Poster: Spatial Audio for Human-Object Interactions in Small AR Workspaces

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

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

ABSTRACT

While spatial audio has been an essential component in Virtual Reality, it has been rarely applied to Augmented Reality. We propose a concept and a prototype to enhance human-object interactions in daily life with 3D audio. We augment real objects in a small workspace around the user with spatial audio notifications.

CCS CONCEPTS

• **Human-centered computing** → *Auditory feedback*; *Sound-based input/output*; *Ubiquitous and mobile devices*;

1 INTRODUCTION

The auditory channel gives us an immediate 360° sense of space, time, and presence. An appropriate 3D soundscape alone enables us to perceive the environment without visual elements. Given this advantage, people usually render authentic spatial audio in VR.

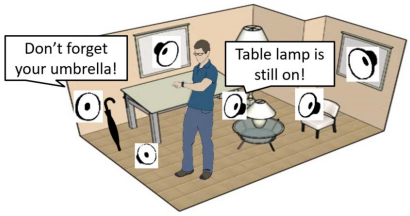

Figure 1: 360° spatial audio notifications. Image credits: SketchUp 3D Warehouse.

We anticipate that a virtual soundscape can also enhance our interaction with real objects. We design a system to attach artificial 3D audio notifications to everyday objects by tracking the user and the objects in real time. Although the objects have no speakers, the user perceives as if the sounds are authentically coming from a particular object. We believe that such a system can help enhance lively and quick 360° interactions in navigation, prompt notification (Figure 1), audio message attachment [1], and general scenarios where visual augmentation is not desired or not feasible.

2 PROTOTYPE SCENARIO

Our prototype scenario is as follows: A user works on a PC at a wooden desk, on which there is a coffee mug and a plate of fruits. We attach virtual audio notifications to the mug and the plate, which creates the impression that they are the sound sources.

MobiSys '18, June 10–15, 2018, Munich, Germany

ACM ISBN 978-1-4503-5720-3/18/06.
https://doi.org/10.1145/3210240.3210811

The proposed system consists of four functional components: object tracking, head pose estimation, environment modeling, and audio rendering. Object tracking determines the positions of the user and the objects. Head pose indicates the user's orientation relative to the objects. Environment modeling provides a digital scene model for sound propagation. It is important to model the geometry and the surface materials. Finally, artificial spatial audio is rendered to provide realistic hearing experience.

All these components can be realized with existing technologies. With wide-angle cameras, a computer can estimate the poses of the user and the objects. The results are mapped into the modeled scene, where an audio SDK will create spatial sound, which can be perceived with a hearing device like Bluetooth headset.

3 PROTOTYPE IMPLEMENTATION

The scene is modeled in Unity[1]: We assume a dominant wooden plane (desk) and another plane (screen) in front of the user. Except for the mug and the plate, any other objects in the environment are assumed to be sufficiently far away so they do not influence sound propagation. We track the user's head using OpenCV[2] with facial landmarks, and track objects by applying the Vuforia AR SDK[3] plugin. For simplicity, fiducial markers represent the objects and the user at the workplace. After mapping the real scene into the virtual model, Google's Resonance Audio SDK[4] renders the spatial audio which is delivered to the user via earphones.

In a preliminary experiment, five participants were asked to judge (1) the sound source distance (near at 15cm / far at 30cm) of four pairs of positions, (2) the sound source direction (front / back / left / right) of 16 static equidistant positions. The experiments were conducted without any visual interference. All participants could judge the distances correctly. As for the directions, although participants found front / back determination difficult since they did not have any visual clue and were not allowed to move the head as assistance, they could differentiate between left and right at a high accuracy of 86.7%. Plus, when the sound moved around, they could correctly tell the movement direction.

So far, we have demonstrated the feasibility of the idea. In the future, we will model indoor spaces considering boundaries as well as materials. We intend to realize the whole system with only head-mounted camera, smartphone, and earplugs. We aim at rendering spatial audio at arbitrary positions so that both the user and the objects can move freely in the environment.

REFERENCES

[1] Alaeddin Nassani, Huidong Bai, Gun Lee, and Mark Billinghurst. 2015. Tag It!: AR Annotation Using Wearable Sensors *(SIGGRAPH Asia '15 MGIA)*.

[1] https://unity3d.com/
[2] https://enoxsoftware.com/opencvforunity/
[3] https://www.vuforia.com/
[4] https://developers.google.com/resonance-audio/