Spatial Audio Empowered Smart speakers with Xblock - A Pose-Adaptive Crosstalk Cancellation Algorithm for Free-moving Users

Frank Wencheng Liu*, Anish Narsipur*, Andrew Kemeklis, Lucy Song, Robert LiKamWa
Arizona State University
Tempe, AZ, USA
{fwliu1,anarsip2,akemekli,lsong26,likamwa}@asu.edu

ABSTRACT

Smart IoT Speakers, while connected over a network, currently only produce sounds that come directly from the individual devices. We envision a future where smart speakers collaboratively produce a fabric of spatial audio, capable of perceptually placing sound in a range of locations in physical space. This could provide audio cues in homes, offices and public spaces that are flexibly linked to various positions. The perception of spatialized audio relies on binaural cues, especially the time difference and the level difference of incident sound at a user's left and right ears. Traditional stereo speakers cannot create the spatialization perception for a user when playing binaural audio due to auditory crosstalk, as each ear hears a combination of both speaker outputs. We present Xblock, a novel time-domain pose-adaptive crosstalk cancellation technique that creates a spatial audio perception over a pair of speakers using knowledge of the user's head pose and speaker positions. We build a prototype smart speaker IoT system empowered by Xblock, explore the effectiveness of Xblock through signal analysis, and discuss future perceptual user studies and future work.

CCS CONCEPTS

• **Human-centered computing**; • **Computer systems organization** → **Sensor networks**; • **Hardware** → **Sound-based input / output**;

KEYWORDS

spatial audio; crosstalk cancellation; algorithm; internet of things

ACM Reference Format:
Frank Wencheng Liu*, Anish Narsipur*, Andrew Kemeklis, Lucy Song, Robert LiKamWa. 2023. Spatial Audio Empowered Smart speakers with Xblock - A Pose-Adaptive Crosstalk Cancellation Algorithm for Free-moving Users. In *Cyber-Physical Systems and Internet of Things Week 2023 (CPS-IoT Week Workshops '23), May 09–12, 2023, San Antonio, TX, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3576914.3589563

* Both authors contributed equally.

CPS-IoT Week Workshops '23, May 09–12, 2023, San Antonio, TX, USA

ACM ISBN 979-8-4007-0049-1/23/05…$15.00
https://doi.org/10.1145/3576914.3589563

Figure 1: Interaural time differences (ITD) and interaural level differences (ILD) provide auditory localization cues to users, as they mimic properties of sound propagation from the source to the user's two ears. Headphones preserve ITD and ILD cues, as binaural audio channels meet their respective ears. When spatialized audio is played from loudspeakers, these cues are not maintained and audio will sound from the loudspeakers directly rather than the intended position. Xblock allows users to experience real-time spatial audio by preprocessing sound signals through time shifts and gain changes based on head and speaker pose - when audio from loudspeakers reaches the user's ears the binaural spatialization ILD and ITD cues are preserved.

1 INTRODUCTION

Smart speakers have seen a rapid adoption over recent years. Imagine the possibilities if a network of these Internet-of-Things enabled speakers was capable of delivering spatialized audio to users. Spatially placed virtual sounds could create sounds in mid-air, or grant audio to objects that would otherwise be unable to produce sound. For example, users could locate misplaced items through spatial audio guidance and/or interact with objects through auditory response.

How we perceive where sounds are in the physical world comes from the combination of two primary auditory cues - interaural time difference (ITD), the difference in time between a sound reaching one ear and the other, and interaural level difference (ILD), the difference in sound pressure that reaches each ear [12]. Through audio spatialization frameworks, game engines can recreate these auditory localization cues through binaural audio over sound channels that are designated for the user's left and right ears. Advanced frameworks can leverage an individual's head related transfer function (HRTF), adapting to how each person's head characterizes sound in their unique way. Spatialized audio is primarily delivered over headphones. Spatialization frameworks typically recommend headphone use for localization cues to be effective [15] [33] [39].

Our project aims to present binaural audio not through headphones, but over loudspeakers in the user's physical environment

[25]. Playing binaural audio over a pair of loudspeakers is challenged by crosstalk, where audio from the left speaker reaches the right ear, and audio from the right speaker reaches the left ear. Due to the intermixing of audio channels and ears, crosstalk reduces the presence of ITD and ILD cues, and destroys the user's perception of audio spatialization. Crosstalk must be "cancelled" to enable spatial audio over loudspeakers, resulting in audio channels that are properly delivered to the user's ears.

In this paper, we describe our work on Xblock, a novel time-domain pose-adaptive crosstalk cancellation technique that preserves the binaural ILD and ITD cues by processing the audio based on the user's relative location to the speakers in real time. When these cues are preserved, users can experience real-time spatial audio when listening to sound signals from loudspeakers. Xblock can empower the Internet of Things smart speakers with the ability to collaboratively render sounds in the physical environment. We implement a prototype of Xblock into a networked smart speaker system. Using visual tracking, we can observe the relative position of the user with respect to the speakers . As a result, an untethered, free moving user can hear the spatialized audio rendered by smart speakers. Neither objects nor the environment need be modified in any way. The requirements are that both the user and objects of interest are in the line of sight to our "main" smart speaker hub. Additionally the other IoT speaker positions are known with a minimum requirement of two speakers for spatial audio generation.

Xblock achieves spatial audio over loudspeakers through: (*i*) a novel crosstalk cancellation time-domain algorithm that dynamically adapts to user and loudspeaker geometry, and (*ii*) a game engine plugin to integrate Xblock processing into application audio. Crosstalk cancellation has been well-studied for static users positioned centrally between loudspeaker pairs[2][5] [8] [10] [12] [18] . To the best of our knowledge, Xblock is the first work to support crosstalk cancellation for dynamically moving users, e.g., walking, ducking, and jumping around a physical environment for a loudspeaker pair system. We discuss the prototype spatial audio empowered smart speaker system utilizing Xblock. We conduct signal analysis in a live environment over a variety of user positions, demonstrating Xblock's ability to preserve ILD and ITD cues. We discuss future perceptual user studies and future work.

2 RELATED WORKS

2.1 Speaker Array Systems

Speaker array system techniques, generally using more than two speakers, are established methods to generate spatial audio. However, these techniques are expensive in cost, set up, and computation making them impractical for at home listening experiences. Wave Field Synthesis produces artificial wavefronts synthesized by a large number of individually driven loudspeakers [9] [36]. The location of a virtual sound source is made from a summation of these waves. Vector-Based Amplitude Panning applies the same sound signal to a number of loudspeakers in different directions equidistant from the listener [34]. Then, a virtual source appears to a direction that is dependent on amplitudes of the loudspeakers. Ambisonics has multiple audio channels which are played through different speakers. Typically the setup is with four different speakers with four corresponding audio channels, but ambisonics works for more speakers as well [11] [12]. Beamforming uses speaker arrays responding to the same input signal in different ways — adding slight delay to the signal, different volumes, or using cancellation effects - to enable virtual sound placement [1] [41]. Implementing beamforming is quite difficult and time intensive to programatically control for real-time virtual reality uses.

2.2 Crosstalk Cancellation

Conventional methods of static crosstalk cancellation utilize filtering via inverse crosstalk transfer function [12] or recursive filtering [14] to cancel out crosstalk. However, the crosstalk cancellation effect only works when the user is perfectly in between the two loudspeakers and limits a user's range of motion. Movements as little as 75mm can completely ruin the spatialization effect [3].

Recent work for dynamic crosstalk cancellation involve updating the inverse transfer function based on a user's HRTF from a filter bank [22] [26] [27] [31]. There have been existing works implementing low-latency dynamic crosstalk cancellation designed specifically for a CAVE environment [23] [24]. While it is very easy to walk into a CAVE and experience spatialized audio, these environments are not something one can simply bring into their living rooms.

There has been existing work towards the development of home-appropriate spatial audio systems over loudspeakers using dynamic crosstalk cancellation through updating filterbanks. However, these works either incur a noticeable delay for the user [22] or require a intensive setup in updating a room response model in addition to the filter banks [35]. The subjective listening tests of [35] had the only slightly varied head rotation or moving a user 20cm to the left/right. These techniques are unlikely to be suitable for an actively moving user in a virtual reality or augmented reality scene.

By comparison, Xblock is implemented as a low-latency software solution, easily integrable into existing systems without additional hardware/setup. The operations are in the time-domain and do not need intensive convolution operations, allowing Xblock to provide real-time spatial audio. Additionally, Xblock enables spatialized audio in a much larger area as the user can move freely compared to a small sweet spot in between the speakers.

2.3 Alternate Solutions

2.3.1 Acoustically Transparent Headphones. Xblock is intended as a software solution to be integrated into existing infrastructure; however, acoustically transparent headphones are a viable alternate solution to providing spatialized audio to the user with additional hardware. Some examples of these hardware peripherals are the Bose Frames [7], the Amazon Echo Loop [13], the AirPods Pro with Transparency mode [40], or bone conducting headphones. There has been existing research exploring these headsets in the context of auditory mixed reality [6] [28], augmented television experiences [29] and search-based tasks [4].

2.3.2 Digital Ventriloquism. The authors of "Digital Ventriloquism" [17] took a different approach in order for a smart speaker to render sounds onto passive objects in the environment. The authors built a device consisting of dense, 2D array of ultrasonic transducers which modulate a 40kHz ultrasonic signal. Upon collision with an object the signal would demodulate to audible frequencies and thus

the object would be the origin of the sound. Digital Ventriloquism is able to render sound on an object 1:1. Given that Xblock cancels crosstalk, multiple sound source object perception is possible if the smart speaker renders them in the virtual environment.

2.4 Building Towards Personalized In-Home Soundscapes

There has been interest towards developing personalized soundscapes within the home [16] [19]. To create the best experience for the user, the smart speaker IoT devices should be aware of the room environment it is in. [20] and [21] estimate the room acoustics and room impulse responses through computer vision. [38] estimates the real-room acoustic materials and captures the geometry through deep learning. While our Xblock algorithm derivation assumes an ideal multipath-sparse environment, Xblock could be combined with RIR filtering based on estimating the room material using these methods for improved crosstalk cancellation.

3 XBLOCK ALGORITHM

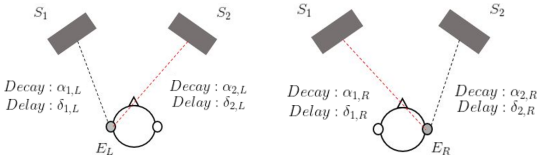
We present our Xblock algorithm, designed to preserve perceptual sound localization cues such that users can perceive spatialized audio played over loudspeakers. For our Xblock algorithm derivation, we make the assumption of an ideal multipath-sparse environment. Additionally, for our derivation, we assume that there is a direct path for the audio to propagate from the speaker to the ear without any obstruction in between.

Xblock aims to replicate the input audio channels I_L and I_R at the ears of the user E_L and E_R, through a matching delay T, such that:

$$E_L(t) = I_L(t-T) \quad \text{and} \quad E_R(t) = I_R(t-T) \tag{1}$$

As audio is delivered to the ears over a pair of loudspeakers, we model the sound that arrives at the ear as a combination of decayed and delayed sound that emanates from speakers S_1 and S_2. Representing the amplitude decay as $\alpha_{source,destination}$, and time delay as $\delta_{source,destination}$, we can model these values as functions of distance between each speaker and ear.

These α and δ values, shown in Figure 2 and modeled in Section 4, serve as inputs to the Xblock algorithm.



(a) Ear Left **(b) Ear Right**

Figure 2: Each ear hears the combination of sound from each speaker, with decay of α and delay of δ specific to the speaker-ear pair

With these, we can model the sound that arrives at E_L and E_R as combinations of propagated audio from both speakers:

$$E_L(t) = \alpha_{1,L}(S_1(t-\delta_{1,L})) + \alpha_{2,L}(S_2(t-\delta_{2,L})) \tag{2}$$
$$E_R(t) = \alpha_{2,R}(S_2(t-\delta_{2,R})) + \alpha_{1,R}(S_1(t-\delta_{1,R})) \tag{3}$$

Our goal is to derive S_1 and S_2 signals that preserve input audio at the ear, as per Equation 1. Usable S_1 and S_2 solutions will not only satisfy the above equations, but also be constructed from previous audio buffer samples (as opposed to future samples) from the input buffer and the speaker output buffer, i.e., $S(t-X)$ and $I(t-X)$, where X is positive.

3.1 Xblock: Relating Speaker Signals and Input Signals

We note a relationship between speaker signals and input signals when adhering to the desired preservation of input signals of Equation 1. When setting $E_L(t) = I_L(t-T)$ in Equation 2, we find that we can solve for S_1 as a function of I_L and S_2:

$$S_1(t) = \frac{1}{\alpha_{1,L}} I_L(t - T + \delta_{1,L}) - \frac{\alpha_{2,L}}{\alpha_{1,L}} S_2(t - \delta_{2,L} + \delta_{1,L}) \tag{4}$$

We can similarly solve for S_2 as a function of I_R and S_1:

$$S_2(t) = \frac{1}{\alpha_{2,R}} I_R(t - T + \delta_{2,R}) - \frac{\alpha_{1,R}}{\alpha_{2,R}} S_1(t - \delta_{1,R} + \delta_{2,R}) \tag{5}$$

These equations provide valid solutions for S_1 and S_2, but potentially request future samples if $\delta_{1,L} > \delta_{2,L}$ or $\delta_{2,R} > \delta_{1,R}$. However, by plugging Equation 5 into the S_2 term of Equation 4, we can arrive at the Xblock equation for Speaker 1:

$$\begin{aligned} S_1(t) = {} & \frac{1}{\alpha_{1,L}} I_L(t - T + \delta_{1,L}) \\ & - \frac{\alpha_{2,L}}{\alpha_{1,L}\alpha_{2,R}} I_R(t - T + \delta_{2,R} - \delta_{2,L} + \delta_{1,L}) \\ & + \frac{\alpha_{2,L}\alpha_{1,R}}{\alpha_{1,L}\alpha_{2,R}} S_1(t - \delta_{1,R} - \delta_{2,L} + \delta_{2,R} + \delta_{1,L}) \end{aligned} \tag{6}$$

Similarly, by plugging Equation 4 into the S_1 term of Equation 5, we can arrive at the Xblock equation for Speaker 2:

$$\begin{aligned} S_2(t) = {} & \frac{1}{\alpha_{2,R}} I_R(t - T + \delta_{2,R}) \\ & - \frac{\alpha_{1,R}}{\alpha_{1,L}\alpha_{2,R}} I_L(t - T + \delta_{1,L} - \delta_{1,R} + \delta_{2,R}) \\ & + \frac{\alpha_{2,L}\alpha_{1,R}}{\alpha_{1,L}\alpha_{2,R}} S_2(t - \delta_{1,R} - \delta_{2,L} + \delta_{2,R} + \delta_{1,L}) \end{aligned} \tag{7}$$

For the third term, the condition of constructing only from previous samples holds when $\delta_{1,R} + \delta_{2,L} > \delta_{1,L} + \delta_{2,R}$. If this is not the case, then we can swap the speaker indexing (Speaker 1 becomes 2 and vice-versa). For the other two terms, the condition holds when $T > \delta_{1,L} + \delta_{2,R}$, i.e., the sum of the propagation delays from speaker to each ear. As T can be arbitrarily set, this does not pose a problem. Altogether, the Xblock equations satisfy all conditions to: (*i*) produce speaker signals from previously buffered speaker and input signals, and (*ii*) reconstructs input audio as it arrives at the user's ears.

4 IMPLEMENTATION

4.1 Xblock Input Parameters

In order for Xblock algorithm to work properly, we need the proper input parameters. To obtain these parameters, we characterized the gain decay and sound delay of the speaker audio as a function of

distance. To characterize the gain decay, we used the inverse square law fall-off to create a linear fit of distance to decibel level. The inverse square law states that the sound intensity will diminish by 6 decibels when doubling the distance from a sound source in a free-field environment. We recognize that in reality the sound sources do not radiate evenly in all directions and room reverberations impact the sound distribution; however, we make this assumption about the behavior of sound in order to generalize the implementation of our Xblock algorithm. In order to find the proper delay values, we divide the distances of each ear relative to each speaker by the speed of sound.

4.2 Prototype Spatial Audio-Enabled Smart Speaker System

We developed a prototype spatial audio-enabled smart speaker system with Xblock capabilities to quantify the performance of the algorithm. This system is made up of four key components: spatial sound synthesis, visual tracking system, network, and smart speakers.

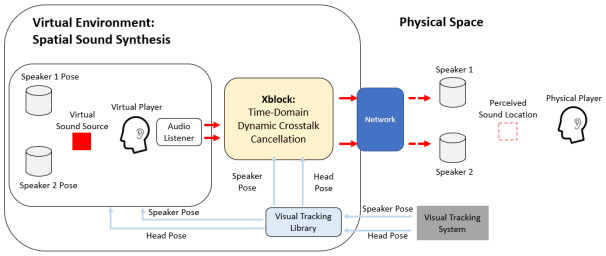


Figure 3: Block diagram of Xblock integration

4.2.1 Spatial Sound Synthesis. For our future vision of the spatial audio-enabled smart speaker system, we imagine a main speaker hub to process the virtual environment (VE) and synthesize the spatial sound. In our prototype, we utilize an Alienware m15 laptop, 3.00GHz, running Unity serving the role of the main speaker hub for the spatial sound synthesis. We integrate the Xblock algorithm into the Resonance Audio source code [15], which builds to a Unity audio plugin. Resonance Audio is an SDK that generates high fidelity spatial audio by simulating audio propagation and interaction with human ears.

We build our modified Resonance Audio Plugin in Unity version 2018.4.14f1 LTS. Generated from the head and speaker poses in the VE, the α gain and δ delay input parameters are sent to the Xblock plugin every frame update cycle. The Xblock plugin uses these parameters to modulate the audio signals, producing audio output that can be routed to a stereo (2-channel) speaker system, or send the audio output to two IoT smart speakers. **Notably, Xblock processing is lightweight, and able to generate audio with negligible latency;** on a 2.3 GHz MacBook (2019), the processing took 105.0 +/- 31.5 microseconds over 200 runs. The plugin also gives the user the option to bypass Xblock processing.

4.2.2 Visual Tracking System. For headsetless tracking of the user and objects in the environment, we envision our future main speaker hub to have visual tracking capabilities. For our prototype, we utilize Azure Kinects for visual tracking which are connected to the Alienware m15 laptop serving as the main hub. The Azure Kinect sensors accurately capture the physical space and maps what is tracked in the physical space into the Unity virtual environment for the spatial sound synthesis. We utilize Light Buzz's Azure Kinect for Unity3D package which utilizes the Kinect body tracking and camera SDKs. One Azure Kinect tracks the user's head and body pose and maps the physical poses in virtual space. A second tracks the objects and maps the physical locations in virtual space.

4.2.3 Networking. In order to send the synthesized spatial audio from the Alienware laptop to the smart speakers, we route the audio output from the Alienware laptop using JackAudio to the smart speakers through the network with Netjack. Netjack allows realtime audio transport over an IP network. Netjack syncs all the clients to one soundcard resolving any resampling, glitches or syncing issues caused by the network.

4.2.4 Smart Speakers. For our smart speaker setup, we use a Raspberry Pi with a stereo speaker system plugged into the Pi. The Raspberry Pi also has JackAudio and Netjack installed. The Raspberry Pi receives the audio transported from the Alienware laptop acting as the main hub.

Our future vision of the spatial audio-enabled smart speaker system, we imagine the smart speakers will know and send their location in relation to the hub automatically to the hub. In our prototype, we have set two configuration routines. The first configuration routine is a simple automatic speaker placement. The Kinects will find the speakers in the room through object detection. If a speaker is found on the left then it will be designated as the left speaker. The same applies for the right. The user can clap and swap the position of the left and right speakers.

We created a manual configuration routine in which the user can set the positions of virtual speakers in the VE to map to the positions of the physical speakers by performing a body action such as clapping - the Kinect finds the midpoint of where the hands are when clapped and places the virtual speaker at that midpoint. To place the location of each physical speaker while in the configuration mode, the user's first clap at the physical location of the right speaker sets the virtual location of right speaker. The second clap at the physical location of the left speaker sets the virtual location of left speaker.

While there are many different methods to localize indoor objects [32][30], such localization techniques are not the focus of this current work.

5 EVALUATION - ITD AND ILD MEASUREMENTS FROM LIVE RECORDINGS

To assess Xblock's ability to produce perceived spatialization and preserve ILD and ITD cues, we captured and measured these cues in the physical environment. Our ground truth baselines are the ILD and ITD cues generated from the audio that directly comes from the VE – the spatialized audio that would be heard through headphones. We compared the ground truth ILD and ITD cues with those coming from normal unprocessed audio (No Xblock) and Xblock processed audio (Xblock) out of a speaker pair. Our

results show that Xblock processed audio maintains ILD and ITD cue patterns that match closer to that of the ground truth compared to normal audio without Xblock processing when measured in an ideal multi-path sparse environment.

5.1 ILD & ITD Generation

To generate the ILDs and ITDs for analysis, we have the sound source object start 1m to the left, 1m behind the user, and at the user's head height and moves in a 2D raster scan pattern of a 2m by 2m square around the user with a step size of 0.1m inside the VE. The sound source object moves to a different position every second. The raster scan pattern is centered around the user's head pose. The audio is played over the loudspeakers.

5.2 Live Recording Setup

We captured live recordings on a physical dummy head with physical speakers to compare Xblock and No Xblock ILD and ITD maps in a real environment. The binaural dummy head microphone was put on a tripod and the eye level was measured to be 1.55m tall. We physically positioned the dummy head with various player positions. The physical room was fairly ideal as it was a multi-path sparse environment as seen in Fig. 6. Additionally, there were sound absorbing curtains to the left and to the right of the binaural dummy head. There was no hum of the air conditioner to influence the recordings.

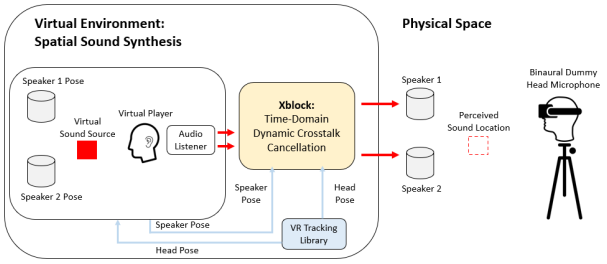


Figure 4: Live Recording Setup

Given that the dummy head has no physical body - we utilized the Oculus Quest tethered via Oculus Link (Fig. 5) for our tracking of the dummy head pose instead of using the two Kinects. For tracking the position of the speakers, we created configuration routines such that we could set the positions of virtual speakers in the VE to map to the positions the physical speakers by placing the left controller at the location of each physical speaker and pressing either "X" and "Y" while in the configuration mode. To demonstrate the capabilities of Xblock spatialization on affordable speakers, we routed our audio to T10 Inspire Speakers, which are commonplace speakers found online for roughly $30 USD.

5.3 Analysis

ILD & ITD Measurement

We measured the left and right audio channels recorded by the binaural dummy head. To create a level versus displacement map, a repeated sound sequence of 0.5 seconds of silence, a 0.3 second pulse of white noise (20 Hz to 20k Hz for full range of human hearing), and 0.2 seconds of silence plays from the sound source object. For each position, the level for each channel is calculated by the sum of squares values of every 1-second window. We obtained the ILD versus displacement map by taking the logarithm of the level map of the left channel divided by the level map of the right channel.

To create the delay versus displacement maps, we estimated the time delay of each 1-second window between the left and right audio channels from the white noise pulse recordings. Our time delay estimation utilized a function which calculated the cross-correlation between two signals and estimated the location of the peak through cosine interpolation [37].

We created ILD and ITD maps for a variety of player positions for the 1.55m tall binaural dummy head where the speaker placement of speakers +/- 0.3m to left and right, at a height of 1.05m.

The different player positions consisted of: (*i*) Where the player was directly facing the midpoint of the speakers at distances 1m, 2m, and 3m away from the speakers (*ii*) Where the player was facing towards the speakers at a distance 2m from the speakers, moving to the left in 0.1m increments until the left speaker position. (*iii*) Where the player was at a distance 2m from the speakers with 10°, 20°, and 30° rotation around the vertical axis. We assume that left and right translations and rotations are symmetric. We also assume that forward facing and backward facing are also symmetric.

Quantitative Comparison

We used root-mean-square error (RMSE) to compare the Ground Truth with the No Xblock and Xblock audio recordings. We took the RMSE of the ILD and ITD maps between the Ground Truth with No Xblock and Ground Truth with Xblock.

Discussion

In our recordings from an ideal room, the live recordings suggest that Xblock preserves ILD and ITD maps, increasing the likelihood that users will perceive spatialized audio. From Figure7, We see that the ILD and ITD maps generated from audio processed by Xblock follows a more similar distribution to the ground truth compared to that of No Xblock. From the ILD RMSE in Table 1 for the various

Figure 5: Dummy Head with Oculus Quest

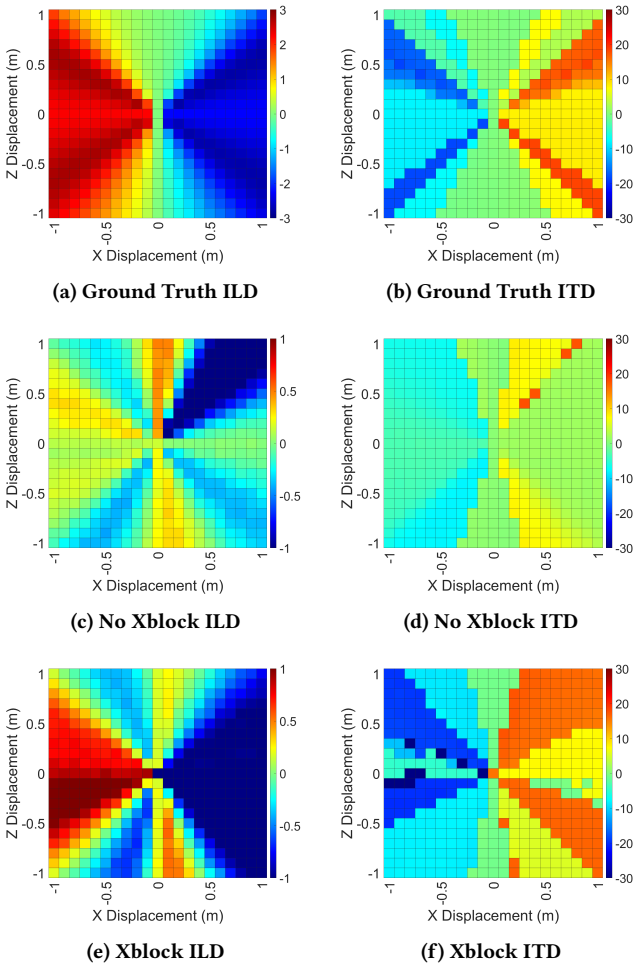
Figure 6: Live Recording Setup Example

Distance from Speakers Facing Towards	RMSE No Xblock	RMSE Xblock
1m away	1.36	1.23
2m away	1.73	1.33
3m away	1.79	1.58
Horizontal Translation	**RMSE No Xblock**	**RMSE Xblock**
0.1m left	1.83	1.50
0.2m left	1.76	1.40
0.3m left	1.83	1.61
Rotation	**RMSE No Xblock**	**RMSE Xblock**
10°left	2.08	1.59
20°left	2.36	2.19
30°left	2.60	2.51

Table 1: ILD RMSE Table

Distance from Speakers Facing Towards	RMSE No Xblock	RMSE Xblock
1m away	6.93	6.52
2m away	7.83	8.14
3m away	8.74	8.94
Horizontal Translation	**RMSE No Xblock**	**RMSE Xblock**
0.1m left	7.77	8.28
0.2m left	7.18	9.88
0.3m left	7.69	10.60
Rotation	**RMSE No Xblock**	**RMSE Xblock**
10°left	8.98	10.32
20°left	11.69	10.68
30°left	16.20	10.22

Table 2: ITD RMSE Table



(a) Ground Truth ILD (b) Ground Truth ITD

(c) No Xblock ILD (d) No Xblock ITD

(e) Xblock ILD (f) Xblock ITD

Figure 7: ILD and ITD Maps from Live Recording for Speaker Position (0.6m apart, height of 1.05m). User Position (2m away). Each square in the chart represents a displacement position of the virtual sound source. The center of each chart is the location of the user.

dummy head positions the Xblock error is smaller than those of No Xblock. From the ITD RMSE Table 2 for the various dummy head positions, the error is larger; however, we note that with Xblock, the values in the ITD map are closer in magnitude (albeit larger) to those of the Ground Truth, while the No Xblock values in the ITD map are much less pronounced than those of the Ground Truth.

6 FUTURE WORK

As future work, we would like to conduct human perception studies to better quantify the performance of the spatial audio processed with Xblock and understand how Xblock impacts the listening experience. Spatial audio processed with Xblock will be compared against baselines of spatial audio played regularly over loudspeakers and spatial audio played over headphones. In order to measure a user's localization ability across the three conditions, we aim to explore general directionality perception tasks, as well as a real world use case of finding a sonically empowered object. The user will be seated in the general directionality perception tasks and allowed to walk around in the real world use case.

To study the performance of the three conditions in different environments, these user studies are to take place in a small room, and large room. To study the impact of speaker positioning, 3 different speaker placements will be explored.

We would like to take a holistic approach in understanding what a user is perceiving. Measurements such as user confidence in localization, perceived audio quality, and testimonials of the listening experience are planned. These measurements could be captured through a series of Likert scale and long response questions.

7 CONCLUSION

We have presented Xblock, a novel time-domain technique that creates spatial audio perception over a pair of speakers using knowledge of the user's head pose and speaker positions to eliminate auditory crosstalk. A prototype IoT smart speaker system was implemented with Xblock capabilities to demonstrate and measure the algorithm's effectiveness through quantitative studies. We discuss future perceptual user studies and future work.

Xblock illuminates new and interesting uses of audio, enabling perceived acoustic spatialization over the existing IoT smart loudspeakers infrastructure for a dynamically-moving user without significant calibration or setup. This acoustic spatialization allows acoustic embodiment for inanimate objects and immersion not currently offered through audio played over speakers. Beyond giving inanimate objects the ability for acoustic feedback, Xblock could create a new world of acoustic interactions for assistive voice agents, room pathfinding for the visually impaired, and a host of immersive auditory interactions.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. 2026920. We also thank the School of Arts, Media and Engineering (AME) for the support.

REFERENCES

[1] Anastasios Alexandridis, Anthony Griffin, and Athanasios Mouchtaris. 2013. Capturing and Reproducing Spatial Audio Based on a Circular Microphone Array. *Journal of Electrical and Computer Engineering* 2013 (21 Mar 2013), 718574. https://doi.org/10.1155/2013/718574

[2] Mingsian Bai and Chih-Chung Lee. 2006. Development and implementation of cross-talk cancellation system in spatial audio reproduction based on subband filtering. *Journal of Sound and Vibration* 290 (03 2006), 1269–1289. https://doi.org/10.1016/j.jsv.2005.05.016

[3] Mingsian R Bai and Chih-Chung Lee. 2006. Objective and subjective analysis of effects of listening angle on crosstalk cancellation in spatial sound reproduction. *The Journal of the Acoustical Society of America* 120, 4 (October 2006), 1976—1989. https://doi.org/10.1121/1.2257986

[4] Amit Barde, Matt Ward, Robert Lindeman, and Mark Billinghurst. 2020. The Use of Spatialised Auditory and Visual cues for Target Acqusition in a Search Task. *Journal of the Audio Engineering Society* (august 2020).

[5] Benjamin b. Bauer. 1961. Stereophonic Earphones and Binaural Loudspeakers. *Journal of the Audio Engineering Society* 9, 2 (april 1961), 148–151.

[6] Valentin Bauer, Anna Nagele, Chris Baume, T. Cowlishaw, H. Cooke, Chris Pike, and P. Healey. 2019. Designing an Interactive and Collaborative Experience in Audio Augmented Reality. In *EuroVR*.

[7] Bose. [n. d.]. Wearables by Bose - AR Audio Sunglasses. (2019). https://www.bose.com/en_us/products/smart_products/sp_frames.html Accessed: 2020.

[8] Duane H. Cooper and Jerald L. Bauck. 1989. Prospects for Transaural Recording. *Journal of the Audio Engineering Society* 37, 1/2 (january/february 1989), 3–19.

[9] Etienne Corteel. 2007. Synthesis of Directional Sources Using Wave Field Synthesis, Possibilities, and Limitations. *EURASIP Journal on Applied Signal Processing* 2007 (01 2007), 188–188. https://doi.org/10.1155/2007/90509

[10] P. Damaske. 1971. Head-Related Two-Channel Stereophony with Loudspeaker Reproduction. *The Journal of the Acoustical Society of America* 50, 4B (1971), 1109–1115. https://doi.org/10.1121/1.1912742 arXiv:https://doi.org/10.1121/1.1912742

[11] Matthias Frank, Franz Zotter, and Alois Sontacchi. 2015. Producing 3D Audio in Ambisonics. In *Audio Engineering Society Conference: 57th International Conference: The Future of Audio Entertainment Technology – Cinema, Television and the Internet*. http://www.aes.org/e-lib/browse.cfm?elib=17605

[12] William Gardner. 2005. 3-D Audio Using Loudspeakers. (09 2005).

[13] Samuel Gibbs. [n. d.]. Amazon launches Alexa smart ring, smart glasses and earbuds. (2020). https://www.theguardian.com/technology/2019/sep/26/amazon-launches-alexa-smart-ring-smart-glasses-and-earbuds Accessed: 2020.

[14] Ralph Glasgal. 2007. 360° Localization via 4.x RACE Processing. In *Audio Engineering Society Convention 123*. http://www.aes.org/e-lib/browse.cfm?elib=14358

[15] Marcin Gorzel, Andrew Allen, Ian Kelly, Julius Kammerl, Alper Gungormusler, Hengchin Yeh, and Francis Boland. 2019. Efficient Encoding and Decoding of Binaural Sound with Resonance Audio. In *Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio*. http://www.aes.org/e-lib/browse.cfm?elib=20446

[16] Gabriel Haas, Evgeny Stemasov, Michael Rietzler, and Enrico Rukzio. 2020. Interactive Auditory Mediated Reality: Towards User-Defined Personal Soundscapes. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference* (Eindhoven, Netherlands) *(DIS '20)*. Association for Computing Machinery, New York, NY, USA, 2035–2050. https://doi.org/10.1145/3357236.3395493

[17] Yasha Iravantchi, Mayank Goel, and Chris Harrison. 2020. Digital Ventriloquism: Giving Voice to Everyday Objects. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/3313831.3376503

[18] J. Blauert. 1999. *Spatial hearing – The psychophysics of human sound localization*. The MIT Press.

[19] Stine Schmieg Johansen and Peter Axel Nielsen. 2019. Personalised Soundscapes in Homes. In *Proceedings of the 2019 on Designing Interactive Systems Conference* (San Diego, CA, USA) *(DIS '19)*. Association for Computing Machinery, New York, NY, USA, 813–822. https://doi.org/10.1145/3322276.3322364

[20] Hansung Kim, Luca Remaggi, Philip J.B. Jackson, and Adrian Hilton. 2019. Immersive Spatial Audio Reproduction for VR/AR Using Room Acoustic Modelling from 360° Images. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. 120–126. https://doi.org/10.1109/VR.2019.8798247

[21] Taeyoung Kim, Youngsun Kwon, and Sung-Eui Yoon. 2020. Real-time 3-D Mapping with Estimating Acoustic Materials. In *2020 IEEE/SICE International Symposium on System Integration (SII)*. 646–651. https://doi.org/10.1109/SII46433.2020.9025860

[22] H. Kurabayashi, M. Otani, K. Itoh, M. Hashimoto, and M. Kayama. 2013. Development of dynamic transaural reproduction system using non-contact head tracking. In *2013 IEEE 2nd Global Conference on Consumer Electronics (GCCE)*. 12–16.

[23] Tobias Lentz. 2006. Dynamic Crosstalk Cancellation for Binaural Synthesis in Virtual Reality Environments. *J. Audio Eng. Soc* 54, 4 (2006), 283–294. http://www.aes.org/e-lib/browse.cfm?elib=13677

[24] Tobias Lentz, Dirk Schröder, Michael Vorländer, and Ingo Assenmacher. 2007. Virtual Reality System with Integrated Sound Field Simulation and Reproduction. *EURASIP J. Adv. Signal Process* 2007, 1 (Jan. 2007), 187. https://doi.org/10.1155/2007/70540

[25] Frank Liu and Robert LiKamWa. 2019. Demo: A Spatial Audio System for the Internet-of-Things. In *Proceedings of the 20th International Workshop on Mobile Computing Systems and Applications* (Santa Cruz, CA, USA) *(HotMobile '19)*. Association for Computing Machinery, New York, NY, USA, 183. https://doi.org/10.1145/3301293.3309567

[26] B. Masiero, J. Fels, and M. Vorländer. 2011. Review of the crosstalk cancellation filter technique.

[27] B. Masiero and M. Vorländer. 2014. A Framework for the Calculation of Dynamic Crosstalk Cancellation Filters. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 22, 9 (2014), 1345–1354.

[28] Mark McGill, Stephen Brewster, David McGookin, and Graham Wilson. 2020. Acoustic Transparency and the Changing Soundscape of Auditory Mixed Reality. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3313831.3376702

[29] Mark McGill, Florian Mathis, Mohamed Khamis, and Julie Williamson. 2020. Augmenting TV Viewing Using Acoustically Transparent Auditory Headsets. In *ACM International Conference on Interactive Media Experiences* (Cornella, Barcelona, Spain) *(IMX '20)*. Association for Computing Machinery, New York, NY, USA, 34–44. https://doi.org/10.1145/3391614.3393650

[30] Anca Morar, Alin Moldoveanu, Irina Mocanu, Florica Moldoveanu, Ion Emilian Radoi, Victor Asavei, Alexandru Gradinaru, and Alex Butean. 2020. A Comprehensive Survey of Indoor Localization Methods Based on Computer Vision. *Sensors* 20, 9 (2020). https://doi.org/10.3390/s20092641

[31] A. Mouchtaris, P. Reveliotis, and C. Kyriakakis. 2000. Inverse filter design for immersive audio rendering over loudspeakers. *IEEE Transactions on Multimedia* 2, 2 (2000), 77–87.

[32] Huthaifa Obeidat, Wafa Shuaieb, Omar Obeidat, and Raed Abd-Alhameed. 2021. A Review of Indoor Localization Techniques and Wireless Technologies. *Wireless Personal Communications* 119 (07 2021). https://doi.org/10.1007/s11277-021-08209-5

[33] Oculus. 2020 Retrieved September 8, 2020. Oculus Audio Spatializer. https://developer.oculus.com/downloads/package/oculus-spatializer-unity/

[34] Ville Pulkki. 1997. Virtual Sound Source Positioning Using Vector Base Amplitude Panning. *J. Audio Eng. Soc* 45, 6 (1997), 456–466. http://www.aes.org/e-lib/browse.cfm?elib=7853

[35] M. Song, C. Zhang, D. Florencio, and H. Kang. 2011. An Interactive 3-D Audio System With Loudspeakers. *IEEE Transactions on Multimedia* 13, 5 (2011), 844–855.

[36] Sascha Spors, Rudolf Rabenstein, and Jens Ahrens. 2008. The Theory of Wave Field Synthesis Revisited. 1 (01 2008).

[37] Linas Svilainis. 2021. GetTOFcos(MySignal,RefSignal). https://www.mathworks.com/matlabcentral/fileexchange/65229-gettofcos-mysignal-refsignal

[38] Zhenyu Tang, Nicholas J. Bryan, Dingzeyu Li, Timothy R. Langlois, and Dinesh Manocha. 2020. Scene-Aware Audio Rendering via Deep Acoustic Analysis. *IEEE Transactions on Visualization and Computer Graphics* 26, 5 (May 2020), 1991–2001. https://doi.org/10.1109/tvcg.2020.2973058

[39] ValveSoftware. 2020 Retrieved September 8, 2020. Steam Audio. https://github.com/ValveSoftware/steam-audio

[40] Chris Welch. [n. d.]. Apple AirPods Pro hands-on: the noise cancellation really works. (2020). https://www.theverge.com/2019/10/29/20938740/apple-airpods-pro-hands-on-noise-cancellation-photos-features Accessed: 2020.

[41] Franz Zotter, Markus Zaunschirm, Matthias Frank, and Matthias Kronlachner. 2017. A Beamformer to Play with Wall Reflections: The Icosahedral Loudspeaker. *Computer Music Journal* 41 (09 2017), 50–68. https://doi.org/10.1162/comj_a_00429