ARVI: Augmented Reality for the Visually Impaired

James Lin
UC Berkeley
james97lin@berkeley.edu

Mengshi Feng
UC Berkeley
fms9424@berkeley.edu

ABSTRACT

Virtual reality (VR) and augmented reality (AR) technologies have seen a surge in popularity over the past couple of years, resulting in a deeper exploration of its potential within various fields. One population, however, that has seen little influence from immersive technologies is the visually impaired community. In a way, this is natural given VR/AR's ocularcentric nature, but in truth, many of the same technologies used for VR/AR can be turned into powerful tools for the blind as well. We introduce ARVI, a crowd-sourced scanning and localization iOS app that enables visually impaired users to learn semantic information from their surroundings. ARVI utilizes persistent map scanning and GPS to localize within a pre-scanned environment, and spatial audio through HRTFs to provide localized information. We provide a visually-impaired friendly application that lets users both access scans from the rest of the community and create new scans of their own.

KEYWORDS

VR/AR, SLAM, accessible technologies, spatial audio

1 INTRODUCTION

Ever since this wave of virtual reality (VR) and augmented reality (AR) began in 2012 with the Oculus kickstarter [7], research on the topic has been predominantly focused on vision: in creating better displays, smarter rendering techniques, and impressive visuals for VR, and in better localization, recognition, and mapping techniques for AR. VR/AR products and applications are similarly focused more on what the user can see than anything else.

As a result, there seems to be little connection between VR/AR and the visually impaired - a blind person is unlikely to be impressed by the latest virtual reality headset or the newest augmented reality app. However, some of the underlying technologies that were developed or improved to enable higher quality VR/AR applications can be retooled to support the visually impaired community as well. For example, mobile SDKs like ARKit [15] and ARCore [2] allow smartphones (a ubiquitous device in today's society) to reason about spatial data in ways that only humans were able to do in the past. When combined with semantic information, they have the potential to substitute for or augment the eyes of a visually impaired user.

To better discover a proper marriage between immersive technologies and the needs of the visually impaired community, we reach out to various people and organizations in the Bay Area, including the Lighthouse for the Blind and Visually Impaired [5] and the Smith Kettle-Well Eye Research Institute [10]. We conduct a series of interviews with the goal of learning two things:

(1) What previous work had been done in this space, and their usefulness/impact on the visually impaired community.
(2) What problems the visually impaired community faced that could potentially be solved or alleviated through the use of immersive technologies.

With what we learn from those interviews, we identify that, rather than tasks like obstacle avoidance or long-distance navigation, a fundamental problem for the visually impaired is that there's no easy way to learn semantic information from silent objects in the environment. Blind person must instead rely on touch or context clues to learn about their surroundings, neither of which are reliable methods in all situations. Given this general issue, we narrow down the problem statement to what we call the **5 meter problem**.

The 5 meter problem refers to the issue that GPS is only so precise. While it's satisfactory in getting a user to a rough location, in crowded urban environments the user can easily end up 5 or more meters away from their intended destination [3]. For a sighted person this is fine, as they can navigate themselves the rest of the way. But for the blind, 5 meters is the difference between catching and missing the bus, or between walking into the correct entrance and finding only a wall.

We introduce **Augmented Reality for the Visually Impaired**, or **ARVI**, as a tool to help combat this problem. ARVI is a multi-user iOS application that allows visually impaired users to create or pull down scans of the environment that are augmented with localized spatial audio cues. A user of our app could, for example, walk into a store and hear an audio cue precisely where the cash register is, allowing them to quickly get themselves situated.

To support this functionality, we decide not to rely on semantic detection algorithms as they're not sophisticated or consistent enough to be used safely. Instead, we take a crowd-sourced approach by building in the ability for anyone to construct and upload audio augmented scans of the environment. Users can place audio cues of their choosing into the physical environment, and the resulting maps are pushed to the cloud so that everyone has access to them. Through this process, work done by someone can benefit the rest of the community, and people can choose for themselves what they deem important enough to scan. The entire application is designed to be easy and intuitive to use by a visually impaired person.

In summary, our contributions are as follows:

- Through research and interviews, we isolate a real problem the visually impaired community faces that immersive technologies can play a role in fixing.
- We develop ARVI, an iOS app that enables users to hear and locate semantically important spatial audio cues in pre-scanned environments.
- We improve ARVI with a blind-friendly scanning and audio augmentation process to support a crowd-sourced approach to collecting data.

2 RELATED WORK

As an augmented reality solution for the visually impaired community, our product relates to many prior works. Many efforts have been made to help visually impaired people over the past couple decades. Researchers in both academia and industry have proposed solutions to help improve the lives of visually impaired people using advanced technologies.

Aira is an intuitive system that allows blind people to "see" the world with the assistance of a real-time aira agent [1]. The smart glasses or phone camera input the real-time video and the agent can talk users through whatever situation they're in. This approach is helpful for the visually impaired but lacks scalability as the human agents are expensive and lack in supply. OrCam, an Israeli-based company, aims to improve the everyday lives of visually impaired people by their wearable artificial intelligence devices [8]. Their product mounts a smart camera on the frames of user's eyeglasses, which reads texts, street signs and recognizes common items based on AI technologies, then dictates words into the user's ear [18]. OrCam improves the ability of reading and recognizes object if users already know the precise location of the interested object. Our proposed solution, on the other hand, helps users find desired pre-scanned objects within a short distance.

A group of researchers at the Technical University of Lodz also made some efforts for visually impaired people emphasizing on obstacle avoidance and spatial orientation using spatial audio cues. They proposed an electronic travel aid(ETA) utilizes segmented 3D scene images, personalized spatial audio [12] and then test it in both virtual and real environment. They utilized the pitch and amplitude of the assigned sound to differentiate the direction and distance information of the obstacle. Our proposed solution adopts similar idea using spatial audio cues not to avoid but to locate the interested objects.

Other researchers have used immersive technologies to augment a user's existing vision. Hindsight aimed to enhance the spatial awareness by sonifying detected objects in a real-time 360-degree video [19]. They use a deep neural network to locate and attribute semantic information to objects surrounding the user through a head-worn panoramic camera. Hindsight utilizes the combination of computer vision and spatial audio to provide 3D surrounding information for cyclists. Our proposed solution uses the same suite of technologies but for the visually impaired.

Audvert is a mobile application that attempts to give a general user a sense of place. Users can infer the direction and proximity of the point of interest (POI) on the map from the spatial audio feedback using Audvert [11]. Both ARVI and Audvert are mobile applications that attempt to locate a point of interest using immersive spatial audio. ARVI, however, is designed for the blind; as such its UI supports blind-friendly features such as voice-over and gesture recognition. Audvert aims to treat an interesting point on the map to be the POI while ARVI aims to aid the visually impaired in finding desired objects.

Soundscape, developed by Microsoft, is also a mobile application that aims to build a richer awareness of surroundings for the blind. It runs in the background in conjunction with navigation apps to provide users with additional context about the environment [6]. Similar to Audevert, Soundscape uses a drop point on map as its POI and utilizes 3D audio cues to aid the user to locate the POI. The limitation of Soundscape (and Advert) is in the precision of the GPS. As mentioned before, the "5-meter problem" can be a pain point for people without sight. ARVI provides a solution for the "5-meter problem" by using pre-scanned objects as POI and computer vision to localize those objects with much greater accuracy.

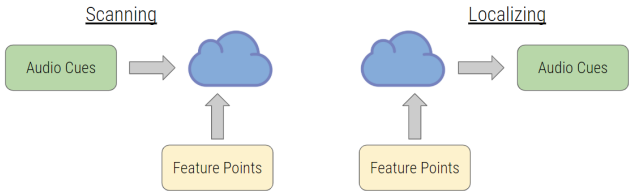

Figure 1: The two components of ARVI.

3 METHODS

ARVI consists of two main modes: scanning and localizing. In scanning mode, users capture a map of their surroundings and annotate it with audio cues, then save the map to the cloud. In localizing mode, users select nearby maps to load and once again capture data from their surroundings. Once a match to the saved map is discovered, the audio cues that were added by the map creator are made available to the user. Fig. 1 presents a visual overview of the two components of our application.

To build the iOS application itself we use Unity, a high-level game engine that allows for quick prototyping and serves as a hub for connecting different tools, SDKs, and platforms together.

3.1 Scanning

Upon starting a new map, users are instructed to slowly move their phone camera to observe their surroundings from different perspectives and distances. They can also drop audio cues in the scene that will be saved alongside the map. UI views of scanning mode are shown in Fig. 2.

3.1.1 SLAM. Simultaneous localization and mapping (SLAM) is the problem of placing a vision system in an unknown environment and asking it to build a map of its surroundings while simultaneously figuring out its location within said environment [14]. As a technology, it has many potential use cases - including its use in augmented reality - and has been a popular research topic in the computer vision and robotics fields for the past decade. We utilize a commercial SLAM solution to achieve high levels of accuracy when replicating audio cue positions that wouldn't be possible by purely relying on GPS, thus solving the the 5 meter problem described in the introduction.

More specifically, we use ARKit, a monocular SLAM solution and library built specifically for iOS devices [15]. For ease of use, we also use the Placenote SDK, which is a wrapper around ARKit designed for its ease of use and persistent mapping integration (further described below). With these libraries, we implement a procedure that performs SLAM on the user's camera feed. ARKit scans a map of the surroundings and identifies feature points in the environment that can be later used to uniquely identify the scan.

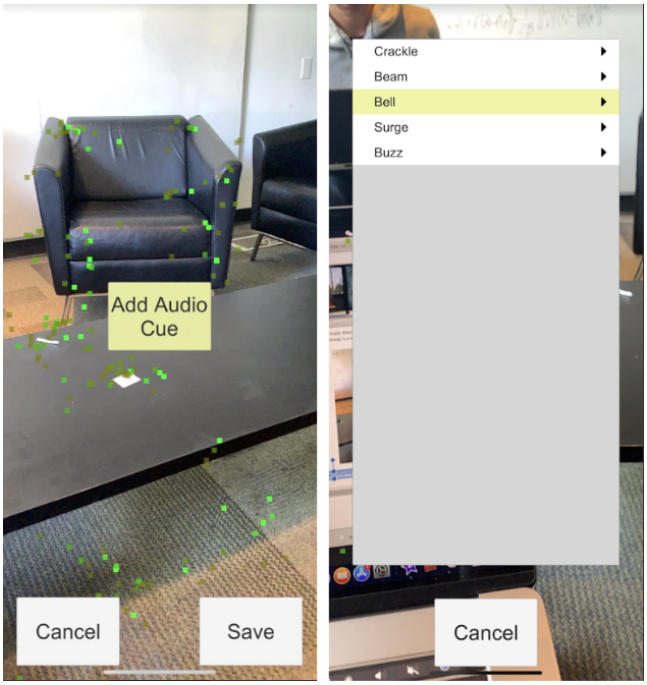

Figure 2: Screenshots of the UI while in scanning mode. Left: the main scanning menu, with options to add audio cues and save the map. The green markers represent feature points detected in the scene. Right: the audio menu, where a sound is chosen to be associated with a cue.

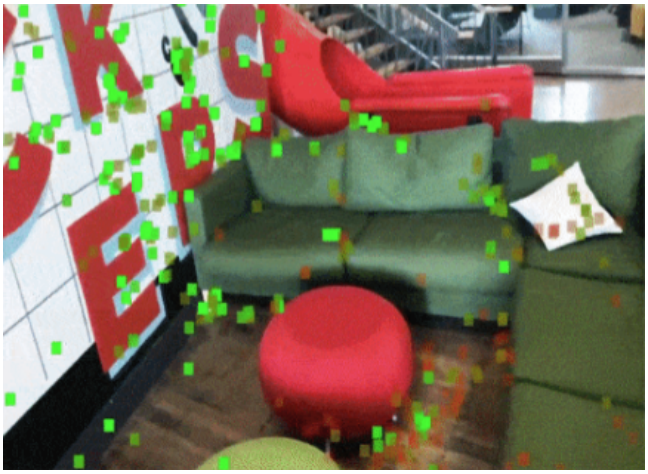

Figure 3: An example screenshot of a scene being scanned using the ARKit and Placenote SDKs. Bright green markers indicate feature points of high confidence, while darker red markers indicate ones with low confidence. Taken from [17]

An example scene with these feature points made visible can be seen in Fig. 3.

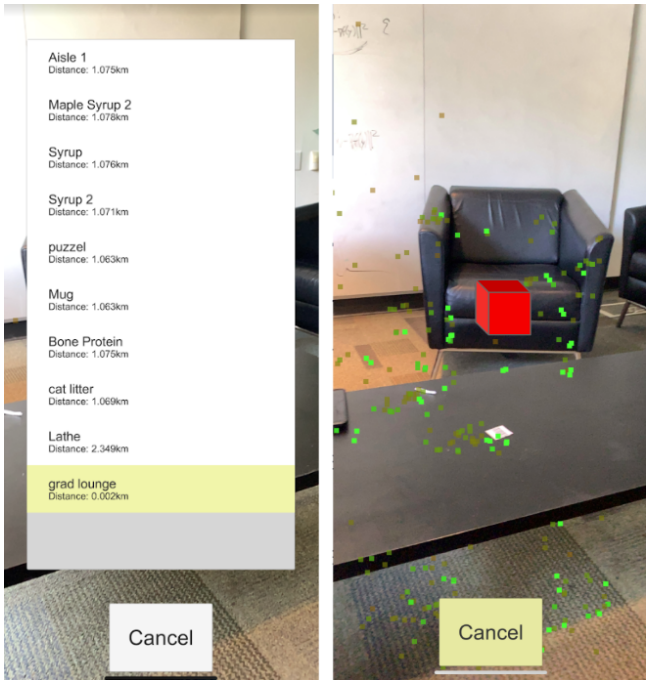

Figure 4: Screenshots of the UI while in localizing mode. Left: the map selection menu, displaying maps that are nearby. Right: the app after a match with the saved map has been made. The green markers are the previously saved feature points.

3.1.2 Audio cues and map storage. At any point in the scanning process, users can position their phone where they wish to add an audio cue and enter the Audio Cue Menu to add one to the map. They're presented with a library of audio clips that they can choose from, with the ability to hear and test out each one in the same conditions as anyone who later localizes to the same map.

Once the user has finished scanning the map and annotated elements of the scene with audio cues, they enter a separate menu where they type in the name of the map and save it. Placenote provides us a method of packaging the feature points in the map and the positions/attributes of any audio cues and uploading them to their cloud servers, where they are made accessible for users in localizing mode.

3.2 Localizing

3.2.1 Map selection and localization. In localizing mode, users are first presented with a list of maps in the area. These maps are all presented with their metadata, such as the name given to them and their predicted location from the user, computed through GPS. We filter out maps we deem too far away for the user to localize to.

Once the user selects a map, they are one again instructed to slowly move their phone camera around as ARKit's SLAM solution gathers information about the surroundings. At the same time, the Placenote SDK attempts to find a match between the map in the cloud and the map being locally constructed. Once such a match is

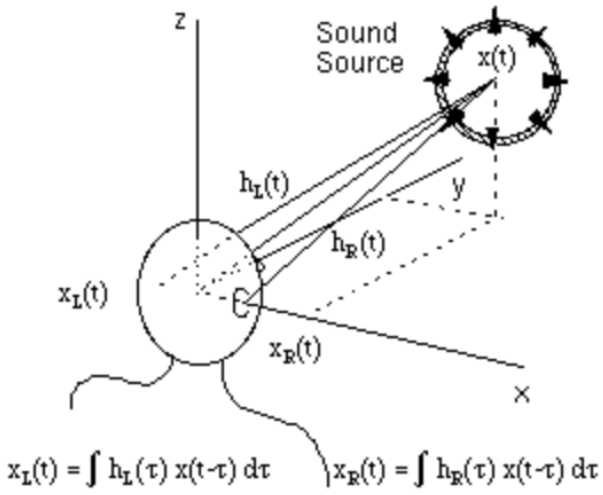

Figure 5: A diagram of the HRTF for a user. Sound from a point source differs in time-of-arrival and volume based off of the shape and positions of the ears. Taken from [4]

.

made, we take any audio cues that were saved with the map and drop them in the appropriate locations around the user.

3.2.2 Spatial audio. A key component of ARVI's usefulness is in allowing a blind user to navigate to any particular audio cue. For this to work, users must be able to distinguish where and from how far away an audio cue is coming from. To achieve this we make use of binaural audio, which is a way of generating or recording audio that seeks to accurate represent our perception of sound in the physical world. Binaural audio is an advancement over the more commonly seen stereo audio. It utilizes two variables: ITD and ILD, where interaural time differences (ITD) are the delays in sound arriving between the left and right ears, and interaural level differences (ILD) are the differences in volume [13]. These variables are adjusted through use of a head-related transfer function (HRTF) (seen in Fig. 5) which models how the ear reacts to sound at different points in space [16].

We use Google Resonance, a dynamic spatial sound toolkit [9], to bring binaural audio support to ARVI. Each of the audio cues are implemented as a spatial sound source, and the position/orientation of the phone is used to approximate the position of the user's ears.

3.3 Visually Impaired Supportive UI

As an application aimed specifically towards the visually impaired, a blind supportive user interface is required. Taking inspiration from VoiceOver, an accessibility plugin integrated within iOS, and our previous talks with blind communities, we identify two key components of such an interface: a gesture based input system and audio-based screen reader.

We design and implement a position-invariant input system that can recognize a user's gesture input regardless of their position on the phone screen. The idea is to allow users to navigate ARVI using gestures commonly found in mobile apps like single taps, double taps, and swipes, on any part of the phone screen. Doing this removes the reliance on position-based input controls such as button clicks or scrolling, which are designed with sight in mind. With this approach, users donâĂŹt have to memorize the relative position of UI elements. For consistency, we designate each gesture to the same action across the entire app. The detailed designations are as follows:

- Single taps plays the audio that indicates the current focused elements so that users can keep track of where they are in the application.
- Double taps select the current focused element.
- Horizontal swipes iterate through all possible UI elements in the current scene.
- Vertical swipes iterate through list items in the current scene, if applicable.

Like many other applications or plug-ins for the visually impaired, our application uses an audio-based screen reader to output information to users. We leverage a text-to-speech library from the Unity Asset Store to convey any text information in our app as synthesized speech, and activate it whenever the user focuses on a new UI element, transitions to a different part of the app, or taps the screen.

4 RESULTS

We use two methods to evaluate ARVI. The first is to conduct trials that test the correctness and accuracy of the application. The second is to gather feedback on the usability and user experience, particularly by those seeing the app for the first time.

4.1 Preliminary study: robustness and accuracy

We conducted an informal experiment to gather preliminary results and thus identify some immediate next steps for ARVI. The purpose of the study was to gain a better understanding of the robustness and accuracy of our app within the context of navigating to pre-scanned objects in a grocery store without use of sight. More generally, we wanted to see how a user felt about navigating through spatial audio alone.

4.1.1 Task and measurements. Participants of the study operated in pairs. One person would select a random item from a grocery aisle and use ARVI to construct a map of the surroundings, with an audio cue placed on the object. The second person would then blindfold him/herself and attempt to locate the item through the information provided by ARVI. Success/failure and time taken were recorded. For reference, we conducted similar trials where a sighted person would be asked to locate an item by name.

4.1.2 Results and observations. The detailed results were gathered into Fig. 6. In two out of four blindfolded trials, the searcher successfully located the item within two minutes. Compared to the times for sighted participants, it cost significantly more time for blindfolded participants to locate the item. This might have been caused by multiple factors: for instance, using spatial audio to identify the direction/distance of an object is by nature harder than using sight. Additionally, the participants were sighted people with their eyes covered, thus they were never trained to use spatial audio and had never relied on their hearing to navigate.

Blindfolded

Searcher	Object	Time
Mengshi	Coffee Mug	1:40
Mengshi	Protein Powder	1:53
James	Detergent	Failure
James	Cat Litter	Failure

Sighted

Searcher	Object	Time
Mengshi	Baby Wipes	0:20
Mengshi	Peanut Butter	0:25
James	Cat Food	0:29
James	Spaghetti	1:05

Figure 6: Tables of results from our preliminary study in using ARVI to navigate towards objects. Left: time recorded for blindfolded searchers to locate the pre-selected object using ARVI. Right: time recorded for sighted searcher to locate the pre-selected object without using ARVI.

There were two failure cases where ARVI failed to recognize the environment and hence was unable to provide spatial audio guidance to the participant. We concluded some insights based on these failure cases. First off, current mobile SLAM solutions are still fragile - they might require a certain perspective or distance when localizing the user, which is very damaging to our application as we cannot control what perspectives users will providing the app. Secondly, the lack of feedback during the period where ARVI is localizing itself is disorienting as users receive no indication of progress and can only move around randomly.

Although there were some flaws in its robustness, ARVI performed surprisingly well in terms of accuracy once it localized itself into a pre-scanned map In the two successful trials, participants were able to locate the audio cue (and hence the selected item) within the space of a couple inches through spatial audio alone.

4.2 Usability and user experience

We showcased our application in the Jacobs Winter Design Showcase and gathered valuable insights regarding the usability and user experience of ARVI. The participants were mostly Berkeley EECS students, including our classmates in the class. Hence we can assume that our audience had some understanding of VR/AR technologies. We placed several objects including headphones, notebooks, ear buds etc. on our demo table, scanned the environment around the table, and dropped an audio cue over the headphones. We showcased the application by explaining the basic idea and introducing the gesture-based UI of ARVI. Then we asked participants to locate the audio cue with their eyes closed. Most participants felt the spatial audio sysyem was intuitive and were able to accurately locate the headphones among all the other objects on the table. A few participants felt confused by sounds at first and were only able to locate the headphones after further explanation of the system. Below we've listed some interesting and useful observations gathered during the showcase:

- It became apparent that users did not realize the importance of matching phone orientation with head orientation. Many participants would just move the phone or just move their heads which prevented them from figuring out where the sounds was coming from.
- Matching what we found in the preliminary study, the lack of feedback in ARVI's localization mode confused a lot of participants. They seemed lost and didnâĂŹt know what was going on during that time period.
- Most of participants ended up opening their eyes to confirm that they had located the object. Several people suggested adding a more obvious indicator in addition to spatial audio cue that would sound when the user was getting close or had reached the cue.

5 CONCLUSION

In this paper we introduce ARVI, an iOS application for the visually impaired where users share and receive spatialized semantic audio cues with higher levels of precision than what GPS can currently allow. We identify the 5 meter problem as a good match between user needs and technology fit, and we provide an overview of the key components that go into making ARVI functional, including SLAM, spatial audio, and blind-friendly UI design. From our preliminary tests, we show that ARVI can be used to successfully identify objects and learn semantic information about foreign environments.

A large amount of refinement and iteration will be required to make ARVI into a tool that the visually impaired can truly use. In the short term, our team hopes to conduct more formal user testing, particularly from those who are visually impaired. There also is much work to be done in fleshing out the user experience: we plan to more tightly incorporate GPS and SLAM such that the transition between the two is seamless, and also implement speech cues (instead of just audio clips) that should greatly improve the semantic flexibility of ARVI. Finally, there is lots of work to be done in bettering both the SLAM system (improving robustness and increasing the range of operation) and spatial audio (emphasizing the directional/distance effects and incorporating additional cues to assist with navigation).

We believe we've identified an important niche that can be benefited by recent advances in immersive technologies, and that we've built the foundation of a tool that can help fill this niche. In addition, the general concept we use here of linking GPS and SLAM to enable world-level annotation and augmentation is a very powerful one and can be applied to many other use cases.

REFERENCES

[1] [n. d.]. aira: your life, your schedule, right now. https://aira.io/
[2] [n. d.]. ARCore. https://developers.google.com/ar/
[3] [n. d.]. GPS Accuracy. https://www.gps.gov/systems/gps/performance/accuracy/
[4] [n. d.]. Head-Related Transfer Functions. https://www.ece.ucdavis.edu/cipic/spatial-sound/tutorial/hrtf/
[5] [n. d.]. LightHouse for the Blind and Visually Impaired. http://lighthouse-sf.org/
[6] [n. d.]. Microsoft Soundscape A map delivered in 3D sound. https://www.microsoft.com/en-us/research/product/soundscape/
[7] [n. d.]. Oculus Rift: Step Into the Game. https://www.kickstarter.com/projects/1523379957/oculus-rift-step-into-the-game
[8] [n. d.]. orcam: help people who are blind or partially sighted. https://www.orcam.com/en/
[9] [n. d.]. Resonance Audio | Google Developers. https://developers.google.com/resonance-audio/
[10] 1970. SKERI: The Smith-Kettlewell Eye Research Institute for the. https://www.ski.org
[11] Liam Betsworth, Nitendra Rajput, Saurabh Srivastava, and Matt Jones. 2013. Audvert: Using spatial audio to gain a sense of place. In *IFIP Conference on Human-Computer Interaction*. Springer, 455–462.
[12] Michal Bujacz, Piotr Skulimowski, and Pawel Strumillo. 2011. Sonification of 3D scenes using personalized spatial audio to aid visually impaired persons. International Community for Auditory Display.
[13] Customers. 2015. What Is Binaural Audio. https://hookeaudio.com/what-is-binaural-audio/

[14] Hugh Durrant-Whyte and Tim Bailey. 2006. Simultaneous localization and mapping: part I. *IEEE robotics & automation magazine* 13, 2 (2006), 99–110.
[15] Apple Inc. [n. d.]. Get Ready for ARKit 2. https://developer.apple.com/arkit/
[16] Doris J Kistler and Frederic L Wightman. 1992. A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction. *The Journal of the Acoustical Society of America* 91, 3 (1992), 1637–1647.
[17] Neil Mathew. 2018. Building an AR house manual for your Airbnb with ARKit Placenote SDK. https://virtualrealitypop.com/building-an-ar-house-manual-for-your-airbnb-with-arkit-placenote-sdk-99422fa6029f
[18] Ari Rabinovitch. 2017. After sale to Intel, Mobileye's founder raises sights on IPO for OrCam. https://www.reuters.com/article/us-intel-mobileye-orcam/after-sale-to-intel-mobileyes-founder-raises-sights-on-ipo-for-orcam-idUSKBN1711V6
[19] 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 CHI Conference on Human Factors in Computing Systems*. ACM, 143.