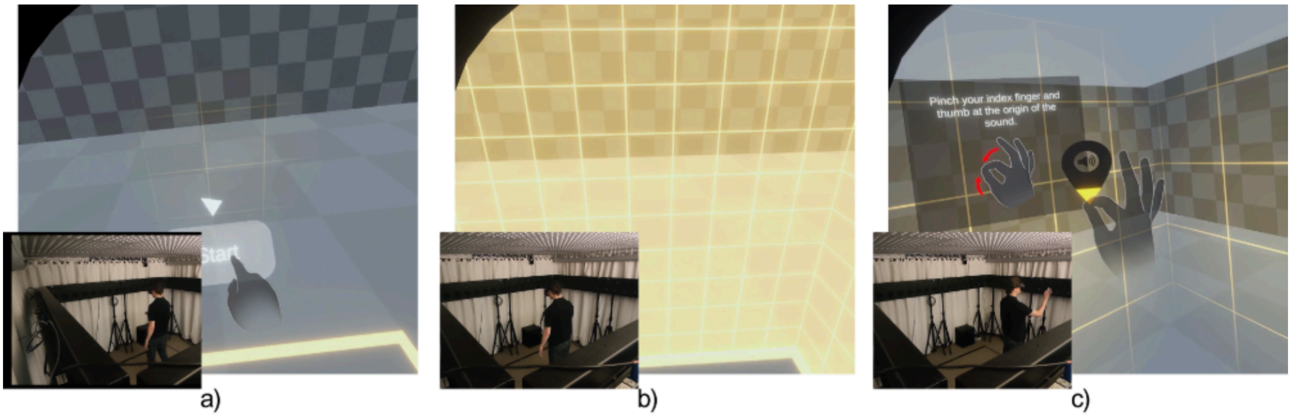
!nteractives

Preprint | 1822.1921

Assessing the Viability of Wave Field Synthesis in VR-Based Cognitive Research

Benjamin Kahl
benjaminkahl19@gmail.com
Max Planck Institute for Human Development


Start
a)
b)
Pinch your index finger and thumb at the origin of the sound.
c)

Abstract: This paper explores Wave Field Synthesis (WFS) as a method to enhance auditory immersion in VR-based cognitive research. While VR environments typically underuse auditory cues, WFS offers realistic, spatially accurate soundscapes. An experiment compared participants' localization accuracy of static and moving sounds between WFS and conventional stereo setups, examining the impact of environment, sound type, and duration. Results showed higher accuracy with stereo but a more natural auditory experience with WFS, especially regarding directional cues. Despite current limitations like poor height localization and user-dependent optimization issues, WFS demonstrates significant potential for specialized auditory perception studies involving complex directional information.

Tags: HCI, Sound, Audio Processing

1 Introduction

In recent years, the use of virtual reality (VR) in cognitive and behavioral science has grown significantly *(Zhu et al., 2021)*, with an increasing number of publications highlighting its advantages in terms of ecological validity, reproducibility, and experimental flexibility.

VR allows researchers to create controlled yet dynamic environments, making it an attractive tool for studying human perception, decision-making, and behavior *(Faria et al., 2023)*. However, a recurring challenge is determining how well findings from VR-based studies generalize to real-world settings. The term of *ecological validity* has thus become recurring in the psychology-vr space, and researchers hope to see an increase in immersion by integrating other senses into their virtual experiments. While most studies focus heavily on visual stimuli, auditory cues are often underutilized despite their critical role in perception and spatial awareness *(Isak de Villiers Bosman & Hamari, 2024)*.

The ability of participants to move freely within VR environments presents a unique opportunity to leverage spatially accurate soundscapes, such as those generated through artificial wavefield synthesis, to create a more immersive and realistic experimental setting.

Wave-Field Synthesis

Wave field synthesis (WFS) is an audio rendering technique for simulating soundwaves conceived by Berkhout in 1993 *(Berkhout et al., 1993)* and refined by Brandenburg et al. in 2009 *(Brandenburg et al., 2009)*. As illustrated in fig. Figure 1, loudspeaker arrays synthesize artificial wavefronts that give the illusion of originating from a given start-point by super-imposing a large set of individual, elementary sound-waves.

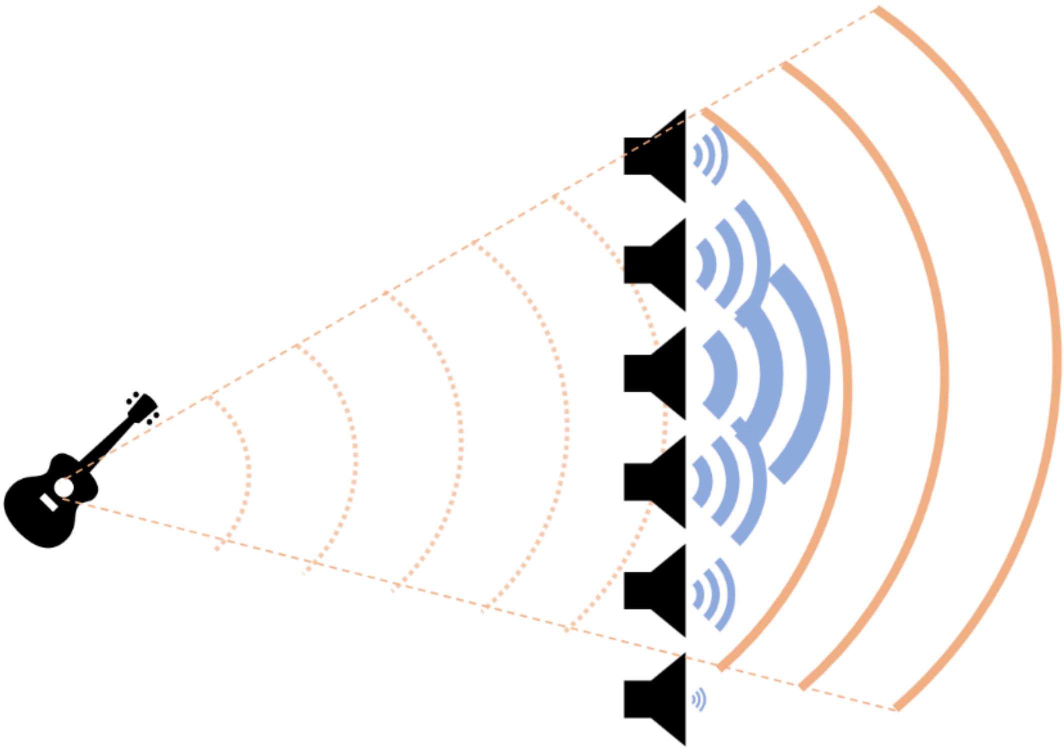


Figure 1: The concept behind wave field synthesis: Each speaker produces an elementary sound-wave which, in conjunction, produce a combined soundwave that approximates the sound of an arbitrary emitter.

These systems allow for the precise placement of sound-sources and allow listeners to experience the artificial soundscapes through their natural sense of hearing, thus providing an unprecedented amount of auditory immersion.

Due to their high cost, complexity and sensitivity to room acoustics, WFS systems are hitherto seldom used and have not yet reached market maturity. This novelty and limited accessibility opens up a new frontier of research thus far untrodden in the field of psychology.

Unlike commonly available, software-based spatial audio solutions such as *Steam Audio* or *Dolby Atmos*, WFS-based systems do not rely on HRTFs. As such, WFS systems are agnostic to the listener and circumvent the inter-subject variability that inhibit the reliability of HRTF-based systems *(So et al., 2010)* in cognitive and behavioral studies.

Goals

This project aims to evaluate the viability of using Wavefield Synthesis (WFS) for VR-based studies, identifying its challenges and limitations, as well as their potential solutions.

To achieve this, we will implement and conduct a sample study that examines the accuracy of perceiving sound origins. Participants will be tasked with locating sound sources in both a WFS-rendered environment and a conventional stereo-headphone setup, allowing for a direct comparison of localization performance.

Additionally, VR will be leveraged to investigate how different virtual environments, sound types, and other experimental parameters influence both accuracy and search behavior.

All in all, we hope to produce a comprehensive evaluation of both the viability of WFS in psychology, as well as an early assessment for the perceived accuracy of sound localization such systems provide when compared to more traditional approaches.

2 Technology and Setup

2.1 WFS Setup

The Wavefield Synthesis system used throughout this project (shown in Figure 2) was developed, installed, and maintained by Fraunhofer IDMT for use at the Max Planck Institute for Human Development (MPIB). The setup consisted of a sound-isolated room housing a dedicated sound-rendering PC and a square-shaped WFS speaker array. The system featured four linear arrays of 16 speakers each, forming a square configuration capable of synthesizing artificial sound fields within a 2×2 meter interaction zone. All speakers were positioned at head height, meaning that virtual sound sources could only be placed within the horizontal plane parallel to the floor. Additionally, three subwoofers on the ground provided complementary low-frequency support.

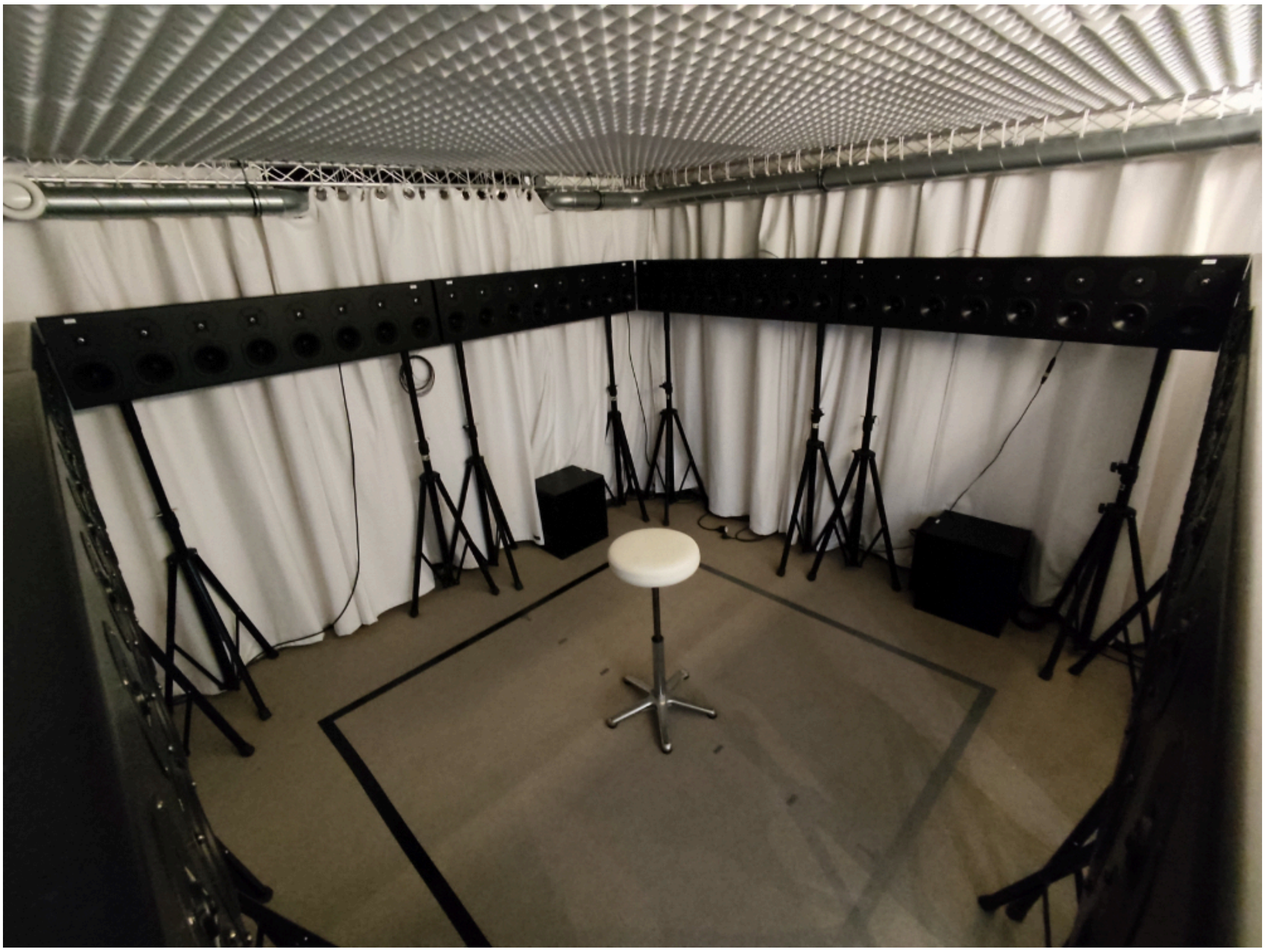


Figure 2: The employed WFS setup at the MPIB. The area marked in black tape corresponds to the WFS-area the participant can move around in.

Audio playback was facilitated via a USB connection to the rendering PC, utilizing the appropriate driver setup through MADIFace. Most fixed system parameters were configured in advance using Fraunhofer's SpatialAudio web interface (see Figure 3), while real-time control over sound source positioning was achieved by sending UDP packets containing OSC messages to the rendering PC (see Section ? for further details). This setup ensured precise spatial audio rendering and allowed for dynamic adaptation of sound source locations during the experiment.

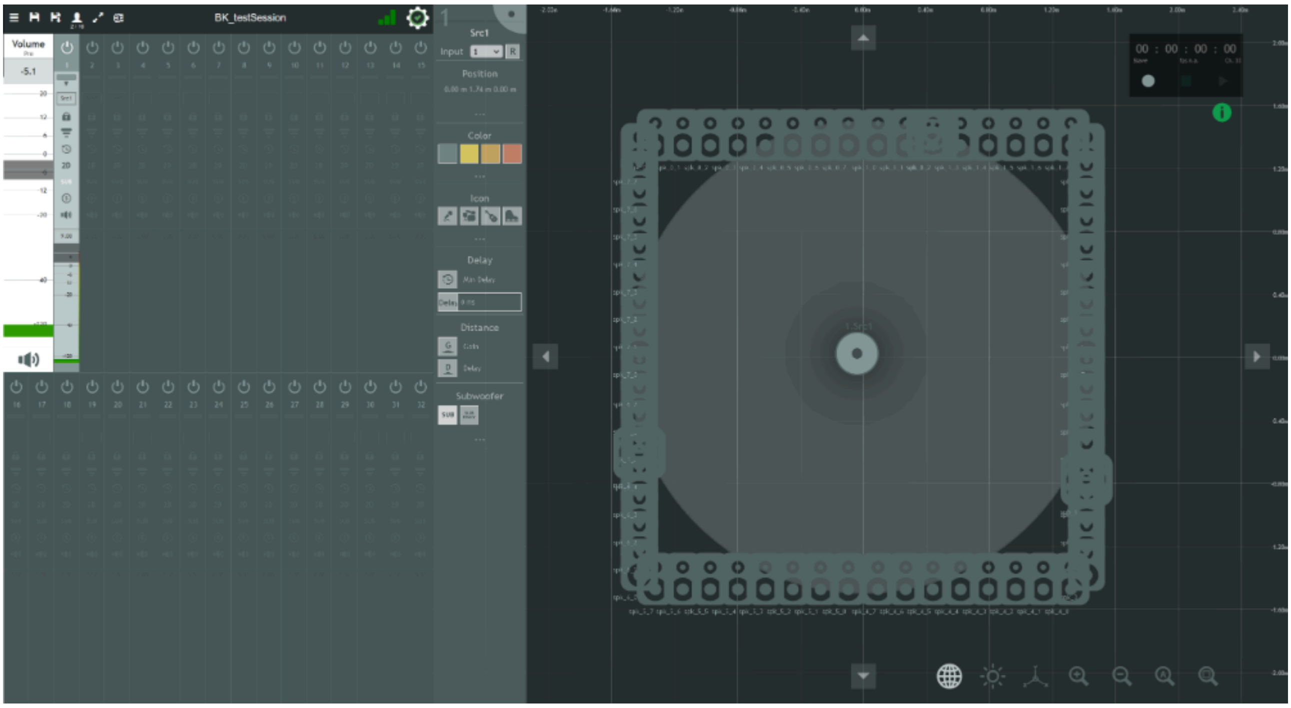


Figure 3: Fraunhofer's SpatialAudio web-interface allowing the WFS system to be configured. We used a single point-like sound source and set it's position by sending OSC command messages to the render PC from out Unity applications.

2.2 VR Setup

We opted to use a Meta Quest Pro as our VR headset, due to its wireless capabilities and built-in hand-tracking, minimizing physical constraints and allowing participants to move freely within the virtual environment and soundscape. Rather than developing a standalone application to run directly on the headset, we utilized Meta's Air Link feature to stream the VR experience from a PC to the headset over a high-speed wireless connection.

To ensure a stable and low-latency VR stream, we deployed a 5GHz Wi-Fi 6 router inside the WFS room, positioned in direct view and close proximity to the headset. The Unity application powering the VR experience ran on a Lenovo Legion 5 Pro equipped with an Nvidia GeForce RTX 3070 Ti (Laptop). This intermediary PC managed the VR rendering and streaming while also serving as the bridge to the WFS system. It was connected to the WFS rendering PC via both USB and a dedicated Ethernet connection, using USB for audio playback and Ethernet for real-time sound source positioning.

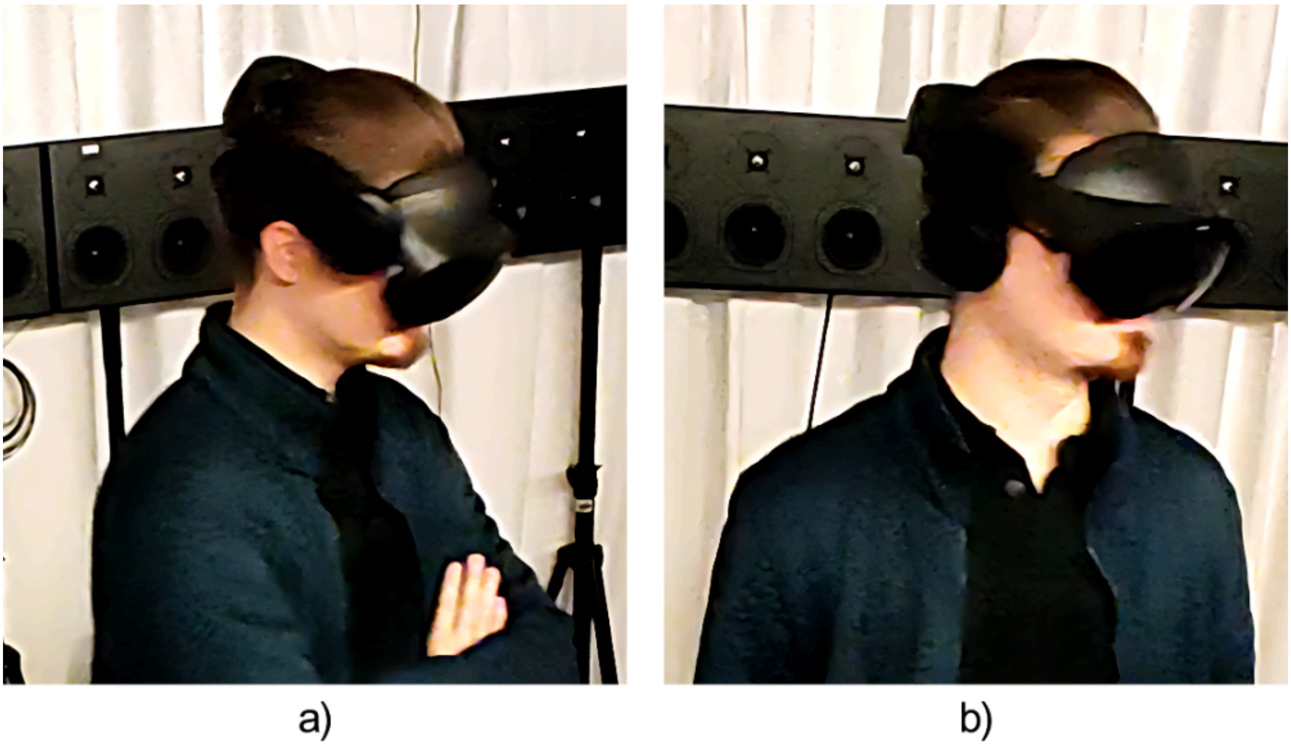


Figure 4: Participant wearing a Meta Quest Pro during WFS trials (a), and stereo trials (b). During WFS trials, the stereo headphones would be folded up and sound would come from the surrounding WFS speakers.

For the stereo headphone condition, we used a pair of Quest Pro-compatible attachable stereo headphones from Globular Cluster, ensuring consistent and high-quality audio output for comparison with the WFS-rendered soundscape. A participant wearing the headphone attachments can be seen in Figure 4.

2.3 Software Setup

The project was developed using Unity 2021.3.17, integrating the extOSC add-on to facilitate communication with the WFS renderer via its OSC interface. Most of the VR components were built using *ARC-VR*, an in-house framework developed at MPIB for VR-based research.

For experiments utilizing individual sound sources, extOSC provided a seamless and straightforward integration with the WFS system, allowing real-time control over sound positioning. However, handling multiple simultaneous audio sources may introduce additional complexity, which has not yet been tested or implemented in this study.

To switch between audio output devices (WFS and stereo headphones) dynamically within Unity, we utilized *SVCL*, a command-line tool by Nirsoft. Additionally, voice instructions for tutorial sessions and transition announcements between experiment blocks were generated using *ElevenLabs*. A custom Unity plugin was developed to allow direct generation of these voice files within the Unity environment.

Models and sound assets were sourced from the Unity Asset Store, CGTrader.com, and Freesound.org, ensuring a diverse selection of experimental stimuli. Data analysis was conducted using a custom Python framework tailored to the needs of the study.

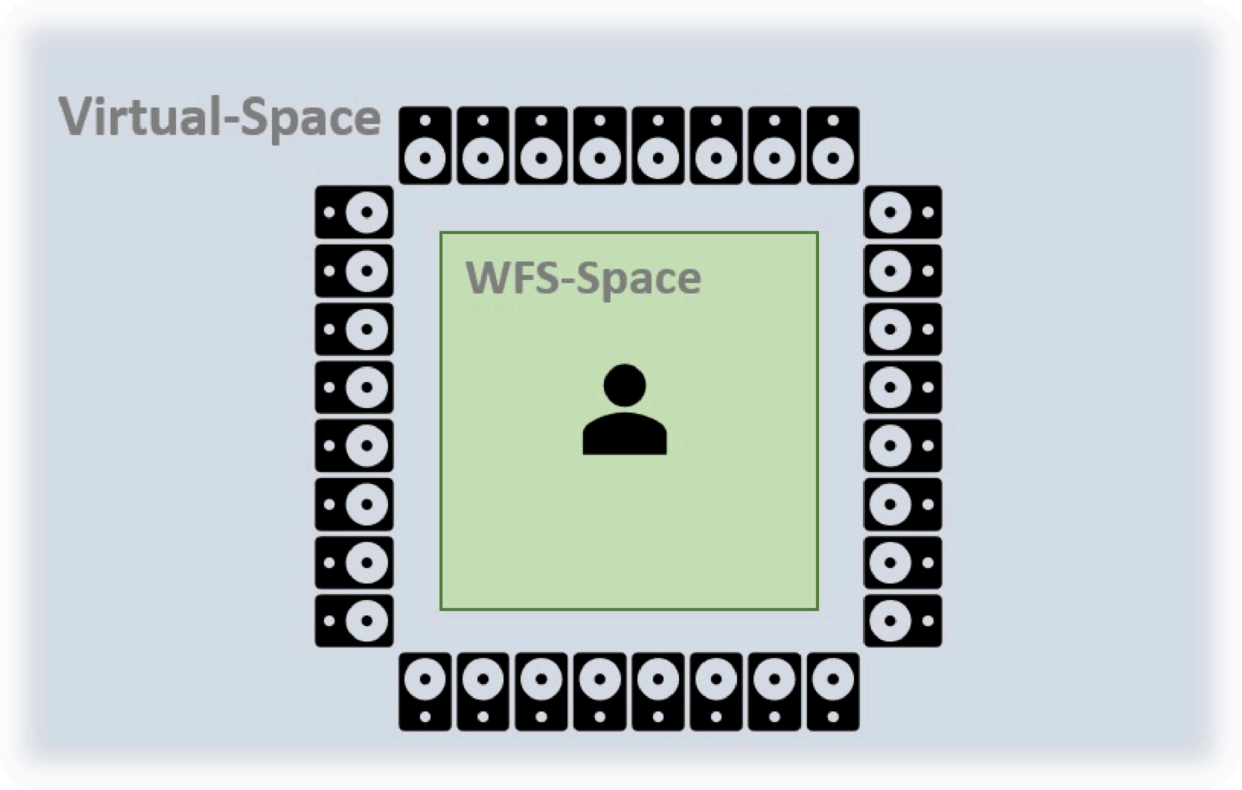


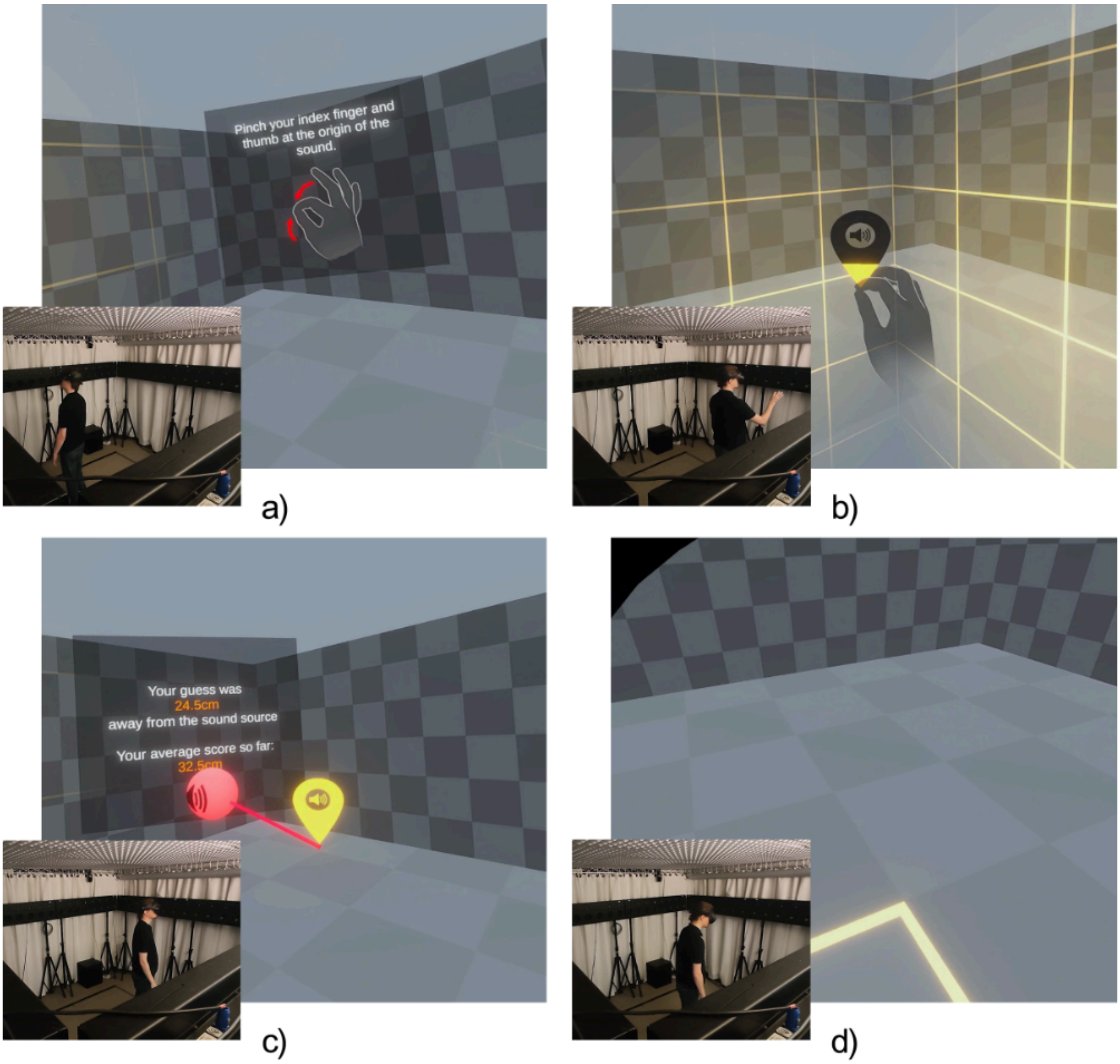
Figure 5: Difference between WFS-space and virtual-space: Whilst the participant could only move around and hear the artificial wave-fronts inside a 2x2m area, the size of the virtual environment (where virtual objects sound sources could be placed) extended far beyond.

3 Final Version

Based on observations and feedback extracted from the above listed prototypes, we built a version that was ultimately used in data collection and analysis. The version introduced some additional complexity and more conditions in the hopes of yielding more meaningful results.

3.1 Overview

Building on insights from previous prototypes, the final version of the experiment was largely based on the *Phone-Placement* prototype, but brought along some additional enhancements.



Pinch your index finger and thumb at the origin of the sound.

Your guess was 24.5cm away from the sound source

Your average score so far: 32.5cm

Figure 6: A single trial in the project's final version: The participant hears a sound (a), moves towards it and places a guess on it's origin by pinching their fingers for one second (b). The participant receives feedback by being shown the sound's real location (c). After a while, the indicators disappear and the next trial begins.

In this version, participants would still be asked to identify the perceived origin of sounds played within a 2x2m sound-rendering area at head height, but perform this task in varying surroundings and with varying sounds. Participants would indicate their guess by performing a *pinch* gesture at the perceived location using hand tracking. In cases where hand tracking was unreliable, regular VR controllers could be used as a fallback input method.

3.2 Experimental Conditions

To examine the impact of different auditory and environmental factors, we selected three distinct sound stimuli, each publicly available and varying in frequency and duration:

- **A rotary telephone ring** (6.12 seconds long, around 4 of which were highly audible)
- **A short piano melody** (6.861 seconds long)
- **A chirping bird in an outdoor environment** (2 minutes and 39.362 seconds long)

Participants would only hear the respective sound once per trial and get no repeat attempts. A localization guess could be placed before the sound finished playing.

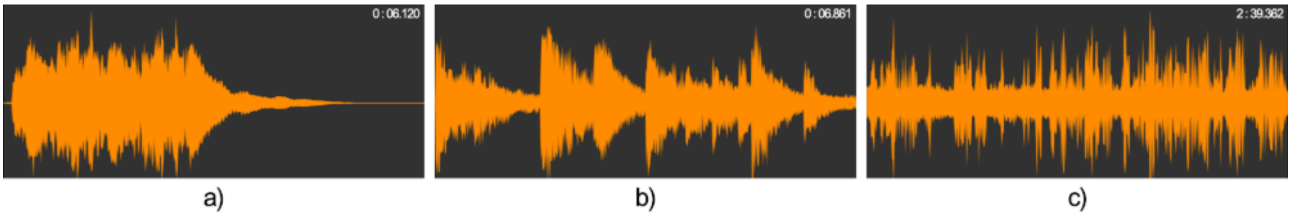
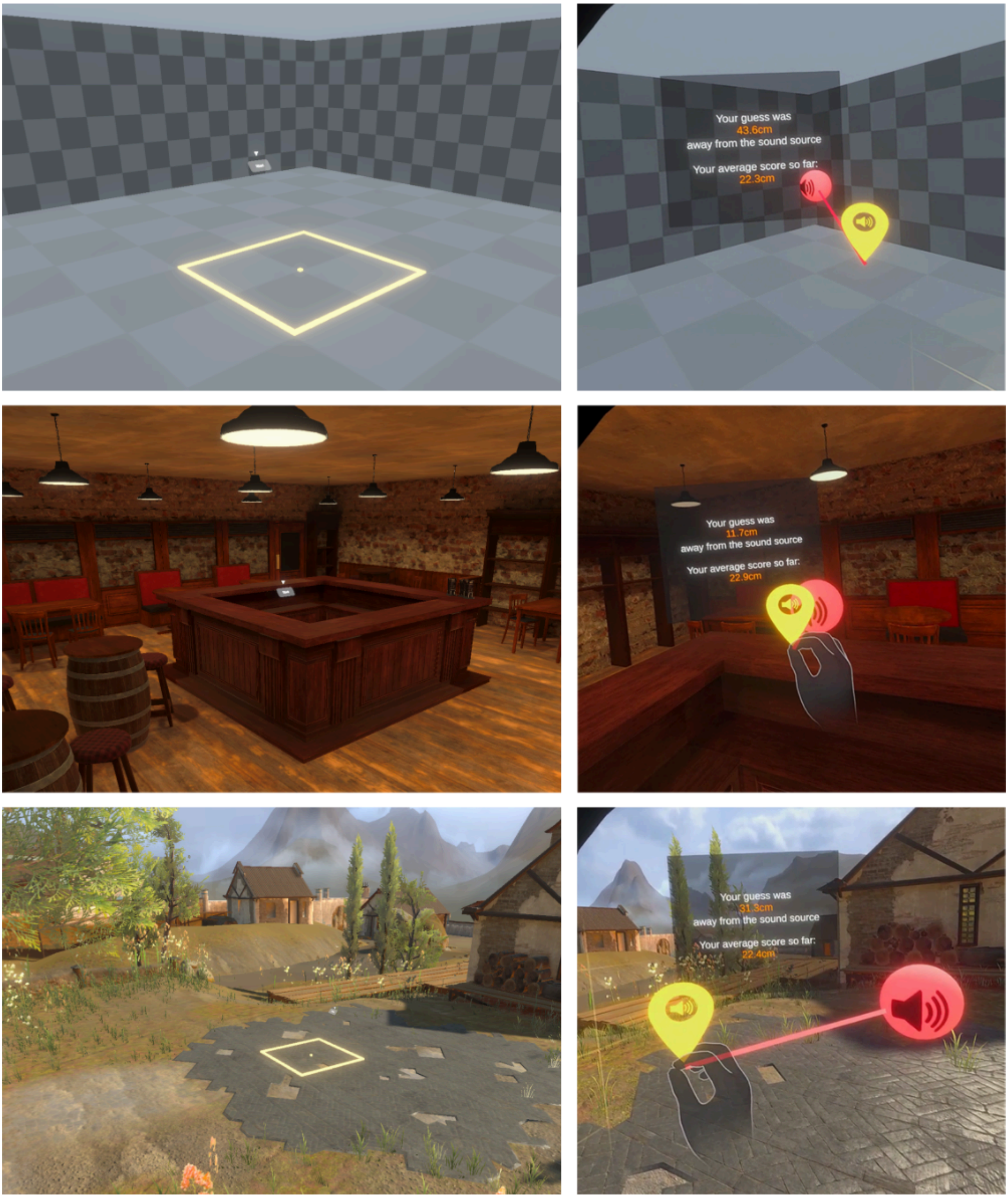


Figure 7: Waveform graphs and duration of the three different sounds used. With a) being the rotary telephone ring, b) the piano tune and c) the birdsong. The clearly audible part of a) lasted four seconds, b) six seconds, while c) lasted for over two minutes, thus usually ending prematurely when the participant placed a guess.

To explore whether the *virtual environment* itself influenced localization performance, three different virtual spaces were created to roughly match the nature of these sounds:

- **A featureless room with checkered walls** (corresponding to telephone ring)
- **An indoor bar environment** (corresponding to the piano melody)
- **A vast outdoor mountain village** (corresponding to the bird chirp)



Your guess was
43.6cm
away from the sound source
Your average score so far:
22.3cm

Your guess was
11.7cm
away from the sound source
Your average score so far:
22.9cm

Your guess was
31.3cm
away from the sound source
Your average score so far:
22.4cm

Figure 8: The three virtual environments that participants solved trials in as seen in the Unity Editor (left) and from a participant's perspective after placing a guess (right). The tutorial and dynamic trials were always conducted in the *blank*environment (top).

Each participant completed *six* trials in each of the three environments (two trials for each sound) in a randomized order. The sound location was chosen at random and would not move throughout the trial.

After completing these 18 *static trials*, participants performed an additional 9 *dynamic trials* (three for each sound), where the sound source moved along a 2-meter trajectory over 1--3 seconds before stopping. These dynamic trials always took place in the neutral, blank environment to isolate movement-related effects.

This resulted in a total of 54 trials per participant, with an average completion time of approximately 25-35 minutes.

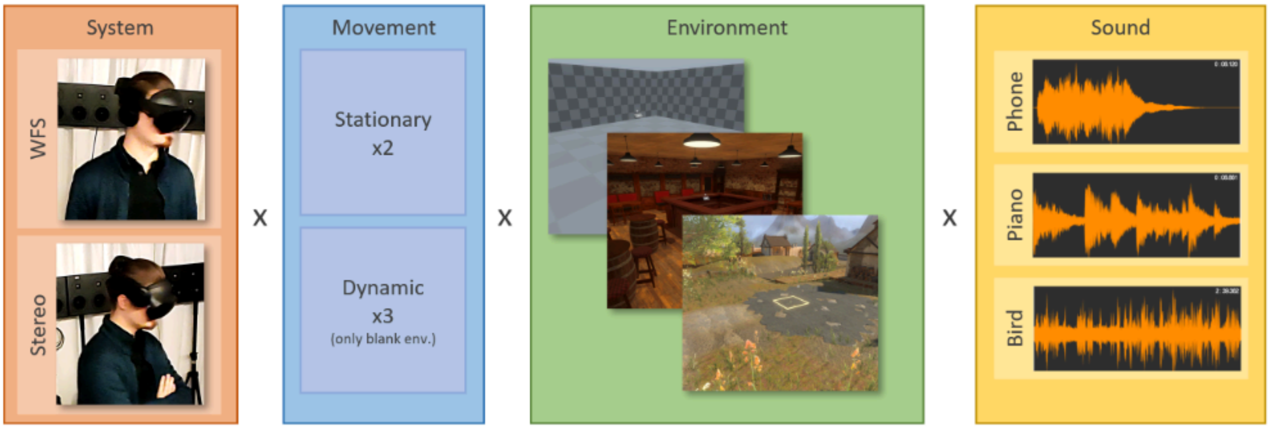


Figure 9: Overview of the individual parameters, producing a total of 24 unique permutations (conditions). Each stationary condition would be performed twice, and each dynamic one thrice by each participant. Trials would be randomized within each block.

3.3 Research Questions

Beyond evaluating the accuracy of WFS vs. stereo playback, the experiment was designed to explore additional research questions, including:

- **Impact of the virtual environment**: Does the visual environment affect sound localization performance?
- **Impact of sound type & duration**: Do different sound types and durations influence accuracy?
- **Environment-sound synergy**: Does performance improve when a sound matches its expected environment (e.g., a piano melody in a bar)?
- **Moving sound sources**: How does localization accuracy change when sound sources are moving, and do these effects differ between WFS and stereo?

By systematically varying these factors, the experiment aimed to provide a comprehensive assessment of the viability of WFS for VR-based auditory research.

3.4 Tutorial

Before starting the experiment, participants were guided through an interactive tutorial designed to familiarize them with the trial procedure and the method for placing guesses.

The tutorial was narrated using AI-generated voice lines (produced with ElevenLabs and provided step-by-step instructions. It introduced participants to the experiment's mechanics, including:

- How to indicate the perceived location of a sound using hand tracking (or controllers, if needed).
- How trials would be structured, including the distinction between static and moving sounds.
- Safety and comfort disclosures, such as showing the virtual barrier participants could not pass through, or explaining that the experiment could be paused or aborted at any time in the case of motion sickness or other reasons.

Following the tutorial, participants completed four test trials, featuring one with explicit instructions, and another three for each sound type. These test trials allowed them to get accustomed to the controls and the process of submitting guesses.

The tutorial always took place in the neutral checkered environment using the WFS system to maintain consistency. Additional audio announcements and instructions were played at key moments, such as switching between virtual environments, transitioning between WFS and Stereo playback or beginning the moving-sound trials.

3.5 Spatial Calibration

As described in privious section, a major challenge in using a mobile VR headset was ensuring that the virtual space and the WFS space were perfectly aligned. Since small discrepancies could significantly impact the accuracy of spatial sound localization, we introduced a calibration step that was performed by the experimenter before each session, prior to the participant putting on the headset.

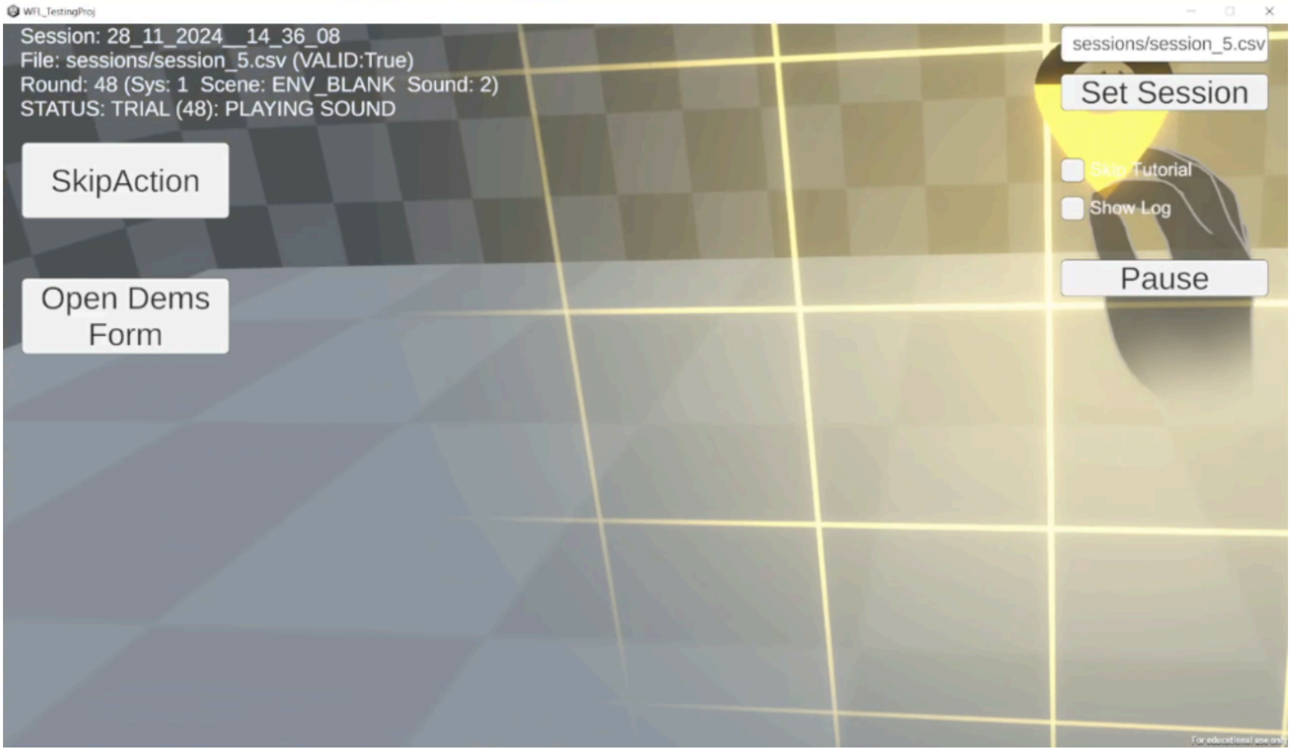
To streamline this process, our Unity program leveraged the Quest Pro's passthrough functionality. At the experiment's starting screen, a specific button combination on the VR controllers would activate calibration mode, temporarily removing the virtual environment and replacing it with the passthrough image of the real world. Over this image, a virtual overlay displayed the boundary and center of the WFS area.

Using the VR controllers, the experimenter could manually align the virtual and real-world WFS zones. The right controller functioned as a laser pointer, allowing the experimenter to indicate the "true" center of the WFS space. Pressing the trigger button would then reposition the virtual zone accordingly. To finalize the alignment, the experimenter could use the left controller's analog stick to rotate the virtual space around this center point until its boundaries matched those of the real WFS area.

This process ensured that the spatial mapping of sounds in VR accurately corresponded to their real-world WFS locations, although we expect that some degree of error remained. The alignment error would only be present in WFS conditions, as stereo playback didn't rely on an external system.

3.6 Experimenter Controls

Because we used an Air Link connection, the experimenter was able to continuously monitor the participant’s view through a mirrored display on the intermediary PC. In addition to this mirrored view, the experiment interface also displayed the current status of the experiment, including which trial was in progress, the sound being played, and the participant’s responses.



```
Session: 28_11_2024__14_36_08
File: sessions/session_5.csv (VALID:True)
Round: 48 (Sys: 1  Scene: ENV_BLANK  Sound: 2)
STATUS: TRIAL (48): PLAYING SOUND
```

sessions/session_5.csv

Figure 10: Screenshot of the application the experimenter could see during a session. The application shows a mirrored view of the participant's left eye, in addition to the current status and some controls.

To maintain control over the session, the experimenter had the ability to manually skip steps or pause the experiment if necessary. The system also included an automatic pause function that would activate if the participant removed the headset, ensuring that no trials continued without their awareness.

At the end of each session, participants were prompted to enter demographic information through a simple form. This form could be accessed and toggled directly from the experiment window, allowing for a smooth transition from the experimental phase to data collection.

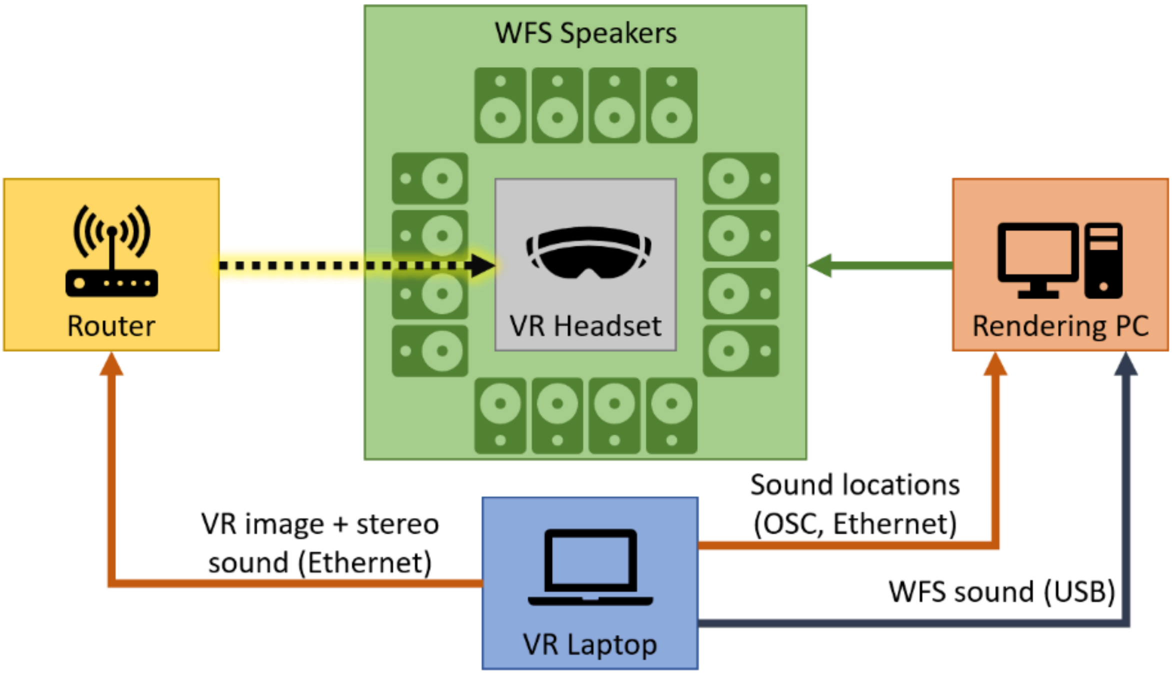

WFS Speakers
Router
VR Headset
Rendering PC
VR image + stereo sound (Ethernet)
Sound locations (OSC, Ethernet)
WFS sound (USB)
VR Laptop

Figure 11: Hardware setup of our experiment. An intermediary VR laptop would render the VR image and send it to the HMD via a router while also managing sound playback of the WFS system via the WFS render PC.

3.7 Participants and Data Collection

The experiment was conducted with six participants (four male, two female) aged between 27 and 40. All participants were recruited internally at the MPIB as volunteers. Most of them reported having either a "regular" or "enthusiast" level of prior VR experience.

Due to technical difficulties with switching between audio devices, the majority of participants (four out of six) began the experiment using the WFS system before transitioning to the stereo condition in the second half of the session.

During each session, we recorded the positions and timestamps of both the played sounds and the participants' guessed locations. Additionally, continuous tracking data was collected for the position and rotation of the head-mounted display and both hands. Since the Quest Pro relies on its built-in cameras for hand tracking, reliable tracking data was not always available. In cases where tracking was lost, default values were assigned, which were later filtered out by our analysis scripts.

The results of data collection are examined and analyzed in the next chapter.

4 Conclusions

The use of WFS in VR-based cognitive and behavioral studies presents both unique challenges and distinct advantages.

While its implementation demands greater effort from developers, it offers a more natural and intuitive interface for participants, potentially enhancing ecological validity. When paired with a mobile headset and hand-tracking controls, WFS minimizes the need for participants to learn complex controls, such as button mappings or specific interaction techniques. Instead, they can rely on their natural body movements and sense of hearing, making the experience more accessible.

As the technology matures, we anticipate a reduction in development effort, but remain skeptical that the additional cost and endeavor is worthwhile for studies where the sense of hearing is secondary. A standardized API for sound positioning and playback would greatly simplify implementation, eliminating the need for OSC-based location setting and separate interfaces for sound playback.

From both the development process and data analysis, we identified several key observations:

- A stereo setup is more flexible, particularly for simulating effects like occlusion. While localization accuracy in stereo can surpass WFS, it requires adaptation to artificial attenuation patterns.
- Participants in stereo trials rely more on perceiving volume attenuation, whereas in WFS, they depend more on their natural binaural hearing.
- The accuracy gap between stereo and WFS is smaller for moving sound sources than for static ones.
- WFS, by using speakers and artificial wavefronts, provides a more natural-sounding soundscape and an intuitive interface. WFS is a clear winner in use-cases where direction matters more than exact location or distance, where as regular headphones/HRTFs are more adequate for use-cases where participants walks close and among the sound emitters.
- WFS systems without *user-dependent optimization* are less accurate than stereo systems when the sounds are *inside* the listening area due to physical limitations. We expect WFS localization to improve substantially by incorporating the user's position into the sound rendering process.
- Longer-duration sounds are easier to localize.
- The visual setting of the virtual environment has no measurable impact on performance.

This study was limited to isolated point-source sounds. We expect WFS to be more suitable for research involving complex soundscapes with multiple directions and sources, such as those studied in soundscape psychology and psychoacoustics. Other experiments utilizing WFS and VR have demonstrated satisfactory localization accuracies.

While we look forward to further advancements in this field, we anticipate that WFS will remain a specialized tool for research focused on auditory perception, rather than a mainstream addition for enhancing immersion in general VR studies.

References

Berkhout, A. J., Vries, D. D., & Vogel, P. (1993). Acoustic Control by Wave Field Synthesis. *The Journal of the Acoustical Society of America*, *93*, 2764–2778.

Brandenburg, K., Brix, S., & Sporer, T. (2009). Wave Field Synthesis. *2009 3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video*, 1–4. https://doi.org/10.1109/3DTV.2009.5069680

Faria, A. L., Latorre, J., Cameirão, M. S., i Badia, S. B., & Llorens, R. (2023). Ecologically valid virtual reality-based technologies for assessment and rehabilitation of acquired brain injury: a systematic review. *Frontiers in Psychology*, *14*, 1233346. https://doi.org/10.3389/fpsyg.2023.1233346

Isak de Villiers Bosman, K. J., Oğuz 'Oz' Buruk, & Hamari, J. (2024). The effect of audio on the experience in virtual reality: a scoping review. *Behaviour & Information Technology*, *43*(1), 165–199. https://doi.org/10.1080/0144929X.2022.2158371

So, R. H., Ngan, B., Horner, A., Braasch, J., Blauert, J., & Leung, K. L. (2010). Toward orthogonal non-individualised head-related transfer functions for forward and backward directional sound: cluster analysis and an experimental study. *Ergonomics*, *53*(6), 767–781. https://doi.org/10.1080/00140131003675117

Zhu, K., Lin, R., & Li, H. (2021). Study of virtual reality for mild cognitive impairment: A bibliometric analysis using CiteSpace. *International Journal of Nursing Sciences*, *9*. https://doi.org/10.1016/j.ijnss.2021.12.007