Augmenting Smart Object Interactions with Smart Audio

Poster Abstract

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

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

ABSTRACT

The auditory output channel is rather under-utilized in smart object to human communication. One reason is that in a smart environment, multiple overlapping audio sources can be disturbing to people. We propose a wearable audio augmentation system which allows people to effortlessly select and switch between sound sources given their interest. Our system leverages visual contact via the head pose as a measure of interest towards a smart object. We demonstrate a prototype implementation in three application scenarios and a preliminary user evaluation.

CCS CONCEPTS

• **Human-centered computing** → *Auditory feedback*; *Mixed / augmented reality*; *Ubiquitous and mobile devices*;

KEYWORDS

Human-object communication; Audio augmentation; Smart objects

ACM Reference Format:
Jing Yang and Gábor Sörös. 2018. Augmenting Smart Object Interactions with Smart Audio: Poster Abstract. In *AH2018: The 9th Augmented Human International Conference, February 7–9, 2018, Seoul, Republic of Korea.* ACM, New York, NY, USA, Article 55, 3 pages. https://doi.org/10.1145/3174910.3174943

1 INTRODUCTION

With the recent advancements in speech recognition and semantic language modeling, speech input to smart appliances may soon become ubiquitous. Yet, auditory output is rather rare in today's smart appliances. One possible reason is that human ears have difficulty in filtering multiple simultaneous sound sources, and therefore sound output from multiple devices can become impractical. However, if we can aid the ears in the sound selection process, we can gain significant new bandwidth for information transfer.

Inspired by the phenomenon of *selective auditory attention* [1], we envision a wearable system that aids our selective attention by enhancing the sound coming from the object of interest while reducing or even shielding other sounds in the environment (see Figure 1). While related works on audio augmentation [2, 3] aim to create more authentic 3D perception based on the relative position between the user and the sound source, our proposed system helps to focus on the sound that the user is really interested in.

Motivated by existing prototypes that detect and respond to human attention [6, 7], in our proposed system, the audio source is selected by people's visual contact with the object. Then, both the selected and unselected objects can react correspondingly such as adjusting volume and starting/stopping speaking. Analogous to human-human communication scenario, our system may contribute to a seamless and more intuitive interaction between humans and objects, complementary to related works [4, 5] which focus on nonverbal control of objects by gaze tracking.

In this poster abstract, we present early results of our ongoing work. First, we describe the concept and advantages of a wearable smart audio augmentation system. We also demonstrate a working but not yet wearable prototype, explore three application scenarios, and describe a preliminary user evaluation.

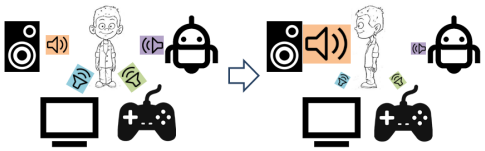

Figure 1: The user selects a sound source by gaze direction.

2 SMART AUDIO AUGMENTATION

The proposed concept consists of three components: human attention estimation, sound source selection, and sound augmentation/attenuation.

We argue that today all required components are available for *user-centric* smart audio by combining existing wearable technology: a smartphone for computing and communication, smartglasses for egocentric camera view and gaze tracking, and beamforming hearing aid technology for sound enhancement or cancellation.

We can also imagine a flipped, *object-centric* view: every smart object is assumed to be equipped with a camera. This camera can perform real-time face tracking, so that the object is able to recognize the user's attention. The second component is a computation device which decides on the sound based on the head pose determined in the previous module. The cooperating smart objects in the environment can increase or decrease their speaker volume accordingly. Alternatively, the rendered sound field can be transmitted to wireless earphones or hearing devices.

It is up to the application scenarios whether to perform user-centric smart appliance recognition or object-centric user recognition. The proposed scheme has several advantages: First, sound

AH2018, February 7–9, 2018, Seoul, Republic of Korea

ACM ISBN 978-1-4503-5415-8/18/02.
https://doi.org/10.1145/3174910.3174943

source selection based on the visual attention is analogous to human-human communication, so it is potentially more intuitive. Second, through selective auditory perception, the user's hearing experience may not only get improved but also personalized. Third, by understanding the user's attention, smart objects may be able to interact with humans more spontaneously. Fourth, this system may help disabled and elderly people who cannot move easily to better interact with objects. Visually impaired people may also find life easier with the aid of sighted objects and accompanying audio guides.

3 PROTOTYPE IMPLEMENTATION

We build our first prototype with object perspective, i.e., the user's head pose is detected by a camera mounted on the object. This is because real-time appliance recognition with wearables is still constrained today, despite promising results in the computer vision community. To examine our concept, we run simulations in virtual Unity[1] scenes. We control the user avatar's movement by the user's head pose, which is estimated in real time with PC webcam and the OpenCV library[2]. 3D sounds are rendered with the Google VR SDK for Unity[3] and then perceived with a hearing device (e.g., earplugs).

4 APPLICATION SCENARIOS

Some application scenarios are demonstrated in Figure 2. Scene 1 shows effortless audio selection in a noisy environment such as an exhibition where various products are presented at different booths. If the user would like to focus on only one of them, with our system, he/she is able to selectively exclude other sounds by facing the product of interest. In Scene 1, We embed the Hi-Fi with a piece of classical music, the TV with a movie sound clip, and the game console with a racing game sound. When the avatar faces an object, its sound is exclusively enhanced while the others are immensely lowered.

Besides, our system can provide smart audio introduction or guide in the scenario of visiting a museum. When an object is viewed by a visitor, our system can locate the focused area and play the corresponding sound clip or music. As shown in Scene 2, two audio introduction clips are attached to the cypress on the left and the blue sky on the right, respectively. When the avatar is heading towards a zone of reaction, the corresponding sound clip is played.

Furthermore, the system is able to support spontaneous and active object-human interaction in daily life. In Scene 3, the coffee machine sees the avatar looking at it and it can then respond using *"Hi! A cup of espresso?"*. The idea is also feasible on other home appliances like cleaning robot, electronic toys, etc. As the usage of these objects already involves interaction like turning a knob and remote control, augmenting them with more active interactions is less abrupt but more natural to users.

5 PILOT STUDY

We invited five people (two female and three male of ages of 25±2) to test the above scenarios. They were asked to control the avatar and describe the hearing experience. All of them agreed that the system could help them better focus on the sound of interest and perceive audio messages in a seamless, natural, and clear way. In Scene 1, one participant suggested to completely suppress the other sounds for even clearer hearing experience. This further inspired us to make a prototype adjustable in sound attenuation level.

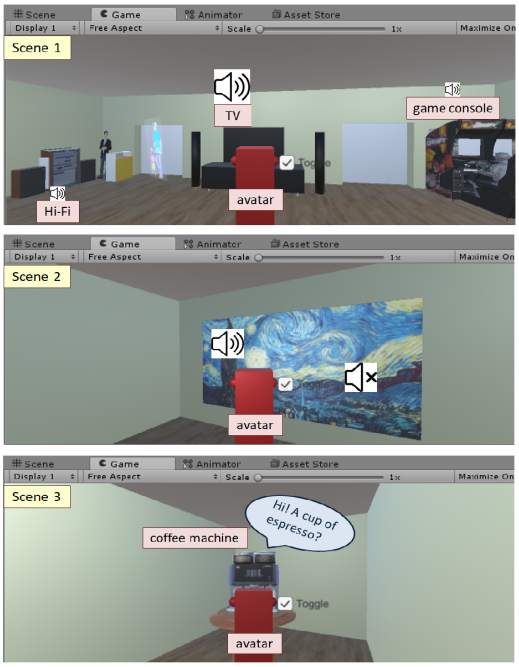

Figure 2: Three application scenarios: *Scene 1:* choosing among multiple smart objects, *Scene 2:* observing a painting, *Scene 3:* spontaneous interaction.

6 CONCLUSION AND FUTURE WORK

We have described the concept of a wearable system for effortless sound source selection based on visual attention and a prototype implementation with head tracking, and discussed a few application scenarios. A preliminary pilot study shows that the concept may enhance the user experience. We also see the potential of this system to incorporate gestures to control the audio signal.

We continue this line of work to make it truly wearable by incorporating wearable gaze trackers for egocentric visual attention tracking, and beamforming earphones for spatial audio augmentation and attenuation. In a future study, we will evaluate how psychologically or physically intrusive users feel when utilizing our system. Also, we will propose methods to handle unwanted sounds when users' gaze aimlessly wanders in the environment. Moreover, cultural differences will be considered when we prepare audio clips for different people in various scenarios.

ACKNOWLEDGMENTS

We thank Prof. Friedemann Mattern, Dr. Vlad Coroamă, and Mihai Bâce for their insightful discussions and help. We also thank Sonova AG for their technical support.

[1] https://unity3d.com/
[2] https://enoxsoftware.com/opencvforunity/
[3] https://developers.google.com/vr/unity/

REFERENCES

[1] Deborah L Arthur, Paul S Lewis, Patricia A Medvick, and Edward R Flynn. 1991. A neuromagnetic study of selective auditory attention. *Electroencephalography and Clinical Neurophysiology* 78, 5 (1991).

[2] Florian Heller and Jan Borchers. 2015. AudioScope: Smartphones as Directional Microphones in Mobile Audio Augmented Reality Systems *(CHI)*.

[3] Florian Heller, Jayan Jevanesan, Pascal Dietrich, and Jan O Borchers. 2016. Where are we?: evaluating the current rendering fidelity of mobile audio augmented reality systems. *(MobileHCI)*. 278–282.

[4] Songpo Li and Xiaoli Zhang. 2017. Implicit Intention Communication in Human–Robot Interaction Through Visual Behavior Studies. *IEEE Transactions on Human-Machine Systems* 47, 4 (Aug 2017).

[5] Shi Qiu, Matthias Rauterberg, Jun Hu, et al. 2016. Exploring Gaze in Interacting with Everyday Objects with an Interactive Cup. In 4^{th} *International Conference on Human Agent Interaction.*

[6] Jeffrey S Shell, Jeremy S Bradbury, Craig B Knowles, Connor Dickie, and Roel Vertegaal. 2003. eyecook: A gaze and speech enabled attentive cookbook. *Video Proceedings of Ubiquitous Computing* (2003).

[7] Jeffrey S Shell, Roel Vertegaal, Aadil Mamuji, Thanh Pham, Changuk Sohn, and Alexander W Skaburskis. 2003. EyePliances and EyeReason: Using Attention to Drive Interactions with Ubiquitous Appliances. *Ext. Abstracts of UIST 2003* (2003).