Audio Engineering Society

Convention Paper 10588

Presented at the 152nd Convention
2022 May, In- Person and Online

Capturing Spatial Room Information for Reproduction in XR Listening Environments

Michael Matsakis[1], Parichat Songmuang[1], and Kathleen"Ying-Ying" Zhang[2]

[1]*New York University, 35 West 4th St, New York, NY 10012*
[2]*McGill University, 555 Sherbrooke St. W. Montreal, Quebec, Canada H3A 1E3*

Correspondence should be addressed to Michael Matsakis (matsakis.michael@gmail.com)

ABSTRACT

An expansion on previous work involving "holographic sound recording" (HSR), this research delves into how sound sources for directional ambience should be captured for reproduction in a 6-DOF listening environment. We propose and compare two systems of ambient capture for extended reality (XR) using studio-grade microphones and first-order soundfield microphones. Both systems are based on the Hamasaki-square ambience capture technique. The Twins-Hamasaki Array utilizes four Sennheiser MKH800 Twins while the Ambeo-Hamasaki Array uses four Sennheiser Ambeo microphones. In a preliminary musical recording and exploration of both techniques, the spatial capture from these arrays, along with additional holophonic spot systems, were reproduced using Steam Audio in Unity's 3D engine. Preliminary analysis was conducted with expert listeners to examine these proposed systems using perceptual audio attributes.The systems were compared with each other as well as a virtual ambient space generated using Steam Audio as a reference point for auditory room reconstruction in XR. Initial analysis shows progress towards a methodology for capturing directional room reflections using Hamasaki-based arrays.

1 Introduction

This research builds on previous work involving "Holographic Sound Recording (HSR)" or holophonic recording techniques. An HSR array utilizes spaced microphone techniques to capture the complex radiation of a sound source with the intent to reproduce an accurate, transferable, and 3D image of the source to a virtual environment with six degrees of freedom (6-DOF) or in a compatible multi-driver and multi-directional speaker system [1].

The goal of this project is to provide preliminary support for a Hamasaki-based microphone system designed to capture spatial auditory room signals for reproduction in extended reality (XR) listening environments. The Hamasaki Square utilizes four outward-facing bidirectional microphones arranged in a physical square [2]. The XR-capture system proposed in this paper replaces these bidirectional microphones with either dual-capsule or first order Ambisonics.

The first version of the array, referred to as the Twins-Hamasaki, employs Sennheiser 800 Twin microphones. These dual-output microphones contain two coincident cardioid capsules that are able to be summed

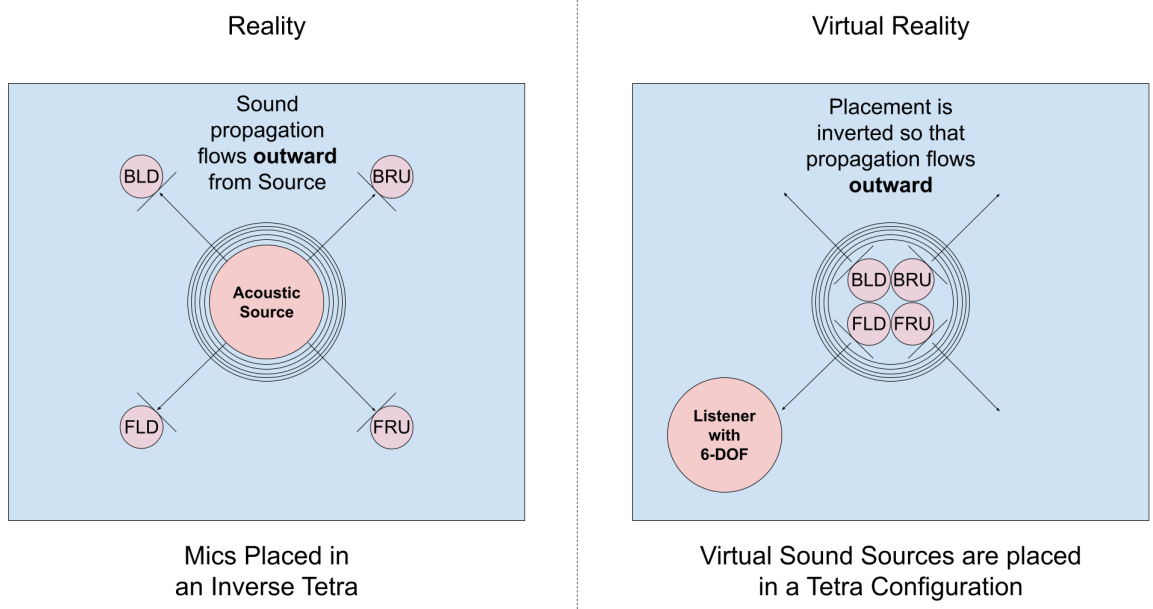

Fig. 1: Reflective audio object placement shows the inverted relationship between physical capture and reproduction in XR.

in post-production to create different polar patterns. The second array, the Ambeo-Hamasaki, consists of four Sennheiser Ambeo first-order Ambisonics microphones.

1.1 Directional Room Acoustics and Reflective Object Placement

Audio for XR applications has unique considerations and challenges in comparison to traditional playback formats. Llewellyn and Paterson note that current listening systems are differentiated by their "linearity" and "perspective," both of which are deeply affected by the interactive environment [3]. As the listener moves planarly or geodesically, there is a tendency to hear the sound timbre change to some degree. This dynamism creates a sense of room immersion. By capturing these diffuse reflections in a directional way, we have the ability to simulate a real-world environment from a specific performance. It has previously been shown that if a holophonic system is recorded in a heavily reverberant space, capturing spatial room information is crucial in order to mask any mis-localizations that might occur between the microphone systems capturing the direct sound of the instruments[1].

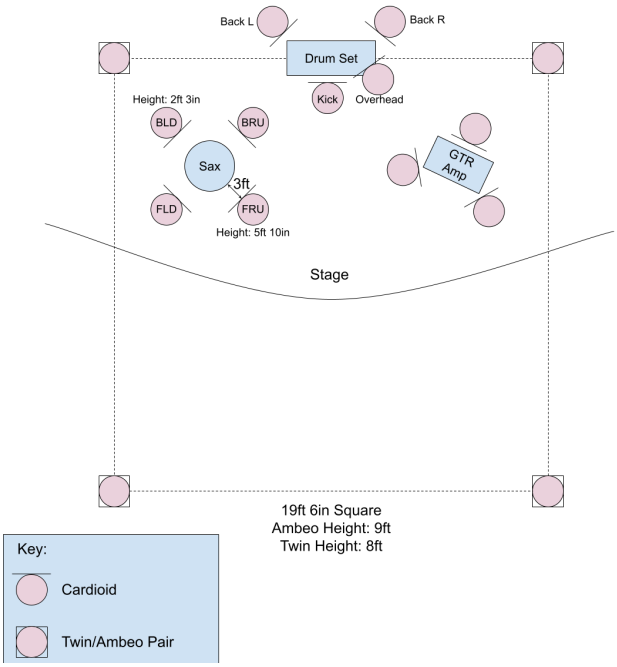

Fig. 2: Diagram showing the total microphone plan for recording the jazz trio including various HSR spot microphone systems and both the Twins-Hamasaki systems

Reproducing acoustically captured sound sources in the virtual world demands a shift in perspective. When recording ambience within a physical environment, the microphone (representing the listener) captures the reflections created by a sound source as they bounce off the enclosing walls. Physically, the reflected sound waves move toward the microphone. When this signal is placed in a virtual environment, the position of the microphone and sound source are reversed: the initial frontal position of the sound source is inverted. Therefore, in this virtual space, the microphone's perspective represents the same position as the sound source in the physical space. Taking this into account at the point of capture creates more opportunities for interaction in XR.

2 Methods

In order to provide proof-of-concept for the proposed capture and reproduction techniques, both the Twins-Hamasaki and Ambeo-Hamasaki systems were set up in a shared session in NYU's Frederick Loewe Theatre. The holophonic array was formed around the players, one half in the audience and the other at the rear of the stage, with all capsules at the same relative height in space. The microphones were placed as coincidently to each other as possible without causing occlusion. To accomplish this, the Ambeo was placed directly above the Twin (in its null point), in order to capture ceiling and floor reflections. The Twin was directly beneath the Ambeo (in its weakest point of response), capturing wall reflections. The arrays were recreated in Unity to both mask the possibility of mis-localized reflections of instruments, and provide a realistic sense of space in the recording (see Figure 2).

Each instrument was recorded using a holophonic spot array that captured the directionality of its radiation pattern. These arrays were designed based on methodologies derived from previous work [1]. The saxophone, having a somewhat spherical radiation, was captured using an inverse tetra. The drum set includes multiple points of interest, each with their own spherical distribution, so a Multi-Timbral Holophone array was used. The guitar amp has a more discrete directional characteristic, with the front of the amplifier being the most musically important face. Therefore, the front of the amp was captured with an "inverse-ORTF" that, when reproduced in XR, becomes an ORTF virtual source.

2.1 Unity Recreation

For this experiment, the Steam Audio Package was used as the spatializer and Ambisonics decoding plug-in. In the Unity 3D engine, the audio was reproduced in order to compare the two room systems to each other as well as an additional simulated room (also created with the Steam Audio Package) for reference. A user interface was designed to let the listener switch between the three systems and move about the scene freely. After the audio objects were positioned, their maximum distances and amplitude/frequency roll-offs were adjusted to fit the size of the room in Unity.

For the Twins-Hamasaki array, each Sennheiser MKH800 Twin microphone produces two signals: one from each capsule. The system was arranged so that one capsule faced a wall *(Signal 1)* and the other towards the center of the room *(Signal 2)*. *Signal 1* was assigned to a cardioid object and placed on a virtual wall running along the outer perimeter of the room. *Signal 2* was assigned to a similar object and placed in the center of the virtual room, creating an inverted figure-8 pattern out of two cardioid virtual audio sources. The relationship between these positions is shown in Figure 3 (L).

The Ambeo-Hamasaki system was reproduced in Unity using the Steam Audio Ambisonic Audio Object. This object takes an Ambisonics A-Format signal and spatializes it around the center of the source, with the reflective audio placement and correct angles automatically taken into consideration. The four Ambisonics objects created from the array were placed in their respective positions on the Hamasaki Square and expanded until enough overlap was obtained to exclude any blind spots.

As a point of comparison, a virtual room was made to be a similar size as the other two with the walls made out of similar materials. The package uses ray tracing to simulate the multitude of delays bouncing off the walls and towards the listeners head. Each instrument was spatialized in Unity using cardioid virtual audio sources shown in Figure 2.

2.2 Preliminary Subjective Testing

A preliminary subjective test was conducted to compare each system based on a set of subjective attributes: envelopment, environmental width, naturalness, presence,

Perceptual Attributes Tested		
Attribute	Definition	Questions
Envelopment	Fullness of the surrounding sound image [4]	• How enveloping is the audio? • Is it all around or is it limited to a specific position (i.e. frontal area)?
Environmental Width	Broadness of the (reflective) environment within which individual sources are located [5]	• Which system do you feel has a larger sense of space? • Is it all around or is it limited to a specific position (i.e. frontal area)?
Naturalness	The similarity of the reproduced sound to its real/original state [5]	• How realistic is the audio? • How realistic are the sound sources? • Is the audio free from degrading artefacts?
Presence	The sense of being inside an (enclosed) space or scene [5]	• Do you feel as though you are within an enclosed space? • Do you feel present within this space or absent from the space? • Are you able to distinguish a boundary (or boundaries) around you?
Preference	The listener's personal satisfaction and/or the "pleasantness" of the recording [5]	• Which recording do you find aesthetically pleasing? • Which recording do you enjoy listening to the most?

Table 1: Preliminary testing subjects were given definitions for each perceptual attribute they were asked to evaluate. Attributes, definitions, and questions are listed above.

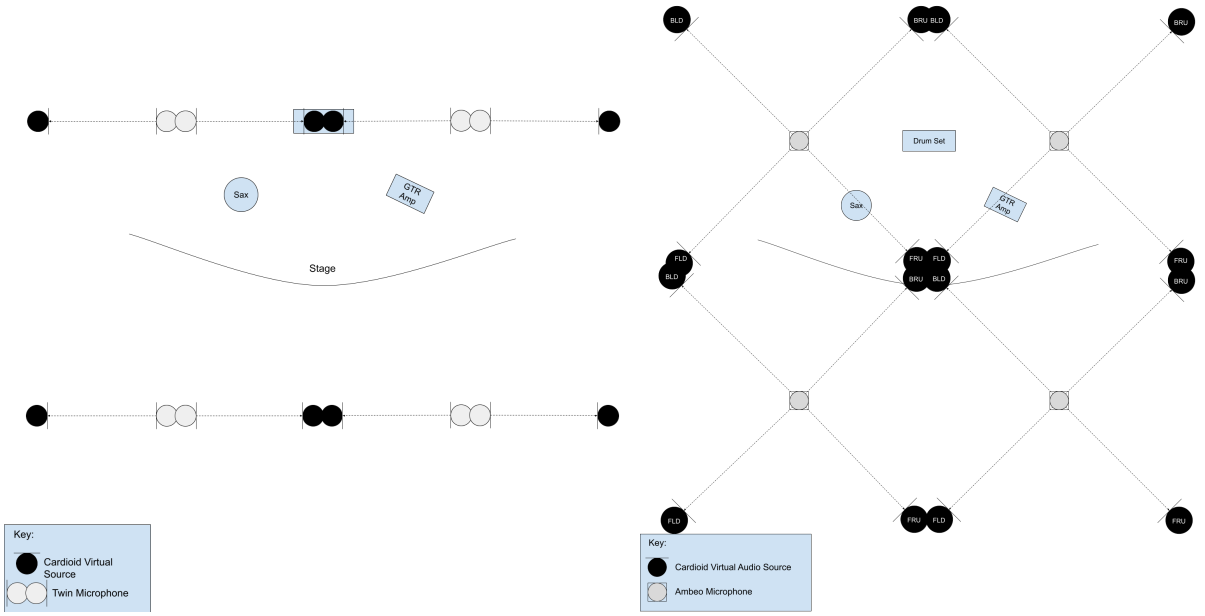

Fig. 3: (L) Diagram showing the relationship between physical Sennheiser MKH800 capture systems and their reproduced virtual cardioid objects.
(R) Diagram showing the relationship between physical Sennheiser Ambeo positions and the virtual first-order Ambisonics objects reproduced..

and preference. Since this testing was preliminary, it was conducted only on a small number of expert subjects who had at least five years of experience in music technology, immersive audio, and/or audio engineering. A total of eight subjects participated in the testing.

During testing, each subject was asked to wear the HTC Vive VR headset and use its controllers to move around the VR environment. When comfortable with the system, they were able to initiate the recording and begin exploring the virtual area. Each ambient playback environment was represented as a Room or *System* in Unity. Room A (System A) reproduced the Ambeo-Hamasaki listening environment, Room B (System B) reproduced the simulated listening environment created with the Steam Audio Package, and Room C (System C) represented the Twins-Hamasaki listening environment.Subjects were free to switch between each environment at will. It should be noted that the visual scenery remained static for each Room. Subjects are not aware of what array was associated with each room. There was no time limit to how long the subject cound remain in the virtual listening environment.

When they were satisfied with their virtual experience, subjects were asked to rank each System in terms of each of the attributes (e.g. from most natural to least natural). Each attribute was defined for the subjects and accompanied by guiding questions to help the subject understand what aspects of the auditory scene they should be analyzing. A list of attributes, definitions, and questions can be found in Table 1.

3 Results and Analysis

Collected results can be found in Figure 4 with *X* representing each system and *Y* representing the ranking of each attribute.

Although System C ranked higher, subjects considered the envelopment of System A (Ambeo-Hamasaki) and System C (Twins-Hamasaki) almost equal. System B, as expected, ranked lowest. Commentary from subjects noted that the reverberation in System B sounded

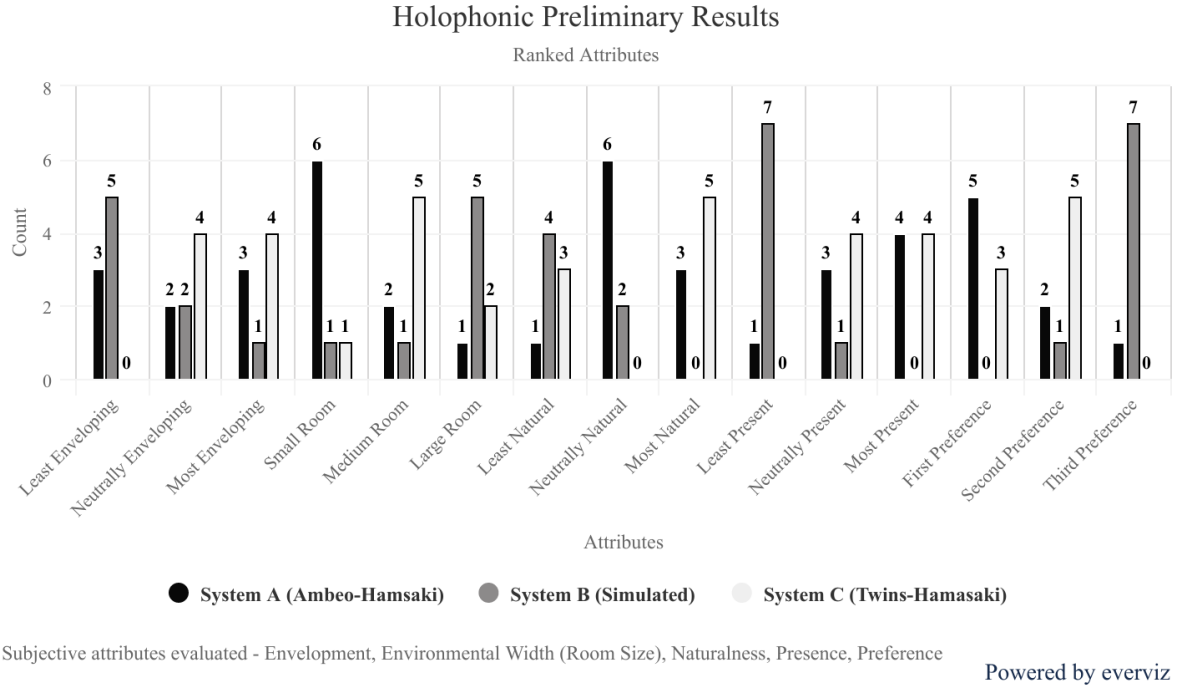

Fig. 4: Preliminary results in a comparison between the Ambeo-Hamasaki array (System A), Twins-Hamasaki arrary (Systems C), and the simulated virtual room as a reference (System B)

diffused which distorted their ability to localize the instruments' direct sound. It is interesting to note that the majority of subjects associated the level of localization with the level of envelopment for each system. Most subjects felt that System A and C had equal envelopment, however, System C performed slightly better in localization of the instruments.

Although listeners felt it lacking in envelopment, System B (the simulated room) ranked highest in environmental width. Subjects felt that widespread reflections caused the instruments to sound stretched out and distant from each other, which had the effect of widening the room. System A was mainly ranked as the smallest room because the direct sound sources felt closer to the subject. Even if the subject stood at the back of the performance area, farthest away from the instruments, their direct sound still felt upfront with less perception of a distance that felt visually accurate to where the instruments were located. For naturalness, System C performed the best in terms of representing the theatre space. Subjects tend to comment that the reverberation sounded natural and acoustically realistic to a large performance space. Some incorporated their commentary that System C performed well in envelopment and environmental width, therefore, influencing the naturalness of the recording. System A was ranked as the second most natural, with listeners citing the clarity of reproduction of each instrument being the most similar to a high-quality stereo recording. This is opposed to the commentary of those that ranked System A lower as it sounded like a stereo recording and did not accurately reproduce an actual acoustic space. The stereo-like quality was due to the close perception of the instruments (low environmental width). For System B, all subjects commented that it sounded simulated and noted it as having minimum to no naturalness.

Due to System B ranking as the largest room, the decay of the reflections sounded long and, therefore, subjects felt that they could not distinguish a boundary or a "stop" to the reflection. This caused the majority of subjects to rank System B as having the least presence. System A and System C were equally ranked with comments revolving around the subjects feeling "encapsulated." Subjects felt that if they could distinguish a boundary, they could perceive a room. Hence, a correlation between presence and environmental width. A majority of individual comments from subjects expressed that they felt the presence in System A and System C were the same with little to no difference.

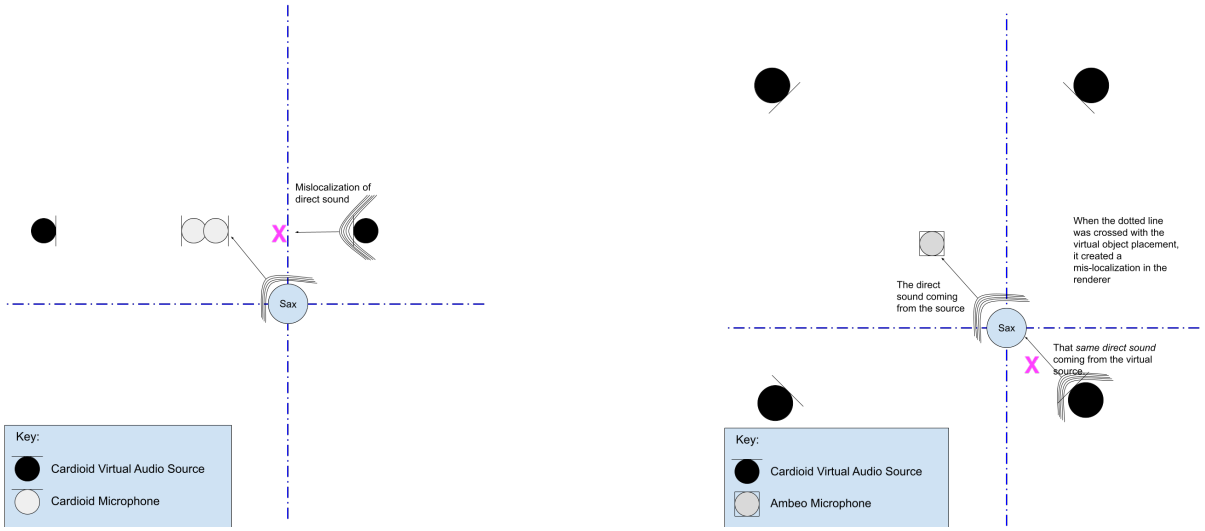

Fig. 5: (L) Diagram showing a mis-localization issue in the XR reproduction of the Twins-Hamasaki system. (R) Diagram showing a mis-localization issue in the XR reproduction of the Ambeo-Hamasaki system.

The split seemed to be influenced by the stability of the reflections based on movement. Interestingly, it was mentioned that, at times, both Systems A and C had moments when a subject was standing still, the reflections sounded defined and stable. But, when the subject moved, this stability was broken and room reflections became less defined. In other words, some reflections of the direct sound source sounded "misplaced." Further explanation of why this have occurred will be discussed in the Conclusion.

With preference, subjects tend to base their ranking on the naturalness of the recordings with some referring to other attributes such as envelopment or presence. System A was the most preferred system in that subjects found it the most aesthetically pleasing and of high-quality, similar to a stereo recording. Some noted System A as having the most clarity of the instruments, a balanced frequency response, and good timbre. However, every subject that chose System A as their first preference mentioned that although they liked System A the most, it did not necessarily represent the actual acoustic space accurately. Some of the subjects expressed that a real-life acoustic space would not sound as clean or clear as it did in System A, which had a quality similar to a stereo recording. Referring to their prior experiences listening to live performances, subjects felt that a real-life performance in an acoustic space, such as a theatre, should have some resonance, imbalanced frequencies, and less definition. Hence, subjects leaned toward System C being the superior system that accurately represented the acoustic space, although it may not have sounded as aesthetically great. Those that chose System A as their first preference mainly chose System C as their second preference for this reason. Also for this same reason, some subjects chose System C as their first preference. System B was the last preference for almost every subject due to the low quality sound of the ambience and the lack of naturalness/realism of the recording that did not represent the acoustic space.

4 Conclusion

Based on these preliminary results, the Twins-Hamasaki capture system and the Ambeo-Hamasaki system performed similarly among listeners, with a slight favoring of the Twins-Hamasaki system, except for the preference attribute, which is based on various factors that determined the listener's aesthetic satisfaction. Certainly, a mixture of both types of systems for XR reproduction would be advantageous in creating a complete virtual experience. Improvements of both systems can be made, especially in increasing the stability of the reflections as first noted in the analysis section. This slight instability caused the subjects to feel as though there was some mis-localization in the sound sources.

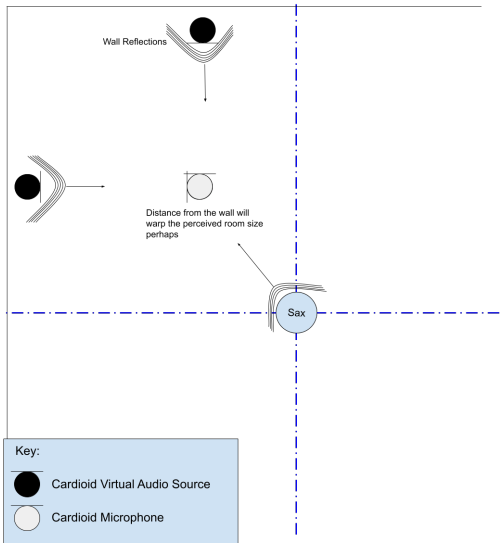

Fig. 6: A proposal to fix sound source mis-localization issues by limiting the placement positions of virtual microphones

In fact, one of the most significant findings from this experiment was regarding mis-localization for listeners when they stood in certain spots in the scene. Only two of the expert listeners tested were able hear that sound sources were shifted in position, with others referring to something "feeling-off" about the source localization. After further testing, was discovered that this was due to an error system design. This finding presents a fundamental starting point for designing holophonic ambience capture microphone arrays. Figure 5 depicts the nature of the issue. When the listener decided to stand near the indicated “X” postion, they heard the saxophone’s direct sound coming from an area where the instrument was not physically present. The dotted lines in the images represent invisible lines that a virtual audio source cannot cross. If they are crossed, mis-localizations will occur. A solution to this issue is to design arrays that capture no direct sound from the instruments. If direct sound is captured by an ambience array, the virtual source for that object can not be placed across the dotted line. Figure 6 presents a potential solution to this problem. This fix will potentially greatly lower the amount of channels required to capture directional room information.

Due to the errors in localization, the Hamasaki arrays implemented in this fashion are not recommended for capturing directional room information. Since Ambisonic microphones pick up the most direct sound, they may not be suitable for live capture unless their pickup patterns are steered to capture only diffuse sound. Instead, a wavefield approach might be the best way to capture directional room information with a minimal number of channels required. Future work could also involve an exploration in height perception by incorporating additional methods for height anchoring or improving the vertical image.

From this preliminary test, it is shown that a live capture of a performance that includes directional spatial room information in a reverberant room is more realistic than a simulated room. The techniques used for holophonic capture are able to recreate an acoustic space for XR environments with nearly accurate spatial perception. Further research and subjective testing is likely to improve the spatial accuracy of these systems.

References

[1] Matsakis, M., Songmuang, P., and Geluso, P., “Multi-Directional Radiation Characteristic Recording Methods and Reproduction in an XR Environment,” in *Audio Engineering Society Convention 151*, Audio Engineering Society, 2021.

[2] Hamasaki, K. and Hiyama, K., “Reproducing spatial impression with multichannel audio,” in *Audio Engineering Society Conference: 24th International Conference: Multichannel Audio, The New Reality*, Audio Engineering Society, 2003.

[3] Llewellyn, G. and Paterson, J., “Augmented and Mixed Realities,” *3D Audio*, p. 43, 2021.

[4] Morimoto, M., “The role of rear loudspeakers in spatial impression,” in *Audio Engineering Society Convention 103*, Audio Engineering Society, 1997.

[5] Rumsey, F., “Spatial quality evaluation for reproduced sound: Terminology, meaning, and a scene-based paradigm,” *Journal of the Audio Engineering Society*, 50(9), pp. 651–666, 2002.