Arizona State University

Real-time spatial audio on the Internet of Things

Robert LiKamWa and Visar Berisha

School of Arts, Media and Engineering + School of Electrical, Computer and Energy Engineering + College of Health Solutions

Arizona State University

likamwa@asu.edu, visar@asu.edu

Funded by NSF CPS-1932377

Introduction

Spatially placed virtual sounds can create sounds in mid-air, or grant audio to objects that would otherwise be unable to produce sound, e.g., books, keys, utensils. Users could locate misplaced items through spatial audio guidance and/or allow users to interact with battery-less objects through auditory response.

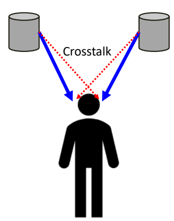
In our envisioned future, the network of speakers on Internet-of-Things devices can provide a fabric of spatial audio throughout homes, offices, and public spaces.

Challenge of Rendering Spatial Audio from Loudspeakers

Delivering spatial audio from loudspeakers is challenging due to crosstalk: the left sound signal

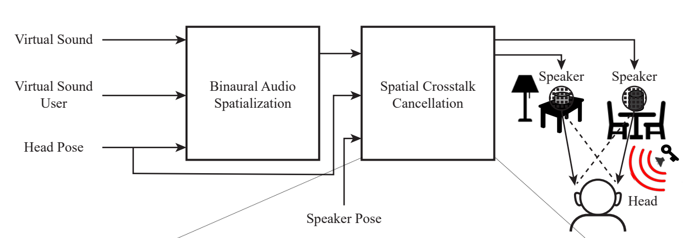
State-of-the-art crosstalk cancellation methods are expensive, requiring predesigned infrastructure and stationary observers.

- *Ambisonics* [1] and *amplitude panning* [2] require many surrounding speakers and require the user be in a specific fixed spot.
- *Dynamic crosstalk cancellers* for frequency domain audio processing have high latency and inaccuracy [3] or take significant calibration for the room and user transfer functions [4].



We propose a distributed spatial audio system that implements a time–domain dynamic crosstalk cancellation technique based on the geometry of a user's head position which uses estimations of amplitude decay and time delay to produce sound signals that become real-time binaural audio at the user's ears.

Our Solution: Geometrically-guided Spatial Audio Processing



Spatial Cross-talk Cancellation

Due to crosstalk, each ear receives a combination of decayed (α) and delayed (δ) signals from speakers S_1 and S_2 to form left and right ear signals $E_L(t)$ and $E_R(t)$.

$$E_L(t) = \alpha_{1,L} S_1(t - \delta_{1,L}) + \alpha_{2,L} S_2(t - \delta_{2,L})$$
$$E_R(t) = \alpha_{1,R} S_1(t - \delta_{1,R}) + \alpha_{2,R} S_2(t - \delta_{2,R})$$

We use estimated α and δ information to create S_1 and S_2 signals that, when combined at the ear, will present left and right inputs $I_L(t)$ and $I_R(t)$ to the E_L and E_R respectively.

$$S_1(t) = \frac{1}{\alpha_{1,L}} I_L(t - \Delta + \delta_{1,L}) - \frac{\alpha_{2,L}}{\alpha_{1,L}} S_2(t - \delta_{2,L} + \delta_{1,L})$$
$$S_2(t) = \frac{1}{\alpha_{2,R}} I_R(t - \Delta + \delta_{2,R}) - \frac{\alpha_{1,R}}{\alpha_{2,R}} S_1(t - \delta_{1,R} + \delta_{2,R})$$

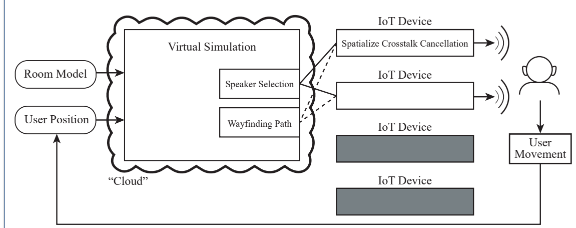
Future Exploration

Calibration: Can we re*tune system to work well in different environments?*
- Gathering and applying room and speaker sound models
- Real-time speaker selection for optimized binaural perception, including optimizing around predicted user movement (see below)

Sound Design : *What sounds work best for spatialization?*
- Understanding efficacy of natural and synthesized sound patterns
- Applying spatial motion patterns to spatial sounds for localization

Integration: *Can we provide real-time device-free tracking?*
- Integrating with visual head tracking in a room
- Distributing audio-based computing across weak and strong IoT nodes

Cloud-based room model simulation for speaker selection



References

[1] D. Artega. 2018. Introduction to Ambisonics. Research Gate. https://www.researchgate.net/publication/280010078_Introduction_to_Ambisonics

[2] V. Pullki. 1997. Virtual sound source positioning using vector base amplitude panning. J. Audio Eng. Soc., vol. 45, pp. 456–466.

[3] H. Kurabayashi, M. Otani, K. Itoh, M. Hashimoto, M. Kayama. 2013. Development of dynamic transaural reproduction system using non-contact head tracking. In 2013 IEEE 2nd Global Conference on Consumer Electronics (GCCE), (Tokyo, Japan).

[4] M. Song, C. Zhang, D. Florencio, H. Kang. 2011. An Interactive 3-D Audio System with Loudspeakers. In IEEE Transactions on Multimedia (Volume: 13, Issue: 5, Oct. 2011)