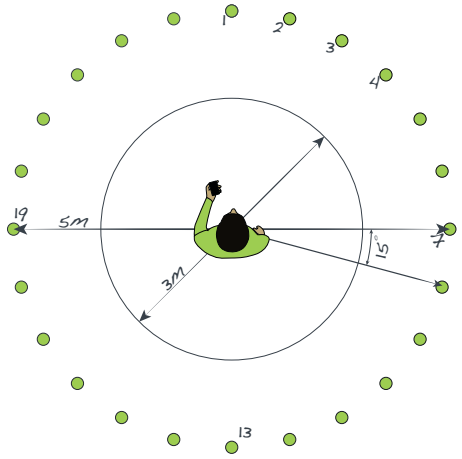

Figure 3. We placed 24 virtual sound sources, spaced by 15° in a circle of 5 m diameter. Participants had to start every trial standing in the center, facing source no. 1. They could move freely within the inner 3 m circle.

1 m. We used an audio sample of a male voice reciting colors at a fast pace[1].

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 from users starting each 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 about their perceived presence in the virtual environment [18] on 5-point Likert scales.

In total, 20 users, 8 female, 12 male, aged 21 to 33 (average 26), participated in the study. None reported a hearing disorder or known problems with spatial hearing. Seven had prior experience with audio augmented reality systems.

RESULTS

The average time users took to navigate to the sound source was 15.71 s (SD=8.51) in the *head* condition and 17.22 s (SD=9.72) in the *device* condition. A mixed model repeated measures ANOVA revealed this difference to be statistically significant ($F(1, 916)$=14.79, p=.0001). At the same time, the rate of correctly recognized sources was 65% (SD=0.28) for head tracking and 69% (SD=0.26) for device tracking. This difference is not significant (p=.91) according to a Wilcoxon Signed Rank test. The recognition rates seem fairly low, but considering that the 15° spacing between our sources is in the range of the localization error of virtual sound sources [11, 17] and that we used a rather simple rendering algorithm, this is not surprising. Most of the errors were only off by one source to the left or right. If we count these "off-by-one" answers as correct, then the recognition rates climb up to 97% (SD=0.1) for head and 98% (SD=0.1) for device tracking. No front-back confusions occurred. While people without prior experience with audio AR where significantly slower in the device condition (M=14.3s, SD=7.5s vs. M=18.8s, SD=10.4s, p=.0003), no significant difference between conditions could be found for participants with prior experience. The average distance travelled only differs by half a step between conditions (*device*: M=9.00m, SD=5.4m, *head*: M=8.7m, SD=6.32m, p=.0335).

The median ratings from the perceived presence questionnaire did not differ by more than one item on the 5-point Likert scale. Not surprisingly, the headphone tracking was perceived more natural than the device (*head*: Mdn=5, IQR=1.75; *device*: Mdn=4, IQR=1). For both conditions, the experience was rated to be consistent with the real world (*head*: Mdn=4, IQR=1.75; *device*: Mdn=4, IQR=2), and participants felt able to localize sounds well (*head*: Mdn=4, IQR=0; *device*: Mdn=4, IQR=0.75). Wilcoxon signed rank tests only revealed the ratings for the *natural interface* to be significantly better (p=.022) for the *head* tracking, all other ratings did not differ significantly between conditions.

The tracking technology did not seem to interfere with the experience as both the responsiveness (*head*: Mdn=5, IQR=1; *device*: Mdn=5, IQR=1) and the perceived delay (*head*: Mdn=4, IQR=1.75; *device*: Mdn=4, IQR=2) received similarly high ratings in both conditions. After the experiment, the participants felt proficient with the interface both with head tracking (Mdn=4, IQR=1) and device tracking (Mdn=4.5, IQR=1). Again, no significant difference was found between the conditions.

DISCUSSION

Overall, the differences between both orientation tracking metaphors are fairly small. Out of the answers on the questionnaire, 85% of the ratings are above 3 out of 5 on a Likert scale (5 being the best). We are thus confident that the acceptance of the device orientation measurement is high. For people with prior audio AR experience, no significant difference in task completion time could be found which suggests that the metaphor is easy to adopt. The fact that the average task completion time was slightly longer (1.5 s) in the *device* condition is not critical in practice. Other studies have revealed that a longer task completion time can also be a result of people enjoying the experience [16], which in case of an audio guide for museums, is the primary focus. Furthermore, we observed that users experimented with the handling of the device tracking even though they completed the 10 training trials prior to the experiment.

The low recognition rates show that the distance between the sources was at the limit of what can be differentiated with our rendering. As stated above, the 15° spacing is in the range of the localization error of virtual sound sources. Participants spent more time in a 1.5 m radius around the active source than in the rest of the field, which indicates that they took a long time to differentiate between two candidate sources. This problem can be solved by either placing the sources further apart, or by providing additional context, e.g., a beacon sound that relates to the physical object.

[1] http://hci.rwth-aachen.de/public/AudioScope/Colors.au

Some participants stated that they felt faster with the device tracking, and one mentioned the head tracking being more difficult immediately after switching conditions. On the other hand, some participants mentioned being confused by the metaphor of the virtual directional microphone, since when moving the smartphone to the right of a source, the left channel becomes louder.

The focus of this paper is the evaluation of the interaction metaphor, for which the use of an external location tracking system is acceptable. However, indoor location tracking is currently a focus of both researchers and smartphone manufacturers. Technologies such as Estimote beacons (estimote.com) support that we can expect significant improvements in accuracy in the near future, making smartphones a complete platform for MAARS.

CONCLUSION & FUTURE WORK

We presented AudioScope, a system that uses the metaphor of a directional microphone to explore virtual audio spaces. The simple mechanism of pointing in different directions to locate sound sources is easy to understand and avoids confusion that might occur when simply replacing the head orientation measurement by device orientation without communicating this. Using built-in sensors of the device, AudioScope reduces the hardware requirements and allows deploying mobile audio augmented reality systems via a simple app download.

Together with the current integration of depth sensing cameras into smartphones, the implementation of assistive systems for the blind such as presented in [14] could be reduced to a simple download. Using AudioScope with an HRTF-based rendering that also simulates elevation, finding an item in a shelf can be done in a simple point-and-pick action.

Future studies should take a closer look at emerging indoor location technologies and measure their impact on navigation performance. For outside navigation, we can rely on assisted GPS which has been used successfully in a series of related projects. Furthermore, current spatial audio rendering algorithms for mobile devices usually do not simulate accurate rooms acoustics. Using these would make the implementation more difficult since the impulse response of the room would need to be measured, but the externalization of sounds and thus localization accuracy should increase [1].

ACKNOWLEDGMENTS

We would like to thank the participants of our study and Thorsten Karrer for his valuable feedback. This work was funded in part by the German B-IT Foundation.

REFERENCES

1. Begault, D. R., Wenzel, E. M., and Anderson, M. R. Direct Comparison of the Impact of Head Tracking, Reverberation, and Individualized Head-Related Transfer Functions on the Spatial Perception of a Virtual Speech Source. *J. Audio Eng. Soc* (2001).
2. Begault, D. R., Wenzel, E. M., Godfroy, M., Miller, J. D., and Anderson, M. R. Applying Spatial Audio to Human Interfaces: 25 Years of NASA Experience. In *AES Conference 40* (2010).
3. Billinghurst, M., deo, S., Adams, N., and Lehikoinen, J. Motion-Tracking in Spatial Mobile Audio-Conferencing. In *MobileHCI '07* (2007).
4. Brungart, D. S., Simpson, B. D., and Kordik, A. J. The detectability of headtracker latency in virtual audio displays. In *ICAD '05*.
5. Dicke, C., deo, S., Billinghurst, M., Adams, N., and Lehikoinen, J. Experiments in Mobile Spatial Audio-conferencing: Key-based and Gesture-based Interaction. In *MobileHCI '08*.
6. Heller, F., Krämer, A., and Borchers, J. Simplifying Orientation Measurement for Mobile Audio Augmented Reality Applications. In *CHI '14*.
7. Kajastila, R., and Lokki, T. Eyes-free methods for accessing large auditory menus. In *ICAD '10*.
8. Loomis, J. M., Hebert, C., and Cicinelli, J. G. Active localization of virtual sounds. *The Journal of the Acoustical Society of America* (1990).
9. 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 *CHI '06*.
10. Mariette, N. Navigation Performance Effects of Render Method and Head-Turn Latency in Mobile Audio Augmented Reality. In *Auditory Display*. Springer, 2010.
11. Middlebrooks, J. C. Virtual localization improved by scaling nonindividualized external-ear transfer functions in frequency. *The Journal of the Acoustical Society of America* (1999).
12. Sander, C., Wefers, F., and Leckschat, D. Scalable Binaural Synthesis on Mobile Devices. In *AES Convention 133* (2012).
13. Stahl, C. The roaring navigator: a group guide for the zoo with shared auditory landmark display. In *MobileHCI '07*.
14. Tang, T. J. J., and Li, W. H. An Assistive EyeWear Prototype That Interactively Converts 3D Object Locations into Spatial Audio. In *ISWC '14*.
15. Terrenghi, L., and Zimmermann, A. Tailored audio augmented environments for museums. In *IUI '04*.
16. Vazquez-Alvarez, Y., Oakley, I., and Brewster, S. Auditory display design for exploration in mobile audio-augmented reality. *Pers. and Ubiqu. comp.* (2012).
17. Wenzel, E. M., Arruda, M., Kistler, D. J., and Wightman, F. L. Localization using nonindividualized head-related transfer functions. *The Journal of the Acoustical Society of America* (1993).
18. Witmer, B. G., and Singer, M. J. Measuring Presence in Virtual Environments: A Presence Questionnaire. *Presence: Teleoper. Virtual Environ.* (1998).