Hearing Is Believing: Synthesizing Spatial Audio from Everyday Objects to Users

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

Yves Frank
ETH Zurich, Switzerland
yfrank@ethz.ch

Gábor Sörös
ETH Zurich, Switzerland
gabor.soros@inf.ethz.ch

ABSTRACT

The ubiquity of wearable audio devices and the importance of the auditory sense imply great potential for audio augmented reality. In this work, we propose a concept and a prototype of synthesizing spatial sounds from arbitrary real objects to users in everyday interactions, whereby all sounds are rendered directly by the user's own ear pods instead of loudspeakers on the objects. The proposed system tracks the user and the objects in real time, creates a simplified model of the environment, and generates realistic 3D audio effects. We thoroughly evaluate the usability and the usefulness of such a system based on a user study with 21 participants. We also investigate how an acoustic environment model improves the sense of engagement of the rendered 3D sounds.

CCS CONCEPTS

• **Interaction paradigms** → Mixed/augmented reality; • **Interaction techniques** → Auditory feedback; • **HCI design and evaluation methods** → User studies.

KEYWORDS

Spatial audio, augmented reality, human-object interactions

ACM Reference Format:
Jing Yang, Yves Frank, and Gábor Sörös. 2019. Hearing Is Believing: Synthesizing Spatial Audio from Everyday Objects to Users. In *Augmented Human International Conference 2019 (AH2019), March 11–12, 2019, Reims, France.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3311823.3311872

1 INTRODUCTION

Let us consider the following scenario: *Monday, 7am, Bob is ready to leave for work after a quick breakfast. As usual, Bob puts on his headphones to enjoy his favorite piece of music on the way. Right before he opens the door, somebody is calling from behind "Hey, Bob, you should take me with you!" Bob turns his head following the sound. "Aha! I almost forgot you." Bob finds the "sound source" which, however, is nothing but a book to be returned that has no loudspeaker but communicates directly through Bob's headphones.*

Wearable audio headsets and earbuds have already blended in our everyday life thanks to their low obtrusiveness. While getting used to enjoying music and telecommunication anywhere on the go, we can further explore the potential of our auditory sense to enhance human communication with the surroundings. In addition to giving information, the auditory sense also provides us with omni-directional engagement through an immediate 360° sense of space, time, and presence, which, together with the ubiquity of audio devices, indicates a significant opportunity for audio augmented reality. As opposed to head-mounted displays, wearable auditory devices enjoy much higher social acceptance, which is the main driving motivation for our research.

In this work, we present the concept and a prototype of synthesizing spatial audio (i.e., sound signals that are perceived to have a pronounced direction and distance) from everyday objects to humans. Specifically, 3D sounds can be created based on the current relative pose between the object and the user, and the simulated audio signals can be played via normal earphones. By this, we intend to add a more immersive notification channel to initiate people's interactions with surrounding objects, especially with those that are not necessary to or cannot have a loudspeaker. To evaluate the concept, we built a system in a $6.6\ m \times 9.1\ m \times 3.4\ m$ instrumented space, utilizing 11 cameras, a helmet with markers, a laptop for all computations, and ordinary earphones.

To explore the usability and the usefulness of such a system, we conducted a user study involving 21 participants. Our experiments cover two typical scenarios. (1) *Interaction with an object at a distance:* In this case, the user orientates the notifying object by looking into the correct direction to get information, remotely manipulates it, etc. (2) *Interaction by reaching the object:* The user is expected to approach and find the notifying object, which can be especially practical if the user tries to find something at an unknown or hidden place. We evaluate participants' accuracy and speed of finding the notifying objects by measuring their localization azimuth/elevation, or the distance errors to the ground truth locations, and the time spent on each test. The angular errors achieved by participants are very small and the median error is even comparable to humans' focused field of view (around 5.2°). The average distance errors are smaller than 25 *cm* in both horizontal directions in a $5\ m \times 8\ m$ area. Also, the participants could walk around at a normal speed to perceive the synthesized sounds smoothly in real time. Regarding usefulness, the participants in general regarded the whole experience very interesting and could imagine using such a system in several specific applications, such as receiving notifications from smart objects, getting alert messages, car infotainment, on-site games, home entertainment, and others.

We further improve the spatial perception of the synthesized 3D sounds, by modeling the environment geometry and materials to simulate more realistic acoustic effects. To evaluate the efficacy of environment modeling, we compare ground truth sounds recorded

AH2019, March 11–12, 2019, Reims, France

ACM ISBN 978-1-4503-6547-5/19/03…$15.00
https://doi.org/10.1145/3311823.3311872

in our room with simulated sounds rendered with and without the environment model. Through a qualitative user study, we find that people generally regarded the sounds simulated with an environment model more immersive. Besides, we quantitatively evaluate the improvement by acoustic metrics.

Our work indicates that spatial audio can be utilized in everyday interactions with a satisfying accuracy, immersive experience, and good user acceptance. Our contributions are as follows:

- We built a prototype system to synthesize spatial sounds for everyday real objects without loudspeakers.
- We conducted a thorough evaluation of users' behavior and experience when using synthesized spatial audio in different notification scenarios.
- We explored the influence of acoustic modeling of the environment on improving people's sense of engagement when perceiving the spatial sounds simulated in the same space.

2 RELATED WORK

Previous research has shown the great potential of *spatial audio* in real-life applications, including museum guides [1, 17], gaming [7], spatial music mixing [8], embedding contextual information in music [5, 6], and navigation [2, 10, 13]. In such systems, the key to create authentic 3D sounds is tracking the pose between the object and the listener, which can be done for example by head mounted cameras, radio frequency modules, or external cameras equipped in the environment. The main goal of our work is to evaluate the concept of synthesizing spatial audio from arbitrary objects to the user, therefore we use a not wearable but highly accurate tracking system to estimate the pose of the object(s) and the listener(s).

In interactive activities, one should be clearly aware of the involved object(s). To this end, there exist a number of works that explore people's localization accuracy based on spatial audio in real environments. Sodnik et al. [14] did experiments on a tabletop space ($100\ cm \times 60\ cm \times 60\ cm$). Tang [16] conducted a study on a planar surface of size $40\ cm \times 40\ cm$. Müller et al. [8] evaluated participants' sound localization performance in their ring-shaped BoomRoom of a 3 m diameter and their sound sources were distributed in the room. Heller et al. [4] also did experiments in a ring-shaped space. Their area was larger ($diameter = 5\ m$) but their 24 sound sources were evenly distributed at the edge of the circle spaced by 15°.

In our work, we intend to investigate people's orientation and localization performance in different scenarios in an indoor space of a moderate size where participants can freely move around, and where we arbitrarily distribute the sound sources. Typical localization measures include azimuth and elevation angles [18], distance offsets [8, 10, 16], or a simple counting of correct sound source identifications. Our study involves all these three measures for different scenarios. Research by [15, 20] indicates that a training session with paired audio/visual feedback can significantly help to reduce the error in the audio-only tests. To focus on the effectiveness of the synthesized audio signals and to generally cover the situations where visual perception is not feasible or acceptable, we did not include any visual feedback or assistance in our experiments.

We also explore the efficacy of environment modeling to enhance the sense of reality of the synthesized 3D sounds. Schissler et al. [11, 12] demonstrated distinct auditory perceptions when the sounds are rendered with the same room geometry but different surface materials. To our best knowledge, among the works that generate spatial audio for real-life scenarios, no one has explicitly modeled the surrounding geometry and materials, probably due to the difficulties in doing so in real time. We would like to find out how much the existence of such an acoustic model influences the user experience, therefore we model the space geometry and the materials offline. We then synthesize spatial sounds with the model, and explore the theoretical improvement in the acoustic effects and the subjective sense of engagement perceived by the users.

3 IMPLEMENTATION

As discussed before, the critical component of 3D audio creation is to accurately calculate the user's (head) pose with respect to the object(s). To this end, we utilize the Vicon[1] motion capture system that in our case consists of 11 cameras suspended from the ceiling. At runtime, users can freely walk around in our space with a helmet on which retro-reflective fiducial markers are attached. The Vicon system tracks the markers and calculates the pose of the helmet with high precision and low latency and with 6 degrees of freedom at 100 Hz. We use the helmet pose to approximate the user's head pose. We build our system using a laptop running Ubuntu 16.04 OS. We leverage the Robot Operating System in C#[2] to communicate with the Vicon and stream the pose data via WiFi to the game engine Unity3D[3] where we have built up a digital copy of the scene and registered the object positions. We then update the user avatar's pose in the Unity3D scene based on the streamed pose data in real time. To synthesize spatial audio, we utilize the Google Resonance Audio SDK[4] which can be integrated in Unity3D. We use this SDK to simulate the sound propagation to the listener using the SDK provided average head-related transfer function (HRTF). We define the registered objects as omni-directional sound sources, and take the user avatar as the listener. The spatial audio is then rendered in the Unity3D scene based on the real-time pose and played to the user immediately via commercial wireless Apple earpods.

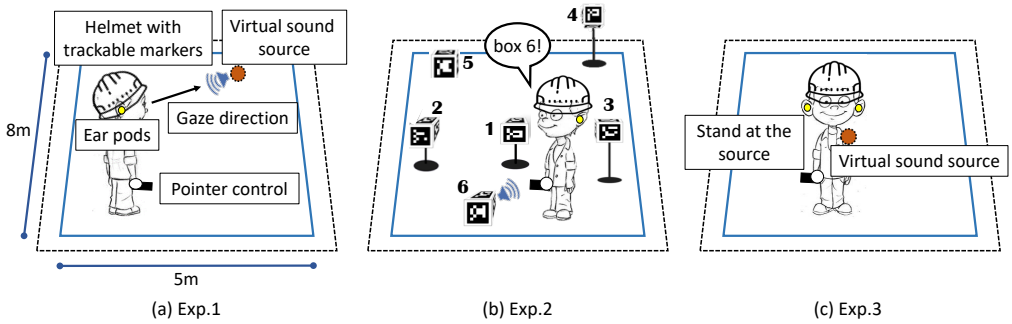
We implemented the whole system in a $6.6\ m \times 9.1\ m \times 3.4\ m$ space, which is surrounded on three sides by heavy curtains and on the last side by a wall with windows. Considering the working space covered by Vicon cameras, users can move freely wearing the helmet and the earpods in an area of approximately $5\ m \times 8\ m$. The working environment and the system components are illustrated in the accompanying video[5].

4 USER STUDY

We conducted three lab experiments to evaluate the concept of synthesizing spatial audio for everyday objects.

We recruited 21 participants (average age 25.8 with a standard deviation (SD) of 2.51, ranging from 21 to 34, five female) for the user study. Seven of them never heard about spatial audio, and 11 participants heard about it but had never tried anything with spatial audio before. Three people had experienced spatial audio in the past when trying virtual reality demos but their main focus was on

[1] https://www.vicon.com/
[2] https://github.com/siemens/ros-sharp
[3] https://unity3d.com/
[4] https://developers.google.com/resonance-audio/
[5] https://youtu.be/_pGjwViGSQI



(a) Exp.1 (b) Exp.2 (c) Exp.3

Figure 1: The illustrations of the three experiments. The blue solid rectangle frames the working area covered by Vicon cameras, while the black dash rectangle depicts the whole space. In the experiments, we marked the working area on the ground for the participants.

the visual components. All participants had no hearing problems and had normal binaural hearing as tested in our pre-experiment.

The participants first read instructions and signed the consent form, then they did three experiments and finally answered a questionnaire. Prior to the experiments, the participants were asked to choose their favorite clip out of three short continuous pieces of music. In each experiment, we started with two trials to help them get used to the sound perception and the system control using a laser pointer, after which they performed eight formal tests. In total, for each experiment, we have 168 test cases from all participants.

In the following, we first elaborate on three experiments and analyze the results in Section 4.1-4.3. then we evaluate the questionnaire in Section 4.4.

4.1 Experiment 1

Scenario. By this experiment, we intended to simulate notifications from objects at a distance. A user does not need to approach the object, but looks into the object's direction. In this experiment, all sound sources were virtual (invisible), because we intend to explore the users' localization performance exclusively based on spatial audio, without any other assistance (or visual clues).

Procedure. We illustrate this experiment in Figure 1(a). We played spatial sounds from different virtual locations in the space. Upon hearing a sound, participants were asked to look at the source direction as perceived. They started the series of tests at the center of the room, and they were allowed to move around to determine the source location. To confirm the location, they were asked to look in the determined direction while standing still and pressing the pointer button. Upon confirmation, we stopped the current sound and recorded their facing direction (based on the helmet pose) and the time spent on the test. Right afterwards we continued with the next test.

The relative sequence of source locations was the same for all participants, i.e., regardless of where the participant stood for test i, they had to turn by (ϕ_i, θ_i) to look in the correct direction for test $i+1$, with (ϕ_i, θ_i) being equal for all participants. Besides, the source of test $i+1$ was always generated three meters from where the participants stood to confirm the answer for test i.

To measure their localization accuracy, we computed the azimuth (horizontal angle ϕ) and the elevation (vertical angle θ) of their facing direction with respect to the virtual sound source direction.

Results. Ideally, both the azimuth and the elevation errors should be 0°. Figure 2 demonstrates the distribution of the absolute azimuth and elevation errors from all participants.

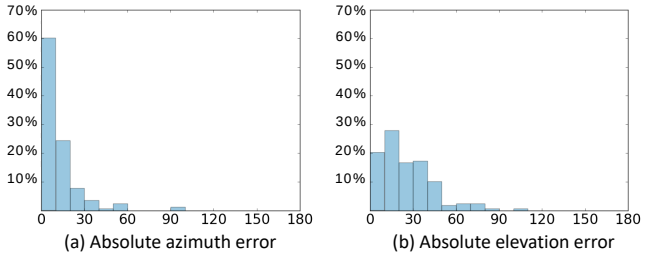


Figure 2: The distribution of the azimuth and the elevation errors in Experiment 1. We show the absolute values in the range (0°, 180°).

Around 90% of the azimuth errors are within 30° while the elevation distribution is flatter, which is reasonable since simulating the vertical differences is more difficult. Note that in the horizontal plane, we can determine (and simulate) the sound direction based on the difference of sound intensities and arrival times in our two ears, which is roughly the same for every person. However, we determine the vertical sound direction based on nuance differences across different paths shaped by our pinna, which might be different for every individual (and needs to be considered in simulations). In our implementation, we utilize an average HRTF, but a personalized HRTF may help to improve the vertical simulation [3] and hence the localization accuracy.

The average azimuth error is 12.07° ($SD = 14.59°$) and the median is only 6.76°, which is even comparable to human beings' focused vision of around 5.2° [19]. The elevation error is larger ($mean = 25.06°, SD = 18.67°, median = 22.81°$) but is still acceptable considering the normal vertical field of view of approximately 135° [9]. We find significant differences in localization accuracy between participants (one-way ANOVA, $p < 0.001$). The smallest mean azimuth error is 3.73° and the largest is 39.54°. The smallest mean elevation error is 13.76° and the largest is 58.82°.

Participants spent on average 13.45s on each test ($SD = 10.66s$). One-way ANOVA indicates significant individual differences ($p < 0.001$) and the fastest participant took only 3.89s on average.

By plotting each participant's viewing angle traces, we found that in a majority of the tests (especially for azimuth), upon hearing the sound, participants could follow it and turn to face the roughly

correct direction very fast, but then they spent much time adjusting their viewing angle to refine the direction as much as possible.

4.2 Experiment 2

Scenario. Similar to the scenario of Experiment 1, here we also deal with objects at a distance, however, we played the synthesized sounds from the locations of real objects. The real "proxy" objects were six paper boxes of size 15 $cm \times 10\ cm \times 10\ cm$. This is more similar to real situations where the audio augmented objects are visible. We anticipated that the participants would be able to localize the sound sources faster than they did in Experiment 1 and they could achieve a very high accuracy.

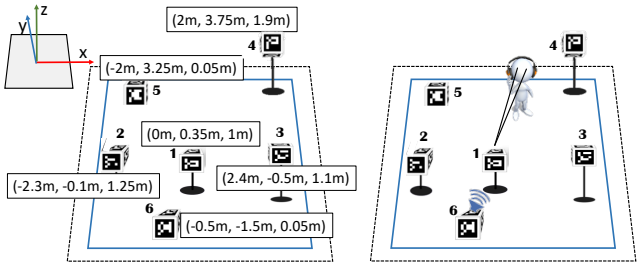


Figure 3: The locations of the boxes in Experiment 2 are depicted in (a). On the top left we illustrate the coordinate directions and the origin is at the center of the room. On the right we illustrate the typical error case. Box 6 is the actual sound source, but when the participants only focused on Box 1 that happened to be collinear with Box 6, they reported the wrong source.

Procedure. Figure 1(b) demonstrates the experiment scenario. We placed the six numbered boxes at fixed locations and their coordinates are illustrated in Figure 3(a). The participants first looked around to see the boxes. In each test, they were asked to determine from which box the sounds were originating. Like in Experiment 1, they started the series of tests at the center of the room, and they were allowed to walk around to decide on the source box. To confirm the answer, they pressed the pointer button and told the investigator the box number, then we continued with the next test. The sequence of the source boxes was the same for every participant. We measured the time spent on each test and counted the correct identifications of the source boxes.

Results. Out of a total of 168 tests there were only four mistakes from four individuals, which were of the same type as depicted in Figure 3(b). In these four cases, when the participants moved around and finally faced the collinear boxes at a distance, they quickly confirmed the answer but did not realize that the farther box was the correct sound source. In contrast, we observed that when people realized the other collinear box, they all walked back and forth and gave the correct answer.

On average, participants spent 8.51s on each test ($SD = 5.61s$). Significant differences are found between participants (one-way ANOVA, $p < 0.001$) of which the smallest average is 4.5s and the largest is 15.68s. As anticipated, participants localized the sound sources much faster than in Experiment 1. We also observed that people were generally more confident about their answers, even if they were several meters away from the source box. This confirms the hypothesis that the visual cues largely improve our perception. We suppose that in a familiar environment they know very well, the localization can be even faster.

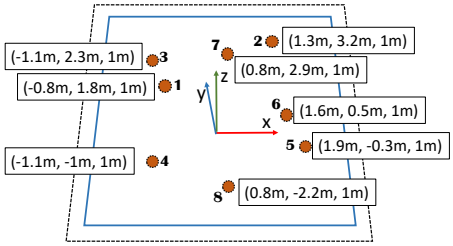


Figure 4: The positions of the virtual sound sources in Experiment 3. As before, the origin is at the center of the room.

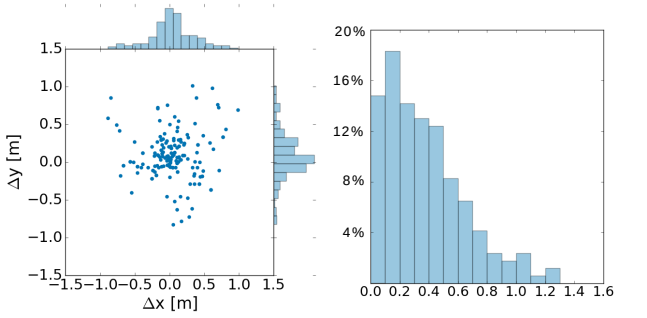


Figure 5: (a) shows the distribution of the distance errors in two horizontal directions in Experiment 3. Most of the distances are within $0.25m$. (b) shows the histogram of the Euclidean distances between the participant and the sound source.

4.3 Experiment 3

Scenario. This experiment was designed to simulate the cases where the user has to really reach the object, instead of simply interacting with it at a distance. We assumed that such a scenario would be very common when trying to find an object at an unknown or hidden place.

Procedure. Figure 1(c) demonstrates this experiment scenario. We played sound signals from arbitrary virtual locations that were all one meter above the ground. Upon hearing the sound, the participants were asked to find out the source and stand at the exact location as they perceived. As before, they started the series of tests at the center of the room, and the test $i + 1$ was played right after they confirmed the answer at the location for test i. When they confirmed their localization by pressing the pointer button, we recorded the spent time and their standing position.

The locations of the eight test sources are depicted in Figure 4. We distributed the sources in a way to cover the space, with some margins to allow localization errors near the boundary of the working area. Since the sources were all of the same height ($z = 1\ m$), we only computed the distance errors in two horizontal directions x (the 5 m edge) and y (the 8 m edge).

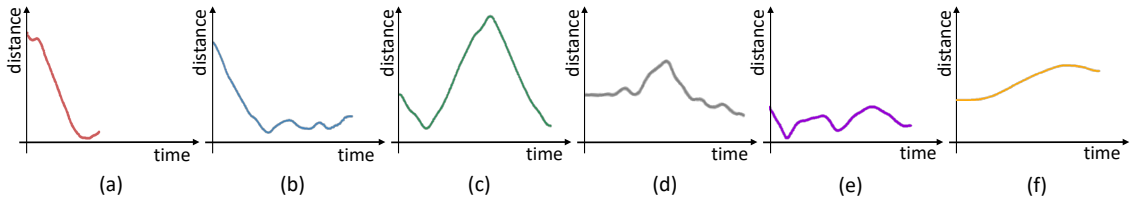


Figure 6: We highlight six most representative movement trajectories in Experiment 3. The horizontal axis represents time and the vertical axis represents the Euclidean distance from the participant to the sound source.

Results. Figure 5(a) shows the error distribution in both directions. Based on the absolute values, the mean Δx is $0.24m$ with a SD of $0.23m$ and a median of $0.16m$ and the mean Δy is $0.23m$ with a SD of $0.22m$ and a median of $0.16m$.

Considering that the participants' movement area is $5\ m \times 8\ m$ (see Section 3), the average error is 4.84% in x direction and 2.88% in y direction. Significant differences are found between participants (one-way ANOVA, $p < 0.002$). The smallest mean Δx is $0.1\ m$ and the largest is $0.62\ m$, while the smallest mean Δy is $0.09\ m$ and the largest is $0.51\ m$. Furthermore, in Figure 5(b) we show the distribution of the Euclidean distances $\sqrt{(\Delta x)^2 + (\Delta y)^2}$ between the participants and the sound sources. The mean Euclidean distance is $0.37m$ with a SD of $0.27m$ and a median of $0.32m$. Again, we find significant individual differences (one-way ANOVA, $p < 0.001$) of which the smallest average Euclidean distance is $0.18m$ and the largest is $0.83m$. The time performance is similar to that in Experiment 1. We get an average completion time of $12.44s$ ($SD = 11.56s, median = 9.52s$).

Based on the pose data recorded and streamed from Vicon, we examined all the participants' movement patterns in each test, and in Figure 6 we highlight six most representative trajectories (Euclidean distance vs. time).

Pattern (a) refers to the fastest movement that the participants followed the sound and quickly confirmed their localization once they believed that they arrived. It can be seen that the distance does not change significantly at the beginning, and many participants share this pattern. In this short period of time, they first turned around to determine the source orientation before walking towards it. The first half of pattern (b) is very similar to pattern (a), but then the participants spent much time walking around the sound source to carefully check their answer. In pattern (c), the participants also first quickly approached the sound source, but then they walked further and finally came back. Pattern (d) is similar to pattern (c) but the participants did not get closer at first. In some cases, the participants kept wandering around the sound source (pattern (e)), and this happened more when the participants were already close to the source at the beginning of the test (e.g., from source 5 to source 6). Finally, there are very few cases of pattern (f) that the participants moved and stayed too far in the end. The patterns (a,b,c) took the majority among all the traces.

Summary. From all the experiments, we can see that the synthesized spatial sounds can generally guide users to the objects accurately. Especially from Experiment 1 and 3, we realize that in most tests, the participants' first intuition to follow the sound was always in the direction to reduce the angle or distance error.

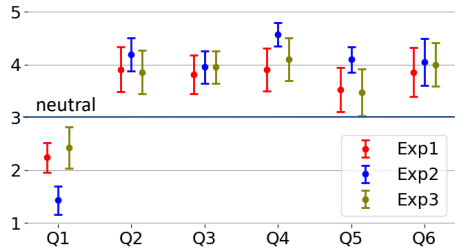


Figure 7: The average Likert scores of each question for all three experiments. Error bars show the 95% confidence intervals.

Experiment 1 and 3 were based on virtual sound sources, which significantly increased the difficulty. Besides, participants tried to make little errors so they carefully checked their localization in many tests. As implied by the movement patterns, we anticipate that in reality where the objects stand out in the environment, users can localize them much faster. And this can be further improved if the user is familiar with the surroundings.

4.4 Questionnaire

After the experiments, participants filled in a questionnaire with two parts. Questions were answered on a 5-point Likert scale from "strongly disagree" (1) to "strongly agree" (5) with the neutral at 3. After finishing the questionnaire, we also discussed potential applications and additional insights with the participants.

Part 1: Experiment Questions. The first part includes six questions for every experiment: (Q1) This experiment was difficult for me. (Q2) The audio clips sounded smooth when I moved in the environment at a normal speed. (Q3) I felt the guidance by spatial audio natural. (Q4) I felt that the spatial sound guided me to locate the sound source correctly. (Q5) I felt that the spatial sound guided me to locate the sound source fast. (Q6) Overall, I could experience immersion (sense of space and presence) by using the system. Figure 7 shows the mean scores with 95% confidence intervals of each question for all three experiments.

Participants in general disagreed that the experiments were difficult for them (Q1), especially Experiment 2. This is as expected since Experiment 2 was the only one using visible objects which stood out in the environment.

Q2 investigates whether the system ran smoothly to synthesize the 3D audio as participants moved around at their normal speed, and the results indicate positive feedback. During the experiments, we observed that people sometimes performed very fast movements such as quickly turning their head, and our tracking system kept stable against these actions and steadily operated at 100Hz.

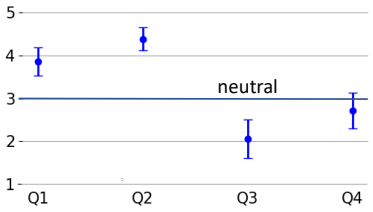


Figure 8: The average Likert scores of four general questions. Error bars show the 95% confidence intervals.

There were, however, several moments when a few participants felt the sound stumbling. This was mainly because the participants moved around the borders of the working area which were not fully covered by the cameras, therefore the pose update was missing. This happened less frequent in Experiment 2, in which people could see the boxes clearly and did not approach the borders.

Regarding Q3, participants generally agreed that they felt it natural to follow the synthesized spatial audio in every experiment scenario. Based on their verbal feedback, in most cases they could perceive the sound signals with clear orientation to follow. Besides, the 3D sound was updated as they expected while walking around. Both of these effects contributed to the positive feedback to this question, as well as Q6 in which they experienced the sense of engagement in the space.

Q4 and Q5 were participants' self-evaluation of their performance. Corresponding to Q1, people in general thought that they performed the best in the least difficult Experiment 2, while they were not that confident with Experiment 1 and 3. Participants were satisfied with their correctness especially in Experiment 2, which matches with the actual results as we analyzed in previous sections. Compared with the correctness, people were less content with their speed. As reported by the participants and also analyzed before, sometimes they were not quite sure about their determination, therefore they took considerable amount of time checking the potential source locations. Except for the object visibility and the user's familiarity with the environment, we assume that it would help if the synthesized spatial audio sounds more realistic with more matching acoustic effects.

By one-way ANOVA tests, we find that there are not significant differences between experiments in Q2 ($p = 0.39$), Q3 ($p = 0.76$), and Q6 ($p = 0.81$), which further indicates that in all three experiments, participants perceived synthesized sounds smoothly when moving around and they generally regarded the whole experience natural and immersive. However, there are significant differences in Q1 ($p = 4.8 \times 10^{-5}$), Q4 ($p = 0.023$), and Q5 ($p = 0.041$), which implies that participants' assessments on the difficulty, the correctness, and the speed vary among different experiments.

Part 2: General Questions. The second part includes four questions regarding their overall experience: (Q1) I got used to the spatial experience quickly. (Q2) The spatial sound experience was interesting for me. (Q3) The sense of audio orientation was not pleasant for me. (Q4) The audio signals sounded too artificial to be true. Figure 8 demonstrates the participants' feedback regarding these general questions.

Our experiments were new to every participant, including the three people who experienced spatial audio before because they only heard 3D signals in virtual reality demos while focusing on the visual rendering. Participants generally agreed that they got used to the spatial experience quickly (Q1), and they tended to strongly agree that the spatial audio experience was interesting (Q2).

We intended to explore the rendering quality of our synthesized sounds by Q3 and Q4. As also indicated in the previous questions, participants could generally experience clear orientations to follow and such a sense brought by the synthesized 3D signals were comfortable for them (Q3). However, even though they experienced the sense of engagement to a certain degree, some of them still thought that the synthesized sounds were more artificial than naturally occurring (Q4). A few participants commented on the sounds to be more "internal" than "external", which indicates that the audio signals were not adequately spatialized. An issue of our system was rendering the spatial sound using an average HRTF instead of a personalized one. This worked properly for most participants, but one of them particularly reported that his 3D experience was poor compared to his expectations. We anticipate that a personalized HRTF would further improve the results.

Potential Applications. In general, the questionnaire answers indicate satisfying user experience which implies that spatial audio based interaction has great potential in everyday applications.

During our discussions with the participants, we proposed several application scenarios and asked for their opinions. They could easily imagine using such a system to receive messages (reminders, notifications, etc.) from normal objects, especially those which do not and are not necessary to be equipped with a real loudspeaker, such as coffee machines, lamps, and even bags. In particular, they believed that using the system to receive alert messages can be really helpful in some emergency situations. A very interesting application which was strongly supported by most participants was on-site games such as escape rooms. They believed that it would be quite interesting if they have to localize hints guided only by spatial sounds.

The participants also proposed their own application ideas. A few of them mentioned that such a system can be utilized in a home cinema, and some would like to use it for navigation. Several participants anticipated that the system can be quite useful for visually impaired people to interact with everyday objects. Such a wide range of application ideas indicates the great potential of the usefulness of such a spatial audio system.

5 ACOUSTIC ENVIRONMENT MODELING

In previous sections, participants in general found the synthesized spatial audio immersive, but some felt the 3D signals somewhat artificial. Therefore, we further explored how to improve the sound rendering with an adequate acoustic environment model.

Our model emulates geometric shapes and surface materials in order to produce authentic acoustic effects. Like before, we use Unity3D and the Resonance Audio SDK which supports modeling of arbitrary geometries and flexible assignment of surface materials. Moreover, it defines parameters such as reverb gain and reverb brightness that are adjustable to fine tune the room acoustics.

We first created a rough environment model based on the dimensions and the materials obtained in reality, and refined this model with real impulse response measurements. The Reverberation Time

Table 1: RT60 comparison in octave band between the real room and the model room. The model is quite close to the reality on these eight frequency bands.

Octave Band (Hz)	63	125	250	500	1K	2K	4K	8K
real RT60 (s)	1.375	1.185	0.925	0.862	0.655	0.61	0.608	0.411
model RT60 (s)	1.4	1.26	1.12	0.82	0.6	0.61	0.68	0.38

60dB (RT60) describes how long a sound takes to decay by 60dB in a space of diffuse sound field, which is an important room acoustic metric. To reproduce the room acoustics as realistic as possible, we measured the RT60 in the real space. We placed a sound recorder in the middle of the room and recorded four balloon blast sounds at four different locations. With help of the acoustics analysis software REW[6], we calculated the RT60s based on each blast impulse, then we averaged the results to approximate the room RT60. Based on these numbers, we adjusted the model to have the same RT60 in order to better approximate the real environment.
The final RT60 of the modeled room is compared to that in reality in Table 1. Next, we compare the ground truth sounds with the synthesized 3D sounds which are rendered with and without the environment model.

Ground Truth (GT) Sound. To record the ground truth sounds in reality, we played a continuous piece of music using a round loudspeaker Jabra Speak 410 (surface diameter=10cm) at four arbitrary locations L1-L4. For each source location, the investigator listened to the sound while (1) standing at the center of the room (*static*) and (2) freely walking around starting from the center (*dynamic*). We utilized the Roland CS-10EM binaural in-ear microphones[7] to record the investigator's sound perception at both ears. We also recorded the investigator's movements using the Vicon system.

System Simulated (SS) Sound. In the Unity3D simulation, we defined the same sound sources, and replayed the listener's movement from recorded trajectories. In theory, by this we reproduced the same listening activities and recorded the synthesized 3D audio with and without the environment model.

5.1 Theoretical Comparison

We first compare the GT sounds with the SS sounds using two acoustic metrics: interaural cross correlation (IACC) and Mel-frequency cepstrum coefficients (MFCCs). IACC measures the difference in signals received by two ears. The values range from -1 (identical but out of phase) to 1 (identical and in phase). The IACC 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. MFCCs are commonly used as features in audio similarity measures.
We extracted 7 seconds from each of the 24 clips. These 7s sounds were then split into left and right channels at a sampling rate of 44.1KHz. We then segmented these signal arrays into seven 1s frames. In each frame, we (1) calculated the IACC, and (2) extracted the MFCCs features (first 12 dimensions), applied max-min normalization, and calculated the cosine similarities between GT and SS sounds in left and right channel respectively. Finally, we averaged the IACCs and the MFCCs-similarities across the seven frames.

[6]https://www.roomeqwizard.com/
[7]https://www.roland.com/us/products/cs-10em/

Table 2: IACCs of the ground truth sounds (GT), the simulation sounds with the environment model (SS^*), and the simulation sounds without the environment model (SS^0). "s" refers to *static* and "d" refers to *dynamic*.

	L1-s	L1-d	L2-s	L2-d	L3-s	L3-d	L4-s	L4-d
GT	0.605	0.462	0.293	0.460	0.562	0.546	0.814	0.557
SS^*	0.629	0.486	0.364	0.402	0.608	0.519	0.838	0.626
SS^0	0.998	0.865	0.796	0.869	0.799	0.858	0.979	0.892

Table 3: Cosine similarities of MFCCs features between the ground truth sounds (GT) and the simulation sounds with/without environment model (SS^*/SS^0). "s" refers to *static* and "d" refers to *dynamic*.

	Left Channel		Right Channel	
	GT vs. SS^*	GT vs. SS^0	GT vs. SS^*	GT vs. SS^0
L1-s	0.892	0.853	0.891	0.856
L1-d	0.891	0.861	0.893	0.868
L2-s	0.904	0.861	0.907	0.891
L2-d	0.893	0.847	0.890	0.862
L3-s	0.899	0.889	0.899	0.863
L3-d	0.866	0.825	0.866	0.824
L4-s	0.903	0.880	0.902	0.878
L4-d	0.899	0.886	0.903	0.878

Table 2 lists the IACC measures of the GT and the SS sounds with/without the environment model (SS^*/SS^0). Compared with SS^0, the IACCs of SS^* are significantly closer to the IACCs of the real sounds. L1 and L4 are almost directly in front of and behind the listener when standing at the room center. Therefore, the IACC measures of SS^0 at L1-s and L4-s are nearly 1. However, the sound perceptions at left and right ears are influenced by the real room acoustics, which is captured by the simulation with the environment model (SS^*).
Table 3 shows the cosine similarities between the GT and two types of SS sounds. It indicates that for both left and right channels, simulations with the environment model are more similar to the ground truth than simulations without the model and their differences are around $0.02 - 0.05$. We argue this is somewhat significant considering that the more similar SS^* sounds are approximately only 0.1 from perfect matching with the ground truth.

5.2 User Study

Theoretical measures have demonstrated the improvement by including the environment model. We also conducted a user study to investigate individual perceptions when listening to different simulations and the ground truth. Different from Section 4, here an on-site experiment was not feasible because in many cases the perception differences were subtle and a user might want to relisten to previous sounds, which is hard to control in reality. Instead, we conducted an online survey, in which participants listened to the aforementioned eight groups of sounds. Each group includes the GT and two SS sounds but the participants were not informed which simulation was with the model. The participants were asked to select the simulations which sounded more immersive to them

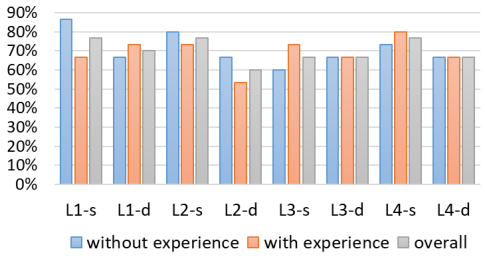


Figure 9: The percentages of participants who perceived the simulations with the acoustic environment model more immersive. The results are shown in three categories: with experience (15 participants), without experience (15 participants), and overall.

and explained how they interpreted "immersion" during the tests. As a reference, the pictures of the real environment and the GT sounds were provided.

We recruited 30 participants (average age 26.8 with a SD of 3.58, ranging from 19 to 34, nine female). 15 people had no experience in spatial audio and 15 people had experience (14 tried spatial audio before and one person had research experience). Ten of them also joined the previous user study. Figure 9 demonstrates that in almost all the tests, a majority of participants perceived the simulations with the environment model more immersive, and this holds true for both experienced and inexperienced people.

According to the participants' feedback, they mainly determined the degree of immersion based on the reverberation or the spatial location they could feel compared with the GT sounds. Eight participants, who always selected the simulations with the model, reported that they experienced the spatial locations and the 3D feeling as they would expect (*more externally originated*), while simulations without the model were more artificial (*more internally originated*). However, since our simulation was only an ideal approximation of the reality which did not involve all the details, therefore the reverberation was stronger, the background noise was louder, and the sense of externalization was not thorough. These issues also influenced the immersion perception for some participants.

Overall, the results have demonstrated the efficacy of simulating the room acoustics. We anticipate by providing more authentic environment model, people can experience further improved sense of engagement, and this advancement cannot be compensated by only applying a personalized HRTF for sound rendering (considering our real-life audio experience in a hall vs. in an open space).

6 CONCLUSIONS & FUTURE WORK

In this paper, we have proposed the concept of utilizing synthesized spatial audio as a new communication channel from everyday objects to humans, even without built-in loudspeakers on the objects. In three experiments, we have demonstrated that people can locate the objects with high accuracy even while walking. By the study about environment modeling, we have shown that people can actually experience different levels of immersion depending on the simulations of the room acoustics. Our findings indicate great potential of this research direction and we are interested to continue with the following future works.

First, we intend to implement such a system using completely wearable components in order to decrease our dependency on the working environment. Second, we are interested to explore real-time methods that create an acoustic model of the working environment in order to generate more authentic audio effects. We anticipate that such a system will not only benefit visually impaired people, but also generally enhance people's everyday interactions with objects in smart environments using *only* personal wearable devices.

ACKNOWLEDGMENTS

We thank all study participants for their time, effort, and feedback.

REFERENCES

[1] Benjamin B. Bederson. 1995. Audio Augmented Reality: A Prototype Automated Tour Guide. In *Conference Companion on Human Factors in Computing Systems (CHI'95)*.

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

[3] Michele Geronazzo, Erik Sikström, Jari Kleimola, Federico Avanzini, and Stefania Serafin. 2018. The Impact of An Accurate Vertical Localization with HRTFs on Short Explorations of Immersive Virtual Reality Scenarios. In *International Symposium on Mixed and Augmented Reality (ISMAR '18)*.

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

[5] Florian Heller, Jayan Jevanesan, Pascal Dietrich, and Jan Borchers. 2016. Where Are We?: Evaluating The Current Rendering Fidelity of Mobile Audio Augmented Reality Systems. In *Proceedings of The 18th International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '16)*.

[6] Florian Heller and Johannes Schöning. 2018. NavigaTone: Seamlessly Embedding Navigation Cues in Mobile Music Listening. In *Proceedings of The 2018 CHI Conference on Human Factors in Computing Systems (CHI'18)*. 637.

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

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

[9] Austin Roorda and David R Williams. 1999. The Arrangement of The Three Cone Classes in The Living Human Eye. *Nature* 397, 6719 (1999), 520.

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

[11] 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 (2018), 1246–1259.

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

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

[14] 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 (OZCHI'06)*. 111–118.

[15] Venkataraman Sundareswaran, Kenneth Wang, Steven Chen, Reinhold Behringer, Joshua McGee, Clement Tam, and Pavel Zahorik. 2003. 3D audio augmented reality: implementation and experiments. In *Proceedings of The 2nd IEEE/ACM International Symposium on Mixed and Augmented Reality (ISMAR'03)*. IEEE Computer Society, 296.

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

[17] Yolanda Vazquez-Alvarez, MatThew P Aylett, Stephen A Brewster, Rocio von Jungenfeld, and Antti Virolainen. 2014. Multilevel auditory displays for mobile eyes-free location-based interaction. In *CHI'14 Extended Abstracts on Human Factors in Computing Systems (CHI EA'14)*. ACM, 1567–1572.

[18] Yolanda Vazquez-Alvarez and Stephen Brewster. 2009. Investigating Background & Foreground Interactions Using Spatial Audio Cues. In *CHI'09 Extended Abstracts on Human Factors in Computing Systems (CHI EA'09)*. ACM.

[19] Brian A Wandell. 1995. *Foundations of Vision*. Vol. 8. Sinauer Associates Sunderland.

[20] P Zahorik, C Tam, K Wang, P Bangayan, and V Sundareswaran. 2001. Localization Accuracy in 3D Sound Displays: The Role of Visual-feedback Training. In *Proceedings of The Advanced Displays and Interactive Displays Federal Laboratory Consortium*.