Audio Augmented Reality for Human-Object Interactions

Jing Yang
jing.yang@inf.ethz.ch
ETH Zurich, Switzerland

Friedemann Mattern
mattern@inf.ethz.ch
ETH Zurich, Switzerland

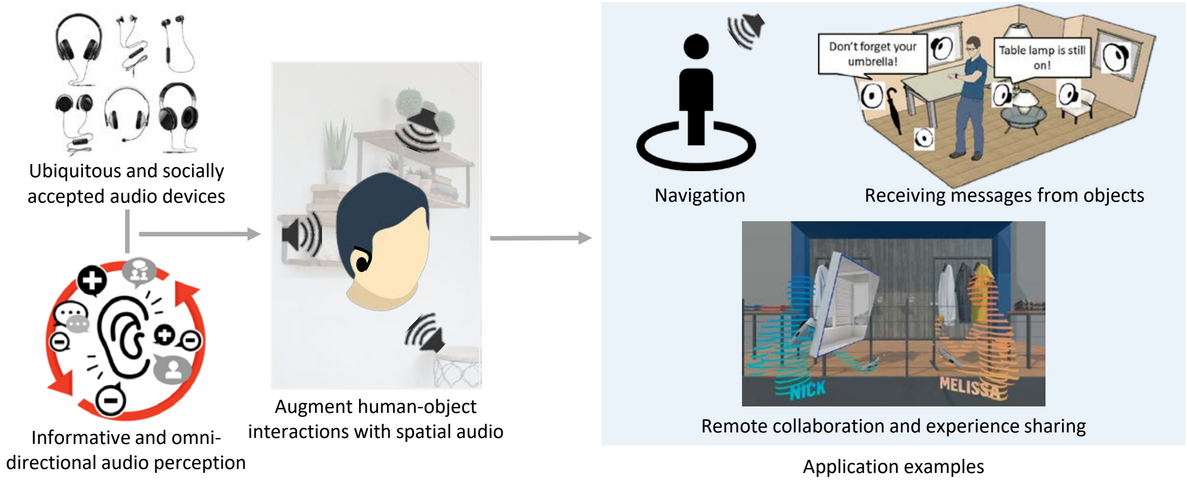
Ubiquitous and socially accepted audio devices
Informative and omni-directional audio perception
Augment human-object interactions with spatial audio
Navigation
Don't forget your umbrella!
Table lamp is still on!
Receiving messages from objects
NICK
MELISSA
Remote collaboration and experience sharing
Application examples

Figure 1: The ubiquitous and socially accepted audio devices and the importance of human beings' auditory sense indicate the opportunity to augment human-object interactions with synthesized spatial sounds. This virtual audio channel is expected to benefit a wide range of applications, including home entertainment, on-site games, remote conference, etc.

ABSTRACT

In the past, augmented reality (AR) research focused mostly on visual augmentation, which requires a visual rendering device like head-mounted displays that are usually obtrusive, expensive, and socially unaccepted. In contrast, wearable audio headsets are already popularized and the auditory sense also plays an important role in everyday interactions with the environment. In this PhD project, we explore audio augmented reality (AAR) that augments objects with 3D sounds, which are spatialized virtually but are perceived as originating from real locations in the space. We intend to design, implement, and evaluate such AAR systems that enhance people's intuitive and immersive interactions with objects in various consumer and industrial scenarios. By exploring AAR using pervasive and wearable devices, we hope to contribute to the vision of ubiquitous AR.

CCS CONCEPTS

• **Human-centered computing** → **Mixed / augmented reality**; **Auditory feedback**; *Collaborative interaction.*

KEYWORDS

Audio Augmented Reality; Spatial Audio; Human-Object Interaction; Wearable

ACM Reference Format:
Jing Yang and Friedemann Mattern. 2019. Audio Augmented Reality for Human-Object Interactions. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and the 2019 International Symposium on Wearable Computers (UbiComp/ISWC '19 Adjunct), September 9–13, 2019, London, United Kingdom.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3341162.3349302

1 PROBLEM STATEMENT

Motivation

The technology augmented reality (AR), referring to a spatio-temporal experience where physical objects are augmented

UbiComp/ISWC '19 Adjunct, September 9–13, 2019, London, United Kingdom
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6869-8/19/09…$15.00
https://doi.org/10.1145/3341162.3349302

with digital information, enhances our interactions with physical surroundings in various industrial, commercial, and home environments. The surveys on the past 20 years of AR research by Dey et al. [8] and Kim et al. [11] revealed that an overwhelming majority of research focused on visual augmentation, while augmenting other human senses has remained less explored. Devices for visual augmentation *(graphics rendering)* such as hand-held devices and head-mounted displays are usually rather cumbersome, obtrusive, and/or socially less accepted, and hence impede the development of ubiquitous AR.

On the contrary, wearable audio headsets and ear buds have already blended in our everyday life for entertainment, noise cancellation, hearing aid, and telecommunication anywhere on the go. Plus, the auditory sense is important for our everyday interactions with the surroundings. As human beings, we are capable of not only interpreting the content of a sound signal, but also localizing its origin in direction and distance thanks to binaural hearing [4]. Therefore, in addition to giving us information, the auditory sense can also provide us with omni-directional engagement through an immediate 360° sense of space, time, and presence, which complements the visual perception inside as well as outside our visual field of view.

From the above, the ubiquity of audio devices *(sound rendering)* and the importance of the auditory sense indicate a significant opportunity for audio augmented reality (AAR) research based on spatial audio, which encompasses the notion of sound signals that can be perceived to have a pronounced direction and distance. Nowadays, spatial audio has been extensively utilized in virtual reality (VR) games to create immersive experiences for players. We anticipate that a virtually created soundscape can also be applied in AR to enhance the way we interact with real objects. However, due to the difficulty of simulating the sound propagation in arbitrary environments in real time, it is still under explored in AR.

In this thesis, we explore new technologies and methods to generate and add virtual 3D sounds to real objects, especially to those that are not equipped with a real loudspeaker, and the sounds are perceived using normal headphones. Upon hearing the spatialized audio signals, users can perceive them as originating from particular locations in the space. This enables a new output channel and new interaction possibilities with real objects.

Research Questions

In this project, our main objective is to design, implement, and evaluate such an AAR system that supports people interacting with arbitrary everyday objects using the virtual spatial audio channel. To promote ubiquitous AR, we set low-obtrusive wearable implementation as a goal. We aspire to address the following research questions (RQ):

1. (a) What are the necessary functional components of such a system? (b) What methods do we need to advance to implement the system?
2. (a) Do people perceive AAR-based interactions as useful and pleasant experience? (b) How accurate and fast are people when guided only by the synthesized sounds to localize the sources?
3. (a) How can we realize such an AAR system utilizing wearable devices in real time? (b) How can we improve the sense of presence for users when using the system?
4. (a) What kind of applications can benefit from such AAR-based interactions? (b) Can we involve other modalities to facilitate more seamless interactions? (c) Can we utilize AAR to enhance the interaction among multiple users?

2 RELATED WORK

In previous research, people have explored the application of virtual spatial soundscape to redirect a user's attention [1, 5, 10, 15], to provide notifications from a specific location [3, 9, 18–20], and to create new music experience [12, 13]. Some projects [9, 19, 20] investigate the functionality of spatial audio cues. They set restrictions on the sound source positions (e.g., equidistantly positioned on a circle) and application space (e.g., on a small tabletop), therefore their setup is not extensible to general scenarios. Some projects propose systems for specific applications, such as outdoor navigation [15], guiding visually impired people [5], or warning about incoming vehicles [18].

We expect to develop an AAR system that can synthesize 3D sounds for arbitrary objects in the surroundings and we focus on indoor environments. Based on our current work, we summarize that basic system components should involve *recognizing the objects to be auralized, tracking the user's (head) pose relative to the objects,* and *rendering and playing back spatialized sound signals*, among which a precise user-object pose tracking is the core to create authentic spatial audio experience. To this end, existing projects utilize environment cameras [13], head-mounted cameras [18], magnetic or radio-frequency modules [15], or headphone-integrated motion sensors [9]. In our work, one focus is to explore light-weight wearable and robust pose tracking methods that work in real time. For the other two components we will mainly apply existing techniques.

Inspired by the work that demonstrates distinct auditory perceptions when the sounds are rendered with the same room geometry but different surface materials [16, 17], we believe one approach to improving the user's sense of presence is to blend the spatialized sound with real environment acoustic effects. Simulating appropriate indoor acoustics has been leveraged in virtual reality (VR) applications, but in real-life scenarios this is barely explored.

Existing projects imply that AAR-based services are usually implemented in everyday situations such as navigation, notification, and gaming. We also see its potential to be leveraged for industrial or even healthcare contexts. Beyond single-user circumstances, we are also interested to explore how AAR can enhance the interactions among multiple users in, for example, remote collaboration and experience sharing scenarios, in which the current research focuses more on the visual sense. We anticipate that the integration of spatialized sounds or spatialized voice can lead to an enhanced immersive experience.

3 METHODOLOGY

As an application-driven project, we will explore relevant sensors and actuators to prototype our designed AAR system. Along with the implementation, we will also advance and integrate current research in the domains of acoustics, human-computer interactions, and even computer vision. Throughout the whole project, we will evaluate the system modules and real-life applications by designing and conducting user studies.

Research Carried Out So Far

After determining the basic functional components of such a system (RQ1(a)), we first justified the usefulness and usability of AAR-based interactions (RQ2). For this purpose, we built a prototype by utilizing a state-of-the-art pose tracking system Vicon[1]. While not wearable, the Vicon system enables precise tracking which is necessary to validate the idea. Based on the user-object poses captured in real time, we spatialized the corresponding virtual sounds using the Unity3D game engine integrated with the Google Resonance Audio SDK. The synthesized 3D audio was played back to the user via off-the-shelf headphones. To simulate indoor scenarios, we implemented the prototype in a room of size $6.6\,m \times 9.1\,m \times 3.4\,m$ where the sound sources were arbitrarily distributed. Figure 2 illustrates the setup.

We designed the experiments to cover two typical scenarios: interaction with an object at a distance and interaction by reaching the object. We evaluated participants' accuracy and speed of finding the notifying objects by measuring the angle and distance errors of their localization and the time spent on each test. Results showed that the median angle error was as small as even comparable to humans' focused field of view (around 5.2°), and the median distance errors were only 16 *cm* in both horizontal directions. In terms of the user experience, participants generally regarded the whole experience as very interesting and would like to use such a system in several applications, such as navigation, car infotainment, on-site games, and more. Details of this work are in our paper [21].

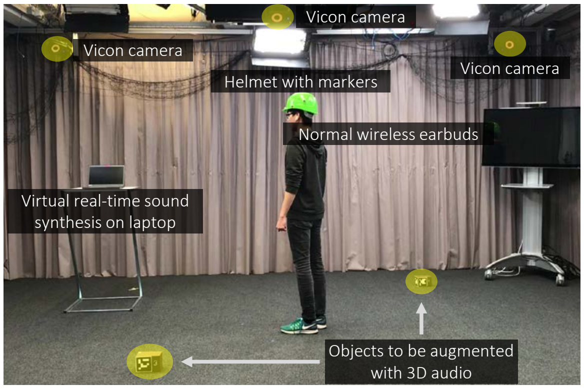
Vicon camera
Vicon camera
Vicon camera
Helmet with markers
Normal wireless earbuds
Virtual real-time sound synthesis on laptop
Objects to be augmented with 3D audio

Figure 2: The setup for exploring RQ2. Participants freely walked around to experience the virtual 3D sound.

This work answers our RQ2. Additionally, in this work we also simulated the room acoustics offline and rendered 3D sounds with auditory effects. Through experiments we have demonstrated the effectiveness of an acoustic model to enhance the sense of presence for users, thus justifying this research direction in the future. Following this work we have been focusing on the wearable implementation of real-time user-object pose estimation (RQ1(b)). We started with exploring inside-out tracking by using light-weight head-mounted cameras running simultaneous localization and mapping (SLAM) or visual-odometry (VO) algorithms [6]. Later, we identified that such tracking approaches tended to drift significantly if people moved their heads fast or abruptly when walking around or responding to unexpected sounds. Therefore, we are integrating motion sensors on top of visual sensors to improve the tracking robustness.

At the same time, to further improve the system's wearability, we are exploring acoustic head tracking approaches by utilizing microphones that are mounted on headphones so they can capture the environment sounds perceived by both ears. By analyzing the feature changes of the input audio signals, we can estimate the user's pose with respect to the object of interest and then simulate the corresponding 3D sound, which is then played to be user via the headphones. For this approach, we need loudspeaker(s) to emit sounds that travel through the room and arrive at both ears with varying patterns along with the user's movement. Previous research has demonstrated the potential of such approaches in room geometry estimation [2, 7] and acoustic indoor positioning [14]. We have obtained preliminary results using a wearable loudspeaker and in-ear microphones to infer head orientations. Further improvement of this approach and its

[1] https://www.vicon.com/

integration into real applications are being investigated at the moment.

Regarding applications (RQ4(a)), we started with museum scenarios where we would like to augment artworks with their content-related spatialized sounds. We have conducted a user study in an on-site simulated gallery scene and explored how the added audio channel could enhance visitors' experience. This work is currently under submission.

Research Planned

Based on the undertaken research and the results so far, one research direction is to improve the sense of presence for users when using the AAR system (RQ3(b)). As discussed above, a promising method is to render virtual 3D sounds with appropriate environment acoustic effects. In our preliminary study, we implemented offline acoustics modeling, but we are interested to explore online approaches. Another aspect to consider is the fusion of the virtual soundscape into the real sounds that already exist in the surroundings, then play the mixed audio signals to the user. Intuitively, this can be done with bone-conduction headsets which leave the ears open to outside sounds. Alternatively, we may combine the audio mixing with indoor acoustics modeling, thus to create more integral sound mixture that can be perceived with normal headphones which cover the ears.

To explore more seamless interactions using AAR with other modalities (RQ4(b)), we are interested to start with integrating the visual sense. One focus is to trigger or tune the sound signals based on the user's gaze direction, which can be detected either from a user-centric view or from an object-centric view, depending on where the camera is mounted. We have summarized the idea together with potential application scenarios in one of our publications [22]. Taking a step further, auditory plus visual augmentation makes a good combination in mixed reality (MR) remote collaboration applications, which would be interesting and useful scenarios to extend our AAR services among multiple users (RQ4(c)). In such scenarios, we may additionally integrate other modalities, such as hand gestures, to further enhance the users' interaction experience.

4 EVALUATION

Theoretical technique assessment makes an important part in the evaluation of our project. One typical approach is to compare with ground truth values. For instance, in our ongoing work on acoustic head tracking, we collect auditory input at a series of head orientations (ground truth). Then we compare the head angles calculated using our algorithm with the ground truth values, to evaluate the performance of our approach. Another assessment method is to compare with state-of-the-art techniques. For instance, as a highly precise motion tracking system, Vicon is utilized to track the user's movement together with our own tracking method at the same time. This way, we can compare both tracking trajectories qualitatively and quantitatively. Furthermore, there exist some typical evaluation metrics. For instance, interaural cross correlation (IACC) is normally used to measure the difference in signals received by two ears. The IACC value will be nearly 1 for monoaural sources directly in front of or behind the listener, while becoming lower if the source is off to one side. Therefore, we can utilize this metric to inspect whether a simulated sound propagation corresponds to its real-life scenario.

Another important evaluation approach is the user study. In user studies, we measure users' performance in real tasks (*usefulness* evaluation) and investigate their general experience through interviews and questionnaires (*usability* evaluation). To make our studies more solid, we typically invite more than 20 or 25 people in a wide age range and of different backgrounds. In questionnaires, we usually design 5-point or 7-point Likert scale questions that cover user experiences about the application's or the system's attractiveness, perspicuity, efficiency, dependability, stimulation, novelty[2], and other system-related metrics. For a thorough analysis, we usually conduct data significance tests such as ANOVA test, Friedman test, and so on.

5 EXPECTED CONTRIBUTION

By the end of this PhD project, we hope to make the following contributions.

1. We will contribute to the AR research field by proposing wearable AAR system that can synthesize spatial sounds from arbitrary surrounding objects to the user. With this we expect to extend human-object interactions with the omni-directional auditory channel.
2. We will develop a novel head tracking method using off-the-shelf headsets, which alleviates the privacy concerns and improves the wearability of such an AAR system.
3. We will spatialize virtual soundscape with real acoustic effects and mix it with existing environment sounds, which creates a more immersive audio experience for the users.
4. We will evaluate the AAR system in real scenarios and verify its usefulness and usability.
5. Our implementation of an AAR system will advance and integrate current research in the acoustics domain with technologies and approaches in the human-object interaction and even pervasive computing fields, thereby helping to bridge the gap between these currently disjoint research communities.

[2] https://www.ueq-online.org/

6 SHORT BIOGRAPHY

Jing Yang is a PhD student in Distributed Systems Group at Department of Computer Science at ETH Zurich. She is supervised by Prof. Friedemann Mattern and advised by Dr. Gábor Sörös. Between June and September 2019, Jing Yang worked as a visiting researcher at Empathic Computing Laboratory in New Zealand, under the supervision of Prof. Mark Billinghurst.

Jing Yang's research interest lies in the field of audio augmented reality (AAR) and the use of AAR in human-object interactions (HCI). In particular, she intends to develop wearable systems that attach synthesized spatial audio to everyday real objects to enhance efficient, intuitive, and immersive interactions between people and arbitrary objects in consumer and industrial contexts. In addition to investigating the AAR application for single users, Jing Yang is also interested to explore its usage for interactions among multiple users in scenarios like remote collaboration.

Jing Yang's research is relevant to several research communities, including ACM UbiComp, ACM CHI, ACM Augmented Human, ACM MobiSys, IEEE ISMAR, etc. Her poster titled "Spatial Audio for Human-Object Interactions in Small AR Workspaces", co-authored with Gábor Sörös, won the Best Poster Award in the conference ACM MobiSys 2018. More information and a list of publications are on her personal website.

REFERENCES

[1] Robert Albrecht, Riitta Väänänen, and Tapio Lokki. (2016). Guided by Music: Pedestrian and Cyclist Navigation with Route and Beacon Guidance. *Personal and Ubiquitous Computing* 20, 20:1, 20(1), 121–145.

[2] Youssef El Baba, Andreas Walther, and Emanuel A. P. Habets. 2018. 3D Room Geometry Inference Based on Room Impulse Response Stacks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26, 5 (2018), 857–872. https://doi.org/10.1109/TASLP.2017.2784298

[3] Amit Barde, Matt Ward, William S Helton, Mark Billinghurst, and Gun Lee. 2016. Attention Redirection Using Binaurally Spatialised Cues Delivered Over a Bone Conduction Headset. In *Human Factors and Ergonomics Society Annual Meeting (2016)*. SAGE Publications, 1534–1538.

[4] Jens Blauert. 1997. *Spatial Hearing: The Psychophysics of Human Sound Localization.* MIT press.

[5] Simon Blessenohl, Cecily Morrison, Antonio Criminisi, and Jamie Shotton. 2015. Improving Indoor Mobility of The Visually Impaired with Depth-based Spatial Sound. In *IEEE International Conference on Computer Vision Workshops (ICCV'15)*. IEEE, 26–34.

[6] Cesar Cadena, Luca Carlone, Henry Carrillo, Yasir Latif, Davide Scaramuzza, José Neira, Ian Reid, and John J Leonard. (2016). Past, Present, and Future of Simultaneous Localization and Mapping: Toward the Robust-Perception Age. *IEEE Transactions on Robotics* 32, 6, 32(6), 1309–1332. https://doi.org/10.1109/TRO.2016.2624754

[7] Alessio Del Bue and Marco Crocco. 2016. 3D room Reconstruction with Sound. https://vgm.iit.it/tutorials/3d-room-reconstruction-with-sound

[8] Arindam Dey, Mark Billinghurst, Robert W Lindeman, and J Swan. 2018. A Systematic Review of 10 Years of Augmented Reality Usability Studies: 2005 to 2014. *Frontiers in Robotics and AI* 5 (2018), 37.

[9] Florian Heller and Jan Borchers. 2014. AudioTorch: Using A Smartphone as Directional Microphone in Virtual Audio Spaces. In *The 16th ACM International Conference on Human-Computer Interaction with Mobile Devices & Services (MobileHCI'14)*. ACM.

[10] Florian Heller and Johannes Schöning. 2018. NavigaTone: Seamlessly Embedding Navigation Cues in Mobile Music Listening. In *The ACM SIGCHI Conference on Human Factors in Computing Systems (CHI'18)*. ACM.

[11] Kangsoo Kim, Mark Billinghurst, Gerd Bruder, Henry Been-Lirn Duh, and Gregory F Welch. 2018. Revisiting Trends in Augmented Reality Research: A Review of the 2nd Decade of ISMAR (2008–2017). *IEEE Transactions on Visualization and Computer Graphics* (2018).

[12] Kent Lyons, Maribeth Gandy, and Thad Starner. 2000. Guided by Voices: An Audio Augmented Reality System. In *Conference on Auditory Display (ICAD'00)*.

[13] Jörg Müller, Matthias Geier, Christina Dicke, and Sascha Spors. 2014. The BoomRoom: Mid-air Direct Interaction with Virtual Sound Sources. In *The ACM SIGCHI Conference on Human Factors in Computing Systems (CHI'14)*. ACM, 247–256.

[14] Mirco Rossi, Julia Seiter, Oliver Amft, Seraina Buchmeier, and Gerhard Tröster. 2013. RoomSense. In *Proceedings of the 4th Augmented Human International Conference*, Albrecht Schmidt (Ed.). ACM, New York, NY, 89–95. https://doi.org/10.1145/2459236.2459252

[15] Spencer Russell, Gershon Dublon, and Joseph A Paradiso. 2016. HearThere: Networked Sensory ProsThetics through Auditory Augmented Reality. In *The 7th Augmented Human International Conference (AH'16)*. ACM, 20.

[16] Carl Schissler, Christian Loftin, and Dinesh Manocha. (2018). Acoustic Classification and Optimization for Multi-modal Rendering of Real-World Scenes. *IEEE Transactions on Visualization and Computer Graphics* 24, 3, 24(3), 1246–1259.

[17] Carl Schissler and Dinesh Manocha. 2017. Interactive sound propagation and rendering for large multi-source scenes. *ACM Transactions on Graphics* 36, 1 (2017), 2.

[18] Eldon Schoop, James Smith, and Bjoern Hartmann. 2018. HindSight: Enhancing Spatial Awareness by Sonifying Detected Objects in Real-Time 360-Degree Video. In *ACM SIGCHI Conference on Human Factors in Computing Systems (CHI'18)*. ACM, 143.

[19] Jaka Sodnik, Saso Tomazic, Raphael Grasset, Andreas Duenser, and Mark Billinghurst. 2006. Spatial sound localization in an augmented reality environment. In *Proceedings of the 18th Australia conference on computer-human interaction: design: activities, artefacts and environments*. 111–118.

[20] Titus JJ Tang and Wai Ho Li. 2014. An Assistive Eyewear Prototype That Interactively Converts 3D Object Locations into Spatial Audio. In *The 2014 ACM International Symposium on Wearable Computers (ISWC'14)*. ACM, 119–126.

[21] Jing Yang, Yves Frank, and Gábor Sörös. 2019. Hearing Is Believing: Synthesizing Spatial Audio from Everyday Objects to Users. In *The 10th Augmented Human International Conference (AH'19)*. ACM.

[22] Jing Yang and Gábor Sörös. 2018. Augmenting Smart Object Interactions with Smart Audio. In *AH'18 Proceedings of the 9th Augmented Human International Conference*. ACM, ACM, Seoul, Republic of Korea, 28:1 – 28:3.