Presence vs. Precision: Spatial Audio Latency in Virtual Opera

Yijhe Li

| DEPARTMENT OF DESIGN SCIENCES
FACULTY OF ENGINEERING LTH | LUND UNIVERSITY
2026

MASTER THESIS

LUNDS UNIVERSITET

LTH
LUNDS TEKNISKA HÖGSKOLA

Presence vs. Precision: Spatial Audio Latency in Virtual Opera

Yijhe Li

LUND UNIVERSITY

Presence vs. Precision: Spatial Audio Latency in Virtual Opera

Published by
Department of Design Sciences
Faculty of Engineering LTH, Lund University
P.O. Box 118, SE-221 00 Lund, Sweden

Subject: Virtual Reality and Augmented Reality (MAMM15)
Division:
Supervisor: Jens Nirme
Examiner: Mattias Wallergård

Abstract

Audiovisual synchronization significantly impacts the viewing experience in streaming environments, yet research on latency sensitivity toward spatial audio in Virtual Reality (VR) remains limited. This study conducted an audiovisual synchronization experiment in a virtual opera house, where 18 participants performed multisensory integration tasks using the Meta Quest Pro. The results indicated a substantial increase in the Just Noticeable Difference (JND) for spatial audio, reaching 461.23 ms, compared to 224.14 ms under the stereo audio condition. Similarly, the Point of Subjective Simultaneity (PSS) shifted from 22.47 ms in the stereo condition to 71.17 ms in the spatial audio condition. Subjective questionnaire results revealed that participants perceived spatial audio as more natural and requiring less cognitive effort, despite the observed decrease in temporal precision. Eye-tracking data further showed that gaze anchors remained concentrated on the conductor's face, with no apparent differences in gaze distribution between audio conditions, indicating that auditory presentation did not fundamentally alter visual attention strategies. Overall, while the data showed a numerical trend toward decreased user temporal sensitivity in the spatial audio condition, this effect did not reach statistical significance. This study provides practical implications for the design of future immersive streaming systems and contributes to the development of latency thresholds that better balance user experience and system performance.

Keywords: Spatial audio, Audiovisual synchronization, Multisensory integration

Sammanfattning

Audiovisuell synkronisering har en betydande inverkan på tittarupplevelsen i streamingmiljöer, men forskningen kring latenskänslighet för rumsligt ljud (spatial audio) i Virtual Reality (VR) är fortfarande begränsad. I denna studie genomfördes ett experiment med audiovisuell synkronisering i ett virtuellt operahus, där 18 deltagare utförde multisensoriska integrationsuppgifter med ett Meta Quest Pro.

Resultaten visade en väsentlig ökning av den minsta märkbara skillnaden (JND) för rumsligt ljud, som nådde 461,23 ms jämfört med 224,14 ms vid stereoljud. På samma sätt skiftade punkten för subjektiv samtidighet (PSS) från 22,47 ms i stereoförhållandet till 71,17 ms i förhållandet med rumsligt ljud. Subjektiva frågeformulär avslöjade att deltagarna upplevde det rumsliga ljudet som mer naturligt och mindre kognitivt krävande, trots den observerade minskningen i tidsmässig precision.

Eye-tracking-data visade vidare att blickfokuset förblev koncentrerat på dirigentens ansikte, utan tydliga skillnader i blickfördelning mellan ljudförhållandena. Detta tyder på att ljudpresentationen inte i grunden förändrade de visuella uppmärksamhetsstrategierna. Sammantaget, även om data visade en numerisk trend mot minskad tidsmässig känslighet hos användarna vid rumsligt ljud, nådde denna effekt inte statistisk signifikans. Denna studie ger praktiska implikationer för design av framtida immersiva streamingsystem och bidrar till utvecklingen av latinströsklar som bättre balanserar användarupplevelse och systemprestanda.

Nyckelord: Rumsligt ljud, audiovisuell synkronisering, multisensorisk integration

Acknowledgments

First and foremost, I would like to express my sincere gratitude to my supervisor, Jens Nirme, for his guidance and support, as well as for providing the equipment necessary for this research.

Secondly, I gratefully acknowledge Lund University Humanities Lab and Malmö Opera for providing the space, equipment, and materials required for this study.

I would also like to extend my heartfelt thanks to everyone who participated in the experiment.

Finally, I am deeply grateful to my family and friends for their unwavering support throughout these amazing two years in Sweden.

Lund, Jun 2026

Yijhe Li

Table of contents

List of acronyms and abbreviations

AR	augmented reality
CI	confidence intervals
FOV	field of view
HMD	head-mounted display
HRTF	head-related transfer function
ILD	interaural level difference
ITD	interaural time difference
JND	just-noticeable difference
MTP	motion-to-photon
PSS	point of subjective simultaneity
SD	spectral differences
SDK	software development kit
TBW	temporal binding window
VR	virtual reality
XR	extended reality

1 Introduction

This chapter serves as an introduction to this thesis, including background, purpose, objective and research questions.

1.1 Background

The digital transformation of traditional opera houses has become increasingly urgent as audiences seek more immersive and interactive viewing experiences[1]. One of possible paths is virtual reality (VR), which offers a powerful solution by recreating the atmosphere of live performance while renewing and broadening audiences.

Spatial audio technology has proven highly effective in providing a sense of presence and has been widely adopted across diverse fields, including music production[2], gaming[3], telecommunications[4], remote training, education[5], and mental health[6]. This highlights both the versatility and the substantial potential of spatial audio for next-generation immersive streaming services. However, for events such as concerts that demand a high degree of acoustic coherence, there is still limited research on how latency affects the user experience.

This research aims to explore the perceptual thresholds and differences in sense of presence between spatial and stereo audio within a concert environment. The study evaluates quality of experience through metrics such as perceptual sensitivity to latency and spatial presence, while simultaneously analyzing visual attention to identify behavioral variations in gaze patterns during perceived technical delay.

1.2 Purpose

The purpose of this thesis is to determine the perceptual thresholds for both stereo and spatial audio in immersive environments and to investigate how audiovisual latency affects the user's sense of presence.

The findings aim to provide practical insights for improving audiovisual synchronization in virtual performances. Specifically, this research serves as a valuable reference for system developers to set appropriate latency thresholds, maintaining an optimal balance between system performance and user experience.

A possible application of the findings is to assist opera house in enhancing the quality of immersive systems.

1.2.1 Research questions

This research aims to answer the following questions:

1. How does audiovisual latency affect user immersion and behavior pattern when using spatial audio compared to stereo audio?
2. Are listeners more sensitive to audiovisual temporal offset in spatial audio compared to stereo audio?
3. Do users detect latency mainly through visual cues or auditory cues?

1.3 Sustainable Development Goals

The United Nations Sustainable Development Goals (SDGs) define 17 goals as part of the *2030 Agenda for Sustainable Development*. This work contributes to *SDG 4: Quality Education* by providing empirical evidence for the development of immersive digital learning environments. The findings suggest that spatialized audio can enhance the naturalness and presence of remote music education without increasing the learners' cognitive load.

Moreover, this research aligns with *SDG 9: Industry, Innovation and Infrastructure* by supporting technological advancement in the cultural and creative industries. By exploring the relationship between audio rendering and human perception, the study provides technical insights for building future smart performance venues. Specifically, it offers a framework for high-quality streaming that balances hardware performance with user experience. This focus on optimization ensures that immersive technologies can be deployed more efficiently, reducing the need for expensive, energy-intensive hardware while still delivering professional-grade results.

Finally, this work supports *SDG 11: Sustainable Cities and Communities* by promoting the preservation and accessibility of cultural heritage. Through high-precision 3D scanning and spatial audio rendering, iconic cultural venues like opera houses can be digitally preserved and shared with a global audience. This technology reduces the carbon footprint associated with international travel while fostering a sustainable ecosystem for the global arts community.

Figure 1.1 SDG Goals.

1.4 Scope

This study focuses on exploring the relationship between a conductor's body movements and the perception of spatial audio. The experiment uses three representative 3-second audio and corresponding video clips from a single musical piece as the main stimulus. During the trials, participants were watched with these clips at random temporal offsets. Their task was to judge whether the video or audio was leading under two different conditions: spatial audio and stereo audio.

To maximize the accuracy of participants' perception of changes in the sound field, the study is designed as an immersive experience with a fixed listening position. Participants are required to complete 162 repeated trials in a virtual opera house. This highly controlled approach was chosen to capture subtle perceptual differences while eliminating potential confounding variables, such as variations in musical style or excessive head movements. By focusing on a specific auditory and visual environment, the experiment ensures that the collected data accurately reflects the impact of audio rendering modes on latency perception.

1.5 Ethics

This study follows the ethics rules of Lund University[7]. Before starting, we explained the experiment to every participant, and they all signed an informed consent form. We also made it clear that they could stop or leave at any time for any reason, without any trouble.

Regarding data management and privacy protection, all collected eye-tracking data, behavioral responses, and subjective questionnaires were anonymized using identification codes to ensure that results cannot be traced back to individual identities. Digital files are stored on password-protected hardware. Additionally, rest breaks were provided during the experiment to prevent fatigue and ensure both data quality and participant comfort.

Specific video and audio materials used in this study are regulated by a Non-Disclosure Agreement. The researchers will not disclose any protected materials to respect the partners' intellectual property.

2 Theory

This chapter provides a comprehensive literature review of the current field from the challenge and the relevant technology in this research.

2.1 Virtual Reality

Virtual Reality (VR) is a rapidly evolving technology designed to immerse users in a fully digital environment, enabling interaction through high-fidelity, multi-sensory experiences[8]. By providing a highly controllable, repeatable, and safe simulated environment, the applications of VR have become extensive and interdisciplinary, spanning fields such as education[9], medicine[10], and arts and cultural heritage[1]. Furthermore, research indicates that VR-based training can effectively enhance users' cognitive functions, physical performance, and social well-being[11].

Depending on the hardware, VR can be classified into several types, including Head-mounted-display (HMD) VR, cave automatic virtual environment, Desktop VR, Workbench VR, and Telepresence. Among these, HMD-based VR is the most recognizable to the public. In recent years, several technology companies have released products such as the Apple Vision Pro, Meta Quest Pro, and Samsung Galaxy XR, reflecting industry mainly focus on HMD.

Some modern VR headsets are equipped with integrated eye-tracking and head-tracking sensors, which seamlessly capture user movements without interruption. These real-time data streams allow algorithms to optimize rendering processes[12], delivering high-fidelity visual and auditory experiences. Real time feedback creates a more realistic and deeply immersive virtual environment.

Latency in virtual environments not only induces sensory conflicts but also directly impairs users' sense of presence and task performance. Audiovisual synchronization represents a critical technical benchmark. Research indicates that the human perceptual tolerance threshold for audiovisual asynchrony typically falls within a range of 160–250 ms [13], [14], [15]. Once latency exceeds this window, cognitive load increases measurably and disrupts the accuracy in spatial localization[16]. Although general thresholds for audiovisual asynchrony have been established, the degree to which these thresholds vary across audio rendering conditions remains to be determined. Equally, the relative weight of visual and auditory cues in latency detection has not been systematically examined within immersive VR contexts.

2.2 Spatial Audio

As audio rendering transitions from fixed-channel outputs to immersive spatial simulations, the increased spatial precision of acoustic data likely modulates user sensitivity to system latency. Understanding the differences between these audio formats is essential for determining how specific auditory cues influence the detection of temporal offsets in immersive environments.

2.2.1 Channel-based Audio

The stereo format utilizes signal differences between channels to construct a horizontal soundstage for the human ear[17]. It allows the left and right ears to receive different audio signals; for example, when listening to music, a guitar might be heard in the left ear while the right ear perceives a base. In practice, signals are often mixed, with only minor differences existing between the two channels.

Surround sound (such as 5.1 or 7.1 channels) builds upon this foundation by adding physical rear and side channels, further enhancing the sense of envelopment in a two-dimensional space. Despite this improvement, users must remain within a specific physical area to achieve the optimal experience[18].

2.2.2 Ambisonics

Ambisonics is a full-sphere surround sound format that employs spherical harmonic decomposition to represent the sound field at a specific point in space[19]. By capturing the comprehensive directional characteristics of incident sound waves, this spatial encoding technology encompasses both a full 360° horizontal plane and vertical height information.

First-order ambisonics utilizes four channels (WXYZ) to define the acoustic pressure (W) and the signals along three spatial axes (X, Y, and Z). Higher-order ambisonics further enhances localization precision by incorporating additional channels[20].

The most notable advantage of this format is its outstanding flexibility: it allows for dynamic decoding and rendering into various output formats, ranging from standard stereo and advanced 3D surround sound to head-tracked binaural audio for VR applications[21].

2.2.3 Object-based Audio

Modern methodologies, such as Dolby Atmos, treat sound as individual objects. An audio object essentially consists of an audio signal coupled with spatial coordinates (X, Y, Z). This positional information, referred to as metadata, tracks the movement and trajectory of sound during playback. Notably, the inclusion of the Z-axis allows for the utilization of overhead speakers, thereby creating a significantly more immersive experience for the listener[22].

The most significant advantages of object-based audio lie in its spatial precision and system scalability. Since the audio is defined by spatial metadata, sounds can be accurately rendered from a specific point in space. Furthermore, the system can dynamically determine the optimal sound distribution based on the actual number and positioning of available loudspeakers[23]. This ensures a consistent and precise immersive experience across diverse reproduction environments, ranging from high-end cinemas to domestic home theaters.

2.3 Physiological and Cognitive Mechanisms of Virtual Hearing

To reconstruct 3D space through headphones, technology must mimic the biological processes of human audition. This synthesis of ITD, ILD, and HRTF enables precise spatial mapping, yet its effectiveness hinges on multisensory temporal integration. Understanding the physiological basis of spatial hearing is essential to the present study, as the precision of binaural localization cues directly determines how sensitively users may detect audiovisual temporal offsets when spatial audio is employed in VR.

2.3.1 Binaural hearing

People’s ability to perceive the direction and distance of sounds in the real world is known as binaural hearing, and it is made possible by having two ears. The shape of the pinnae, along with the geometry of the head and shoulders, interacts with incoming sound waves and creates several key cues: interaural time difference (ITD), interaural level difference (ILD), and spectral differences (SD).

ITD is the difference in the arrival time of a sound at the two ears. It is crucial for sound localization because it provides information about the direction or angle of the sound source relative to the head. ITD is most effective for low-frequency sounds. However, above about 1.5 kHz, it becomes much less reliable.

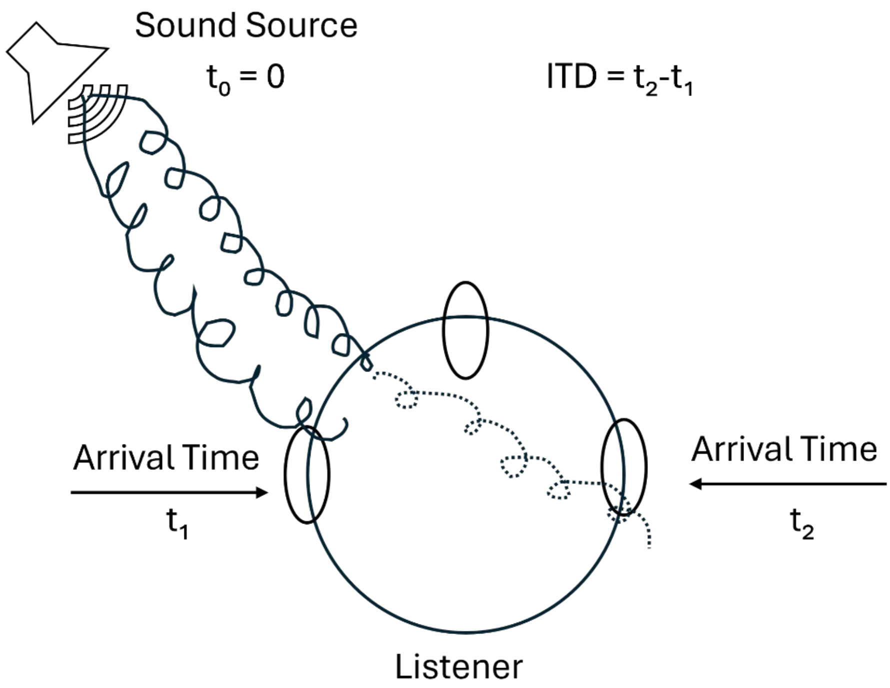

Figure 2.1 Mechanism of ITD for localized sound sources.

ILD refers to the difference in sound level reaching each ear. When a sound source is not directly in front of the listener, the ear closer to the source receives a louder signal. This cue mainly comes from the head-shadow effect, where the head partially blocks the sound. At frequencies above roughly 1.5 kHz, the sound wavelength becomes shorter than the diameter of the head, causing a significant reduction in sound energy at the opposite ear[24].

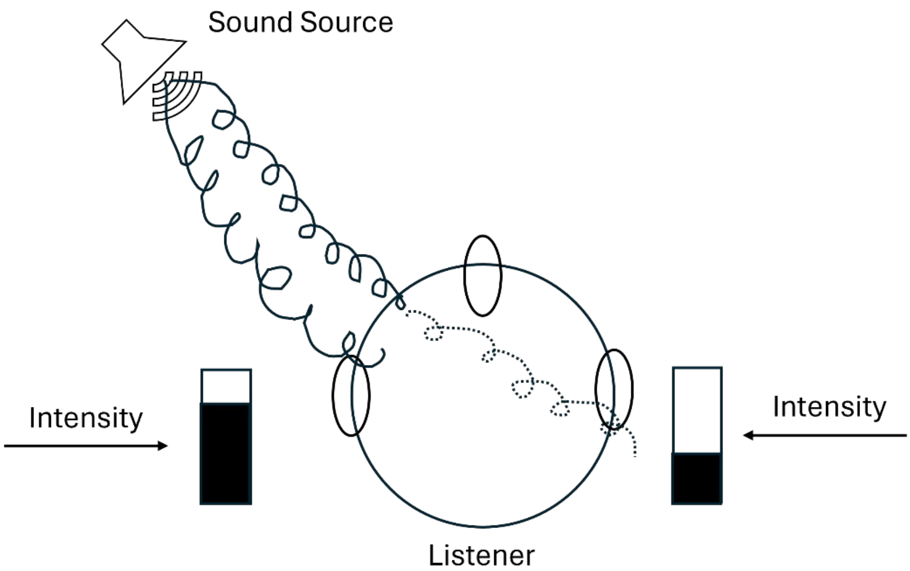

Figure 2.2 Illustration of head shadow effects on ILD.

SD are essential for resolving front-back ambiguities and determining the elevation of a sound source[25]. As sound waves interact with the complex geometry of the pinna (outer ear), head, and shoulders, certain frequency components are selectively attenuated or boosted, allowing the brain to distinguish height and depth.

2.3.2 Head-Related Transfer Function

Head-Related Transfer Function (HRTF) is the primary mechanism for achieving spatial audio, particularly in distinguishing elevation and front-back positioning[26]. HRTF employs a complex set of filters to simulate how sound waves interact with and are shaped by a listener's physical features before reaching the ear canal.

The auditory system utilizes these subtle SD to infer the height and depth of a sound source. In a spatial audio system, a mono signal is convolved with the HRTF for the left and right ears to produce a binaural signal. When played back through headphones, this signal convincingly simulates sound originating from a specific location in 3D space[27].

2.3.3 The Cocktail Party Effect

The cocktail party effect refers to the brain's ability to focus one's auditory attention on a specific stimulus[28]. It originates from our innate capacity to concentrate on a single conversation while filtering out background noise in a crowded social setting. This psychological mechanism allows us to extract meaningful information from chaotic environments, thereby enhancing our spatial awareness of sound sources[29].

2.3.4 Localization Errors

Although various physiological mechanisms assist us in localizing sound sources, humans still experience errors in judging sound direction and distance. Such errors frequently occur when the ITD and ILD of a sound are equal, a phenomenon known as the cone of confusion[30]. Under these conditions, listeners must rely solely on SD to determine the location of a sound, which often leads to front-back reversal errors.

However, these errors can be overcome by turning the head, as head rotation dynamically alters the relative position between the ears and the sound source[31]. This movement breaks the symmetry of the cone, allowing the brain to track changing acoustic cues and pinpoint the exact location of the sound[32].

2.4 Eye Tracking

Eye tracking technology measures and analyzes eye movements to determine where a user's visual attention is focused. Gaze pattern data serves as a behavioral measurement for examining how audiovisual latency and audio conditions jointly influence visual attention in VR.

2.4.1 How eye tracking works

Mainstream eye-tracking technologies primarily employ the pupil center corneal reflection method. This technique utilizes near-infrared light sources to project non-intrusive reflection patterns onto the eye such as the pupil and the cornea.

High-frame-rate sensors, sensitive to the near-infrared spectrum, capture detailed images of these reflections. Advanced image-processing algorithms then analyze the relative position between the pupil center and the glint to construct a 3D model of the eye in space. This allows the system to precisely determine the gaze angle and map it onto the subject's field of view (FOV). Notably, modern iterations of this technology can achieve high accuracy without requiring head restraints, providing a more naturalistic environment for user experience research.

2.4.2 Type of devices

Commercial eye-tracking devices are generally categorized into screen-based and wearable systems[34]. Screen-based trackers like Tobii Pro Spark are typically mounted on or near a computer monitor to track a participant's gaze while they observe stimuli from a fixed seated position. However, a significant limitation of this setup is the restricted range of motion, as participants must remain relatively stationary within the camera's FOV throughout the session.

Wearable trackers primarily take the form of VR headsets or smart glasses. These devices allow users to move freely[35], enabling diverse application scenarios and the observation of natural behavior. While the proximity of the tracker to the eyes typically yields high-precision gaze data, this mobility introduces specific challenges. Abrupt movement can shift the device's fitness, reducing data accuracy and consistency. Similarly, physical activity and environmental transitions often lead to significant variations in lighting or physiological states[36].

2.5 Visual Attention Metric

In visual attention studies, fixations, blinks, and pupil size are three frequently used metrics. As the core metric, the frequency and duration of fixations are heavily influenced by the nature of the stimuli and the observer's background.

Studies indicate that dynamic videos shorten fixation durations than static images due to rapid visual transitions[37]. Certain attributes or regions of a stimulus are more likely to capture an observer’s hidden or overt attention. These features often include distinctive color, motion, orientation, or size[38]. Furthermore, individual expertise, emotional states, and aesthetic preferences significantly alter fixation distribution. For instance, orthodontists spend more time examining dental and oral regions than normal people[39], and a participant’s preference for artworks or faces correlates strongly with total fixation count and duration[40].

Beyond fixations, blink rate and pupil diameter serve as vital indicators of cognitive states. A reduced blink rate often shows intense focus, as observers minimize eye closure to maximize information intake[41]. Meanwhile, variations in pupil diameter serves as an indicator of cognitive load and psychological reaction[42]. While pupil size typically increases with task difficulty, experienced experts demonstrate more stable pupillary changes than novices when performing the same tasks[43].

2.6 Congruence of Audiovisual Stimuli and Asynchrony

Multisensory integration depends on how well auditory and visual stimuli match each other. Previous studies in psychology and neuroscience show that the brain checks both spatial and temporal consistency to decide whether two signals come from the same event [44]. When auditory and visual stimuli happen at the same time, people detect them faster and process them more efficiently than single-sense stimuli[45]. These findings suggest that temporal congruence helps perceptual integration.

Several studies have further examined how partial incongruence influences perception. The McGurk effect demonstrates that conflicting visual lip movements and auditory speech sounds can be perceptually fused into a novel syllable[46]. This phenomenon shows that visual information can strongly influence how people hear speech.

In spatial processing, the ventriloquism effect provides evidence of visual capture[47]. When auditory and visual stimuli come from different locations, people often perceive the sound as coming closer to the visual source. Together, these studies suggest that multisensory integration does not need perfect congruence, but it works within acceptable spatial and temporal limits.

Among the various factors of congruence, temporal alignment is arguably the most critical constraint on integration. This specific temporal relationship is often characterized by the temporal binding window (TBW), defined as the range of delays within which auditory and visual signals are perceived as belonging to the same event.

Previous studies suggest that the TBW typically falls between approximately 160 and 250 ms [13], [14], [15]. However, this range is not symmetric, as observers generally show greater tolerance for auditory lag than auditory lead[48]. When temporal discrepancies exceed this window, audiovisual signals are more likely to be perceived as separate events rather than integrated perceptually[49].

In immersive VR, audiovisual timing may interact with additional sensorimotor factors, including head-tracked visual updates and spatialized audio rendering. While only few studies have directly quantified TBW under VR conditions, research on system latency suggests that increasing delays can negatively affect presence. For example, delays exceedingly approximately 100 ms have been reported to reduce the presence in immersive environments[50].

3 Methodology

This section introduces the experimental design and setup used to compare audiovisual synchrony between spatial and stereo audio. It outlines the data collection methods, analytical metrics and statistical tests.

3.1 Experiment Framework

The overall technical pipeline, including audio rendering, timing control, and data acquisition, is illustrated in Figure 3.1.

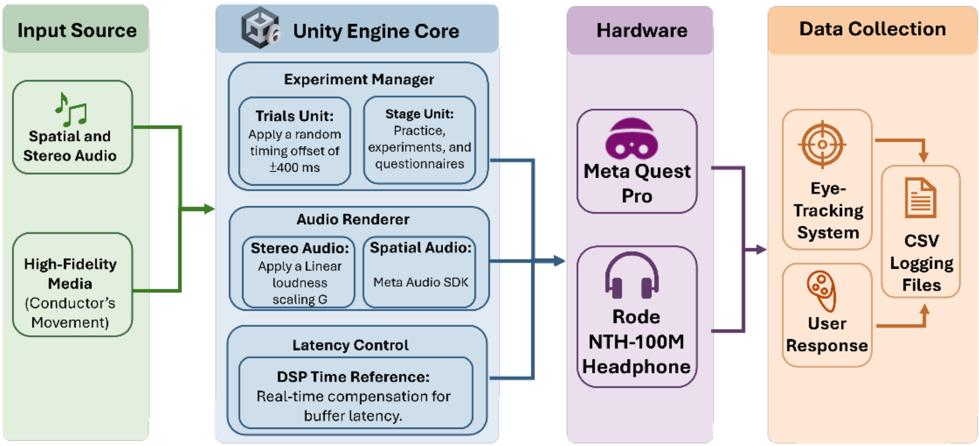

Figure 3.1 Technical framework and data pipeline. The diagram illustrates a DSP-based timing clock to precisely control latency, a dual-path audio rendering process, and a multi-modal data collection setup that combines VR controller inputs with eye-tracking.

3.1.1 Study Design

This study applied a within-subject design to eliminate the influence of individual differences on sensory judgment. Each participant completed a detection task under two audio rendering modes: stereo and spatial audio (see section 3.2.2), using the same hardware and environment. The musical material was selected from Robert Schumann's Overture, Scherzo and Finale, Op. 52. This piece was chosen for its significant dynamic variations, which allow participants to intuitively predict changes in musical intensity through the conductor's body movements. The core task for participants was to observe the conductor's actions and press a response button when they perceived an inconsistency between the video and audio.

To ensure data accuracy and reliability, audio and video files were pre-loaded before each trial. This guaranteed that the programmed latency settings were strictly accurate during the experiment.

Furthermore, to counter the order effect, the sequence of the two audio modes was counterbalanced across different participants. Before the formal experiment began, each participant completed a practice session to ensure they fully understood the task logic. This step was crucial to minimize invalid data caused by unfamiliarity with the interface or procedures.

3.1.2 Platform and Hardware

This study utilized Unity6 (6000.0.53f1) as the primary experimental development platform in conjunction with the Meta Quest Pro headset (Figure 3.2). This engine was selected primarily for its native support for stereo rendering and its integration with the Meta Audio SDK for implementing spatial audio. All experimental logic, including randomized test sequences, latency induction mechanisms, and the data logging system, was implemented using the C# programming language.

Figure 3.2 Meta Quest Pro.

The experiment was executed on an ASUS TUF Gaming A16 high-performance gaming laptop running the Windows 11 operating system. The system was equipped with an AMD Ryzen 9 7940HX processor and an NVIDIA GeForce

RTX 4060 graphics card. To ensure visual consistency and eliminate rendering jitter, the system frame rate was strictly locked at 60 FPS via Unity control.

Participants wore Rode NTH-100M headphones during the experimental trials. These headphones were connected via a wired interface to ensure zero additional transmission latency.

3.1.3 Software

Unity is a game engine developed by Unity Technologies, widely used for 2D/3D game development, interactive experiences, and VR/AR applications. It applies C# as its primary programming language and provides cross-platform capabilities, supporting deployment across mobile devices, PCs, gaming consoles, and VR headsets[51]. Furthermore, Unity features a comprehensive asset store, which offers resources that significantly lower the barrier to entry for creators and accelerates the development process.

Blender is a powerful open-source 3D creation suite that supports the full 3D pipeline, including modeling, rigging, animation, simulation, and rendering. Its flexibility and free access make it a popular choice for researchers and creators who need customized, high-quality virtual stimuli for experimental studies[52]. It is officially maintained by the Blender Foundation and offers excellent compatibility with game engines like Unity for VR asset integration.

3.2 Virtual Environment and Stimuli

3.2.1 Virtual Scene and Opera Model

The 3D model of the opera house was established through high-precision on-site scanning and kindly provided by Malmö Opera. (Figure 3.3). To provide a high-fidelity visual experience without compromising system performance, Blender 4.2 was used for detailed geometric optimization. The primary focus was on mesh decimation to reduce the complexity of the 3D models. To ensure efficient scene management and rendering within Unity, all assets were categorized and grouped into 18 distinct classes based on their functional properties, such as the ceiling, seating, and the stage area.

Custom materials were created and applied to each group to maintain visual realism while keeping the texture memory usage low (Figure 3.4).

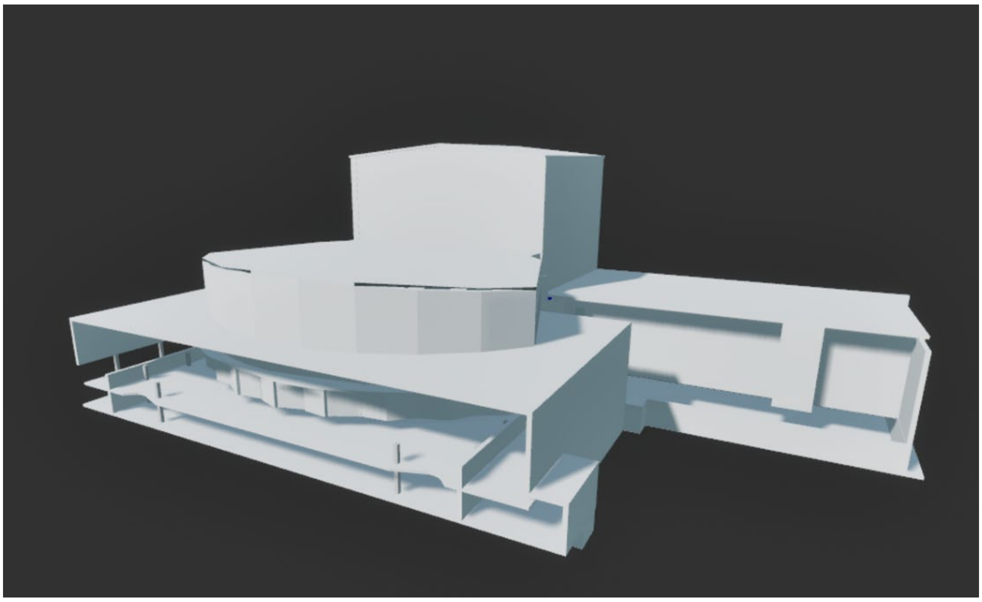

Figure 3.3: High-precision 3D scanned model of the opera house.

Figure 3.4: Opera house with final materials.

Participants were seated in a quiet testing environment at a fixed distance of 10 meters directly in front of a virtual screen. The conductor's performance was projected onto this screen, with the audio source precisely aligned with its geometric center to ensure spatial consistency between auditory and visual cues. Regarding the attenuation model for 3D audio, this study applied a linear rolloff, with the minimum distance set to 10 meters and the maximum distance set to 50 meters.

The primary objective of this configuration was to eliminate volume discrepancies arising from the spatialization process. Because the participant was positioned exactly at the threshold of the attenuation threshold, the system ensured that the output loudness remained consistent when switching between stereo and spatial audio modes. Reaction time differences caused by loudness were reduced by this control.

3.2.2 Experimental Stimuli

The musical material used in this study is based on Robert Schumann's Overture, Scherzo and Finale, Op. 52. Three highly representative excerpts were selected and produced in two versions: stereo and spatial audio. This piece was chosen primarily for its rich dynamic variations, as illustrated in Figure 3.5. The conductor's body movements in the video provide clear visual cues, allowing participants to intuitively predict changes in musical volume and intensity. The strong correlation between these visual cues and auditory signals provides a stable reference for the audiovisual synchronization task, ensuring that participants can accurately identify temporal offsets.

All experimental stimuli were processed with FFmpeg [53] to preserve audio quality and timing accuracy. The audio tracks were extracted using the stream copy mode, which allows for bit-perfect extraction without re-encoding. This preserved the original signal integrity and prevented any quality loss.

Both audio formats adhered to high-resolution professional standards. The spatial audio tracks utilized a 4.0 quadraphonic channel configuration, encoded in pcm_s24le (24-bit Little-Endian) with a sampling rate of 48,000 Hz. This configuration provides a bit rate of 4,608 kb/s, ensuring a broad dynamic range and sufficient spatial data for the Meta Audio SDK to perform binaural rendering.

The Stereo Audio tracks were maintained at an identical resolution (pcm_s24le, 48,000 Hz). This ensured that any observed differences in participant performance were attributable to the spatial rendering mode rather than discrepancies in basic audio quality or bit depth. Storing 24-bit audio in 32-bit containers preserved the sound details for precise audiovisual synchronization judgments.

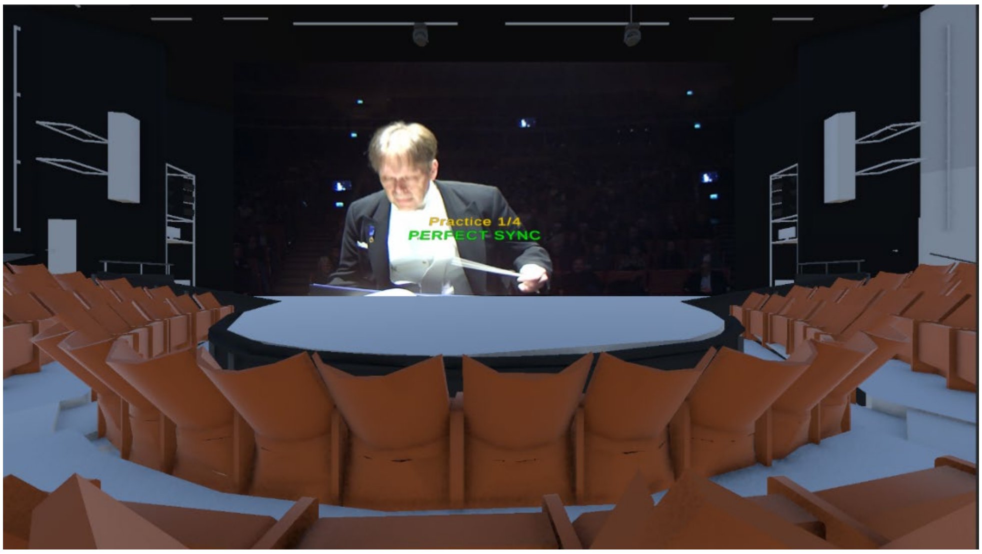

Figure 3.5: Virtual opera house used for the VR environment.

3.3 Audio Timing Mechanism and Volume

To overcome the timing uncertainties in a game engine's traditional rendering loop, this experiment replaced frame-based coroutine control with Unity's underlying Digital Signal Processing (DSP) time technology. DSP Time provides a precise timeline derived directly from the audio hardware's sampling counter, operating independently of CPU rendering loads or frame rate fluctuations. This allows the system to achieve microsecond-level precision for scheduled audio playback at the audio sampling level. The advantage of this configuration is its stability. Even when FPS fluctuates during complex scenes, audio triggers still follow the preset latency settings.

To control loudness-induced bias, stereo and spatial signals were balanced through independent output channels. Initial measurements indicated that the stereo signal was approximately 10 dB louder than the spatial audio. Therefore, a linear scaling factor G was applied to the stereo output, calculated as:

$$G = 10^{((-10)/20)} \approx 0.316 \qquad (1.2)$$

This adjustment ensured that the perceived intensity remained consistent across both audio formats.

3.4 Experimental Procedure and Protocol

3.4.1 Pilot Study

Before the formal experiment, a pilot study was conducted with three participants to evaluate the experimental procedure, hardware stability, and participants' ability to perceive spatial audio. The pilot results led to several refinements in the experimental design:

1. **Spatial Audio Perception:** All participants successfully distinguished spatial audio from stereo in all trials. This confirmed that the spatial cues generated by the Meta Audio SDK were perceptible within the virtual opera house environment.
2. **Hardware and Participant Selection:** During the pilot phase, it was observed that framed glasses interfered with the VR headset's internal spacing, causing the ray to miss the colliders. Therefore, the formal experiment recruited only participants with normal vision or those using contact lenses to avoid this technical issue.
3. **Training and Instruction:** Initial data showed that accuracy at extreme offsets was below 80%, likely due to participants' unfamiliarity with the VR controller. During the experiment, some participants were uncertain about how to use the VR controller to indicate whether the video or audio was leading, which led them to seek guidance. This may have increased their nervousness or caused distractions. To address this issue, the training phase was extended to include additional hands-on instruction and visual demonstrations (Figure 3.6), where participants were shown the controller interface with labeled response options and were also given verbal instructions on how to perform the task before wearing the VR headset.

Figure 3.6: Visual instructions and cues displayed on the VR controllers.

3.4.2 Assessment Protocol

To measure the impact of varying audio configurations on concert streaming experiences, this study was conducted at Lund University Humanities Lab using a Meta Quest Pro headset. Participants were seated on chairs to ensure a quiet and undisturbed environment. Each session lasted approximately 40–50 minutes and was structured into the following phases:

1. **Information and Screening Phase (10 mins):** Participants receive a briefing and sign informed consent forms. This is followed by screenings for visual acuity, auditory acuity, and color perception to ensure their sensory capabilities meet the experimental baseline.
2. **Calibration Phase (5 mins):** Participants underwent equipment calibration to minimize adaptation errors during testing.
3. **Training Phase (5 mins):** Participants familiarize themselves with the VR environment and establish consistent judgment criteria through a repeatable introductory video.
4. **Core Testing Phase (25-30 mins):** Three musical stimuli are presented with audiovisual latency ranging from −400 ms to +400 ms. Two audio conditions (spatial and stereo) are tested, with the counterbalanced order across participants, such that half experienced the spatial condition first and the other half the stereo condition first. A double-blind design is applied, with randomized playback order, latency offsets, and lead–lag direction. After each trial, participants indicate whether the audio or video appeared first, while gaze behavior is continuously recorded during the formal testing phase.
5. **Questionnaire Phase (5 mins):** After each audio condition block, participants completed a subjective survey evaluating their experience across five indicators: presence, naturalness, confidence, perceived difficulty, and mental effort. (Figure 3.7).

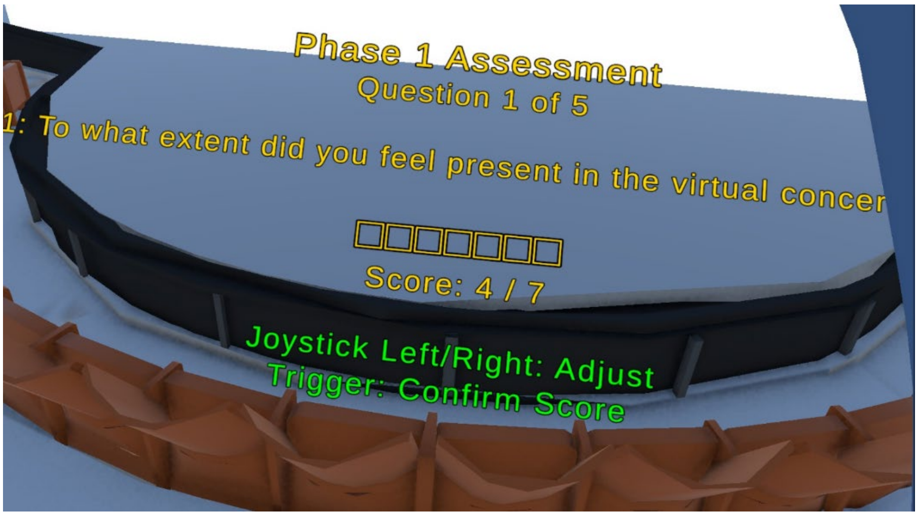

Figure 3.7: The subjective test questionnaire during the experiment.

3.4.3 Participants

The study was conducted with 18 participants (ages 21 - 30), including 9 females; no participants identified as non-binary or preferred not to disclose their gender. All participants provided written informed consent and reported normal or corrected-to-normal vision. Data from one participant was excluded from the analysis as they reported hearing impairments or an imbalance between both ears. Participants also reported their level of musical training and familiarity with orchestral conducting, showing the participants come from diverse training backgrounds (Figure 3.8).

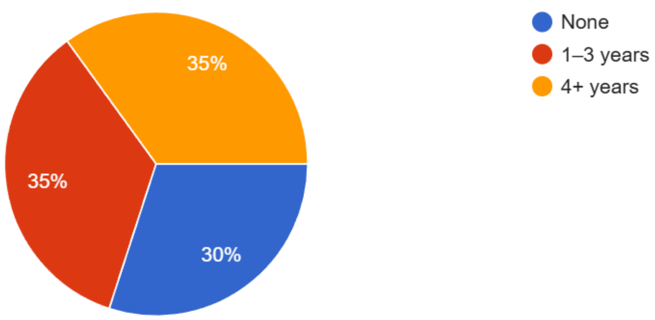

Figure 3.8: Years of formal musical training among the participants.

60% of the participants identified as beginners regarding the use of VR headsets and 25% reported being regular users or developers. This distribution provided a diverse background for the study (Figure 3.9).

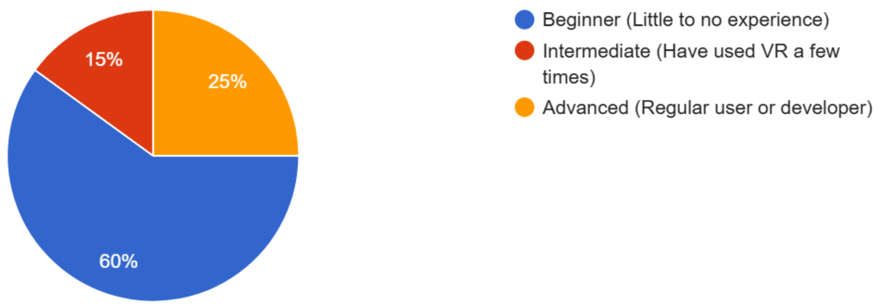

Figure 3.9: Distribution of participants' familiarity with VR headset.

3.5 Data Collection and Analysis

Behavioral responses collected via hand controllers (binary temporal judgments and questionnaire ratings) and continuously recorded eye-tracking data formed the main dataset. To ensure spatial accuracy in the virtual environment, colliders were applied to the display screen. The system used ray-casting to convert raw gaze data into UV coordinates on the screen. These coordinates were visualized as heatmaps for analyzing gaze patterns across experimental conditions.

Temporal perception of audiovisual stimuli was measured by fitting a psychometric function to video-first response rates across seven latency points for each participant and condition. Under the assumption of a cumulative normal distribution, the response probability P(x) was modeled as follows:

$$P(x) = \lambda + (1 - 2\lambda) \cdot \Phi(x; \mu, \sigma) \tag{1.2}$$

Where:

- P(x) is the probability of a video leading response.
- λ denotes the symmetric guessing and lapse rates.
- $\Phi(x; \mu, \sigma)$ is the cumulative normal distribution function.
- μ represents the point of subjective simultaneity (PSS).
- σ represents the just noticeable difference (JND).

The PSS indicates the temporal interval at which the participant perceives the two stimuli as being maximally simultaneous, whereas the JND reflects the sensitivity of temporal discrimination, defined as the stimulus increment required to shift the response probability from 50% to 75%. Finally, the estimated PSS and JND values across spatial and stereo audio conditions were subjected to a one-way repeated Analysis of Variance (ANOVA) and the subjective ratings were conducted the Mann-Whitney U test to determine the statistical significance of the experimental result.

4 Results

This chapter compares spatial and stereo audio performance in a VR environment using psychophysical fitting and eye-tracking analysis. The methodology combines behavioral metrics such as PSS and JND with gaze distribution to examine how participants process visual and auditory cues.

4.1 Behavioral Performance

To systematically analyze the identification performance of the participants, this study divided the testing points into four intervals: extreme latency (400 ms), moderate latency (200 ms), short latency (100 ms), and near-threshold latency (50 ms).

The results show that the overall accuracy rates for spatial audio and stereo audio were similar across all participants, at 0.66 ± 0.09 and 0.67 ± 0.09, respectively. Under both audio conditions, the accuracy improved gradually as the latency increased, demonstrating typical psychophysical experimental characteristics.

In the spatial audio condition, the data revealed a clear perceptual asymmetry. At -400 ms (audio leading), the accuracy reached 88%, the highest in the entire experiment; however, at the corresponding +400 ms, the accuracy dropped to 73% (Figure 4.1). This trend appeared across all latency intervals. For example, the accuracy for audio leading under moderate latency was 70%, much better than the 58% for video leading. Nevertheless, within the near-threshold latency interval, performance was close to chance level, with the video-leading side (56%) slightly higher than the audio-leading side (54%), suggesting that participants found it difficult to make stable judgments within this range.

In contrast, the asymmetry in the stereo audio condition was less obvious. At extreme latency, the accuracy rates for audio leading and video leading were 84% and 75%, respectively. In the moderate latency interval, the gap between the two remained at approximately 10%. Notably, under the short latency interval, the gap significantly narrowed to within 3%, showing a more symmetric perceptual characteristic compared to spatial audio.

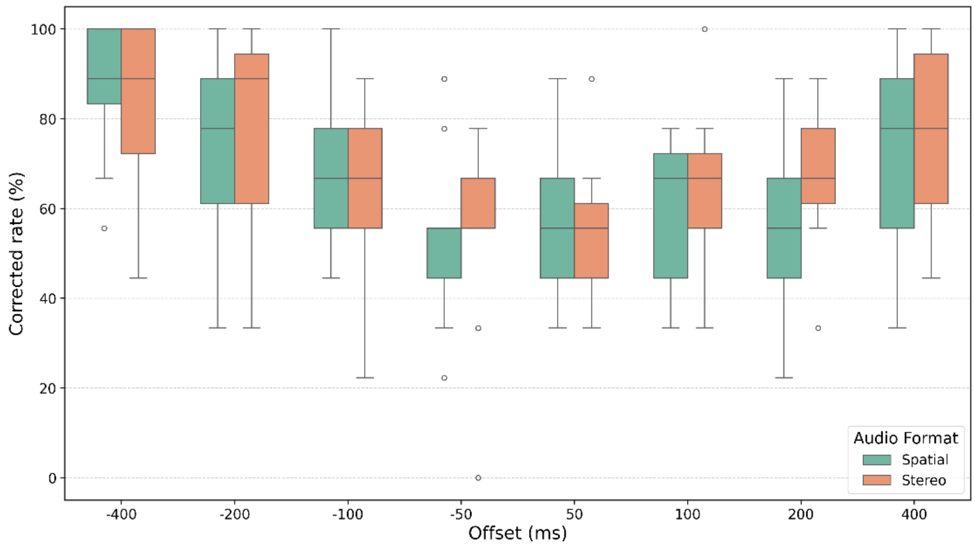

Figure 4.1: Behavioral performance under spatial and stereo audio conditions across different testing points.

4.2 Analysis of PSS and JND

By fitting the psychometric functions to the participants' response data, this study extracted the PSS and the JND as quantitative measures of perception. The fitting results, including the 95% confidence intervals (CI), are illustrated in Figure 4.2.

In the spatial audio condition, the calculated PSS was 71.17 ms and the JND was 461.23 ms. The psychometric curve for this condition exhibits a wider shaded region, representing the 95% CI, which indicates the range of statistical uncertainty in the model's estimation.

In the Stereo Audio condition, the PSS and JND were 22.47 ms and 224.14 ms, respectively. Compared to spatial conditions, the JND value in stereo is approximately 50% lower. Furthermore, the 95% CI in the stereo condition is notably narrower across the most offsets.

This study performed repeated measures ANOVA on both PSS and JND across the two audio formats. The results showed that the effect of audio format on PSS did not reach the threshold for statistical significance $F(1, 16) = .07, p = .79$. Similarly, the effect on JND did not reach a significant level $F(1, 16) = .21, p = .65$.

There were 10 participants who reached the maximum JND limit of 550 ms under the spatial audio or stereo conditions, which affected the overall fitting results. Among them, 4 participants reached the threshold in both audio conditions. In the stereo and spatial conditions, 4 and 2 participants respectively reached the upper limit. After excluding participants who reached the maximum JND limit and reanalyzing the data, the effects remained non-significant for both PSS, $F(1, 7) = 2.08$, $p = .19$, and JND, $F(1, 7) = 4.25$, $p = .08$. Although the JND result showed a trend toward significance, it did not meet the conventional threshold for statistical significance.

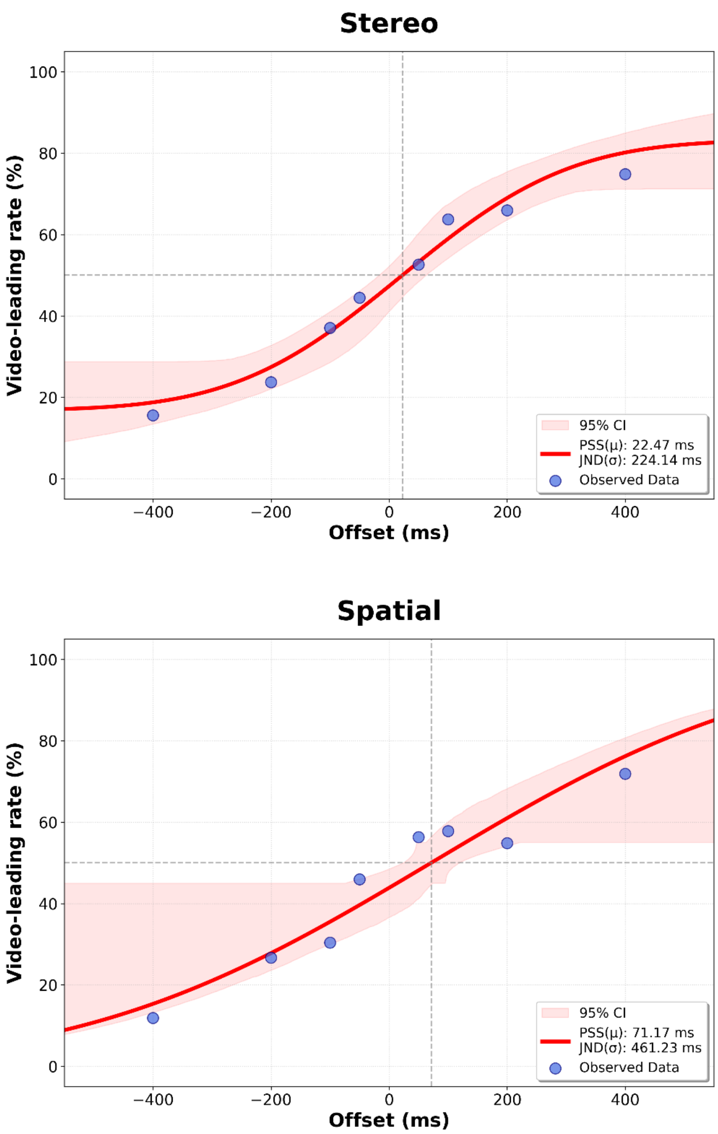

Figure 4.2: Psychometric curves of aggregated participant responses for Stereo and Spatial conditions. The PSS and JND are derived from the fitted Gaussian cumulative distribution function. The red shaded area represents the 95% confidence interval."

4.3 Testing Order and Learning Effects

An exploratory post-hoc descriptive analysis was conducted to examine potential effects of testing order on participant performance. No inferential statistical tests were performed. During the initial testing phase, participants in the spatial audio condition achieved an average accuracy of 67%, compared to 63% in the stereo condition. The average response time was 134 s under spatial audio and 183 s under stereo audio.

In the second testing phase, participants who completed the spatial audio condition first and the stereo condition second showed an increase in accuracy from 67% to 70%, while their average response time decreased to 95 s. In contrast, participants who completed the stereo condition first showed a reduction in response time from 183 s to 102 s, whereas their accuracy remained relatively stable at 63%.

Because these observations were based solely on descriptive comparisons and were not part of the original planned statistical analysis, the findings should be interpreted cautiously.

Table 5.1 Performance across testing orders and audio conditions

Testing Orders	Phase 1 (% / Time)	Phase 2 (% / Time)	Difference (Time)
Spatial → Stereo	67% / 134s	70% / 95s	+3% / -39s
Stereo → Spatial	63% / 183s	63% / 102s	0% / -81s

4.4 Eye-tracking Analysis

This study analyzed gaze distribution under different audio conditions by recording the participants' UV coordinate data during the experiment. The data indicates that the participants' fixation areas showed a high degree of consistency and concentration across both audio conditions (Figure 4.3).

Under both audio conditions, fixation points remained nearly identical, with average coordinates of ($X = 0.46 \pm 0.10$, $Y = 0.67 \pm 0.22$) and ($X = 0.45 \pm 0.10$, $Y = 0.67 \pm 0.24$), respectively. By mapping these coordinates to the spatial layout of the experimental scene, it was found that these fixation points correspond precisely to the head of the conductor in the center of the frame.

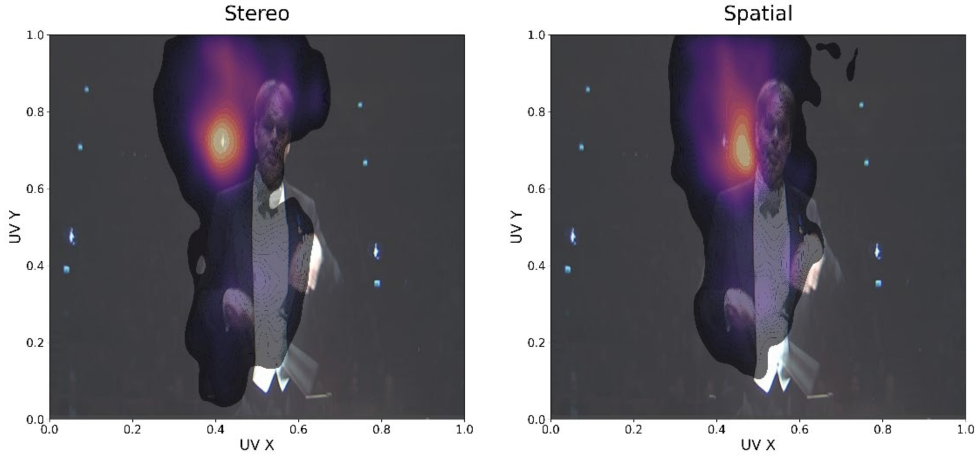

Figure 4.3: Heatmap of visual gaze distribution across spatial and stereo audio conditions.

4.5 Subjective Experience

Following the completion of the two audio conditions, a 7-point Likert scale survey was conducted to assess the participants' subjective experiences across five indicators: presence, naturalness, confidence, difficulty, and mental effort. The results are summarized in Table 4.1.

Regarding task-related indicators, Mann–Whitney U tests were conducted to compare subjective ratings between the two audio formats. The results showed a high degree of consistency across conditions. Participants reported similar levels of confidence in their judgments under the spatial audio and stereo conditions (M = 3.50, SD = 1.45 vs. M = 3.57, SD = 0.94). Ratings of perceived task difficulty were identical across conditions (M = 5.07, SD = 0.92 for spatial audio; M = 5.07, SD = 1.00 for stereo audio). The mean ratings were above the midpoint of the 7-point scale in both conditions. For mental effort, participants reported a slightly lower score in the spatial audio condition (M = 5.00, SD = 1.41) than in the stereo condition (M = 5.36, SD = 1.36); however, this difference did not reach statistical significance.

For the experiential indicators, spatial audio received higher average ratings than stereo audio on both naturalness and presence. The mean naturalness rating increased from 3.71 in the stereo condition to 4.43 in the spatial audio condition. Similarly, the mean presence rating increased from 4.14 (SD = 1.23) to 4.36 (SD = 1.34). Although the observed differences did not reach statistical significance, the descriptive data consistently favored the spatial audio condition for both experiential measures.

Table 4.2 Descriptive statistics of subjective ratings for spatial and stereo audio conditions.

	Spatial Audio		Stereo Audio	
	Mean	**SD**	**Mean**	**SD**
Presence	4.36	1.34	4.14	1.23
Naturalness	4.43	1.28	3.71	1.44
Confidence	3.50	1.45	3.57	0.94
Difficulty	5.07	0.92	5.07	1.00
Mental Effort	5.00	1.41	5.36	1.36

5 Discussion

This chapter provides an in-depth discussion on the final findings of data analysis, limitations and future work.

5.1 Impact of Spatial Audio on PSS and JND

In the fitting process of the psychometric functions, this study found around two times difference in JND, showing a clear trend indicating that participants were much less sensitive to audiovisual delay when spatial audio was used.

Notably, over half of the participants reached the upper limit threshold (550 ms) in at least one audio condition, with most of these saturation cases occurring in the stereo condition. This distribution suggests that spatial audio elevated the collective perceptual threshold across the group, even though stereo induced more extreme cases.

This implies that the complex multi-sensory rendering of spatial audio may introduce higher cognitive load and cross-modal uncertainty, thereby making it more difficult for participants to detect small temporal differences in a complex conducting task.

A similar pattern appeared in the PSS results. The PSS for stereo audio was close to the point of physical synchrony, whereas the PSS for spatial audio shifted strongly toward the visual-lead direction. This suggests that spatial audio changes the perceived point of synchrony, requiring visual signals to occur earlier than audio signals for participants to perceive them as synchronized. Beyond the shifts in mean PSS and JND values, this study observed that the 95% CI for spatial audio was consistently wider across most offsets compared to the stereo condition. This reflects greater uncertainty under spatial audio.

These findings are consistent with increased processing demands during multisensory integration. Spatial audio provides more complex auditory information, such as direction and distance, which increases perceptual and attentional load during multisensory integration [44], [45], [54]. This increased load is probably associated with shifts in PSS and a higher tolerance for audiovisual latency.

In this experiment, participants had to combine detailed visual information with matching sounds in a dynamic situation. This task required effective coordination between vision and hearing. Spatial audio provides clearer directional cues, which may improve perceptual organization and create a more natural experience. However, this also increases multisensory processing demands, which may reduce the attentional resources available for precise temporal judgments. Consistent with previous studies on attention and multisensory integration[55], participants showed lower temporal precision, reflected by higher JND values.

5.1.1 Visual Attention and Gaze Behavior

The gaze data showed a high level of consistency across both audio conditions. Participants naturally focused on the conductor's head when making audiovisual synchrony judgments, regardless of the audio format used. This is likely because the conductor's baton movements and head gestures are the most time-predictable visual events in the scene, making them a natural reference point for cross-modal timing comparisons. Previous study suggests that people are especially sensitive to facial orientation[56], which may help explain why gaze was anchored to the conductor's face rather than to hand movements.

More importantly, the fact that gaze patterns were so similar across conditions suggests that the differences found in PSS and JND were not caused by changes in how participants visually sampled the scene. Participants did not shift their visual attention in response to spatial audio. This means that the perceptual differences observed between audio conditions most likely reflect changes in auditory processing rather than changes in visual information gathering. This rules out an important confounding factor and supports the interpretation that spatial audio itself was responsible for the shifts in JND and PSS.

5.1.2 Subjective Experience and Objective Performance

The questionnaire results revealed an interesting gap between how participants felt and how they performed. The relatively high difficulty ratings observed in both conditions further suggest that the audiovisual synchrony judgment task imposed substantial perceptual demands on participants, regardless of audio format. Although ratings for task difficulty and confidence were similar across both conditions, spatial audio received more positive scores for naturalness, presence, and mental workload. This implies that richer spatial cues created a more convincing sense of presence and reduced perceived cognitive effort.

However, this positive subjective experience did not translate into better objective performance on temporal judgement tasks. Participants felt the task was more natural and easier under spatial audio, yet their sensitivity to timing offsets decreased. This points to an important limitation of relying solely on subjective preference ratings when evaluating synchronization quality, as users may still rate an experience positively even when it contains objectively measurable timing offsets.

Subjective questionnaires alone may overestimate the actual synchronization accuracy of a system when used in spatial audio environments. A more reliable evaluation should combine subjective measures with objective temporal sensitivity indicators such as PSS and JND, providing a fuller picture of how well a system performs.

5.1.3 Potential Learning Effects of Spatial Audio Exposure

The descriptive analysis of testing order revealed a pattern that may be relevant to task familiarization. Participants who experienced spatial audio during the initial testing phase achieved slightly higher accuracy and shorter response times than those who began with stereo audio. One possible explanation is that the richer directional information provided by spatial audio offered more intuitive cues in an unfamiliar environment, thereby reducing the cognitive demands associated with multisensory integration.

Furthermore, participants who completed the spatial audio condition first showed improvements in both accuracy and response time during the subsequent stereo condition. This pattern may suggest that exposure to spatial audio facilitated the development of effective task strategies that transferred to the later stereo condition. In contrast, participants who completed the stereo condition first primarily showed reductions in response time without improvements in accuracy, which may reflect increased procedural familiarity rather than enhanced judgment precision.

Because these observations were exploratory and based solely on descriptive comparisons, future studies should employ larger sample sizes and inferential analyses to determine whether spatial audio can reliably facilitate learning and strategy acquisition in audiovisual synchrony tasks.

5.2 Summary of Research questions

5.2.1 RQ1: How does audiovisual latency affect user immersion and behavior when using spatial audio compared to stereo audio?

This study suggests that spatial audio may positively influence user immersion and behavioral responses during audiovisual synchrony judgment tasks. Although no statistically significant differences were observed in the subjective ratings, participants consistently reported higher levels of naturalness and presence under the spatial audio condition.

At the same time, confidence ratings remained relatively low, while difficulty and mental effort ratings were relatively high across both audio conditions, indicating that the task was generally challenging. This interpretation is further supported by the fact that 10 participants reached the upper limit of the JND fitting range (550 ms) in at least one condition, suggesting substantial variability and difficulty in detecting audiovisual latency.

The exploratory testing-order analysis also indicated shorter response times and potentially faster task familiarization under spatial audio. It shortens the reaction time in judgment tasks and speeds up the learning curve, allowing participants to reach a stable level of accuracy with lower cognitive effort.

5.2.2 RQ2: Is spatial audio more perceptually sensitive to audiovisual latency than stereo audio?

The experimental results reject the hypothesis that spatial audio increases sensitivity to audiovisual latency. Instead, the data showed that participants demonstrated lower latency sensitivity in spatial audio condition. Specifically, the data indicates that both the Just JND and the PSS shift were larger for spatial audio than for stereo audio.

This suggests competition for cognitive resources during multisensory processing. When the brain processes rich spatial information, it appears to reduce the precision of timing judgments, which effectively expands the TBW. These observations suggest a possible tradeoff between spatial information processing and temporal judgement. This relationship still requires further verification through future studies with larger sample sizes.

5.2.3 RQ3: Do users detect latency mainly through visual cues or auditory cues?

The experimental results suggest that participants did not rely on a single sensory modality when detecting audiovisual latency. Instead, they appeared to integrate both visual and auditory information during synchrony judgments. Eye-tracking data revealed that participants consistently relied on the conductor's head movements and facial expressions as visual anchors, while auditory information was used to evaluate whether the perceived timing of the audiovisual events was synchronized.

Regarding sensory sensitivity, the accuracy for detecting audio-lead trials was higher than that for video-lead trials. This suggests that participants were particularly sensitive to situations in which auditory signals arrived before their corresponding visual cues. Because such temporal relationships are relatively uncommon in everyday perception, audio-leading events may be more noticeable and therefore easier to detect.

5.3 Limitations

5.3.1 Data Collection

The eye-tracking system was sometimes unstable when it first started. Although it was initially suspected that individual physiological traits, such as eye shape or pupil characteristics, caused sensor failure, it was later confirmed that the issue could be resolved by restarting the system. Consequently, eye-tracking data for some early participants were not fully recorded.

Moreover, the current gaze-track range is restricted to the screen. When participants explored other parts of the opera house model, their visual paths could not be effectively captured, limiting the depth of analysis regarding comprehensive visual exploration behavior in immersive environments.

5.3.2 Repeated Test and Interface Issues

The long and repetitive trials could also cause fatigue, which might have reduced judgment accuracy over time. This fatigue effect may have become more noticeable in the later stages of the experiment.

Regarding the interface, although instructions were shown on the VR controllers, some participants felt confused at the start. This initial learning curve may have led to inconsistent data in the early trials. As participants became more familiar with the system, their performance likely became more stable.

5.3.3 Latency Measurement

In VR development, syncing audio and video is limited by the hardware. While we used software to optimize the timing, the HMD itself can cause extra delays. Also, we did not have special hardware to measure the exact latency in real-time. We could not use the built-in recording tool in Unity for verification because it caused the screen to lag during testing.

5.4 Future Work

5.4.1 Spatial Positioning and Social Presence

To control experimental variables, this study focused on a single observation point directly in front of the stage. However, in real performance venues, audience members are distributed across various locations, such as side boxes or rear seats, where acoustic reflections and visual perspectives differ significantly. Future research should compare how perceptual thresholds for synchrony change across different spatial coordinates. Furthermore, as virtual environment technology evolves, incorporating multi-user social scenarios is a vital direction. Investigating whether social presence in shared virtual spaces causes shifts in attention and subsequently affects the brain's tolerance for audiovisual errors is essential for building highly interactive virtual opera houses.

5.4.2 Long-term Exposure

Most participants in this study were using spatial audio systems for the first time, meaning their judgments were likely influenced by daily listening experiences with non-spatialized audio, such as traditional headphones or stereo speakers. Future research could observe whether the brain undergoes sensory recalibration after long-term and repeated exposure to spatial audio systems. It is possible that the brain eventually treats spatial cues as the "new normal," which might change how it integrates sight and sound. Through neural adaptation, studies could determine if the temporal integration window converges, leading to a narrower JND or a shift in PSS. Such findings would have high application value for developing professional virtual music training systems, where high precision is required for musicians to perform together effectively in a digital space.

5.4.3 Increased Diversity of Participants

The current study primarily involved university students, which limits the generalizability of the findings across different demographics. It is recommended that future studies recruit participants from various age groups to see if the tolerance for audiovisual latency increases or decreases with age. Additionally, involving participants with professional musical backgrounds, such as conductors or orchestral musicians, could provide deeper insights. These professionals are trained to detect very small timing differences in acoustic environments. Further investigation could show how professional experience and specialized auditory training influence the temporal integration window in spatial audio environments compared to the public.

6 Conclusions

VR together with spatial audio is one possible solution for opera houses seeking to develop digital opera experiences, and this thesis addressed several fundamental questions regarding audiovisual latency in digital system. Through an audiovisual temporal judgment experiment, this study estimated a PSS of 71.17 ms and a JND of 461.23 ms. In addition, an automated experimental framework was developed using the Meta Quest Pro, Rode headphones, and the Unity engine, providing an efficient approach for collecting synchronized behavioral and eye-tracking data in virtual environments.

This study demonstrated that spatial audio influences audiovisual synchrony perception differently from what was originally hypothesized. Although spatial audio was expected to increase sensitivity to audiovisual latency, the results suggested a tendency toward reduced temporal sensitivity, reflected by larger PSS and JND estimates compared with stereo audio. These findings indicate that the processing of rich spatial auditory information may increase tolerance to audiovisual asynchrony during multisensory integration.

The eye-tracking results further showed that participants adopted a multisensory strategy when judging audiovisual synchrony. Rather than relying on a single sensory modality, participants combined visual and auditory information, using the conductor's head movements and facial expressions as stable visual anchors while evaluating temporal relationships through auditory cues. This finding contributes to a better understanding of how users process audiovisual information in complex virtual performance environments.

Overall, this thesis provides practical reference regarding audiovisual latency perception in VR opera experiences and highlights the importance of considering both perceptual performance and user experience when designing future audiovisual systems. The findings offer practical insights for future virtual performance platforms, remote music applications, and spatial audio rendering systems.

References

[1] A. Striner, S. Halpin, T. Röggla, and P. Cesar, "Towards immersive and social audience experience in remote vr opera," in *Proceedings of the 2021 ACM International Conference on Interactive Media Experiences*, 2021, pp. 311–318.

[2] F. Rumsey, *Spatial audio*. Routledge, 2012.

[3] C. S. W. Anderson and M. Thorogood, "A grounded study in user preference for immersive sound in extended reality environments," *Virtual Real.*, vol. 30, no. 2, 16 2026, doi: 10.1007/s10055-026-01348-9.

[4] K. Nowak, L. Tankelevitch, J. Tang, and S. Rintel, "Hear We Are: Spatial Audio Benefits Perceptions of Turn-Taking and Social Presence in Video Meetings," in *Proceedings of the 2nd Annual Meeting of the Symposium on Human-Computer Interaction for Work*, in CHIWORK '23. New York, NY, USA: Association for Computing Machinery, 20 2023, pp. 1–10. doi: 10.1145/3596671.3598578.

[5] A. Afonso-Jaco and B. F. Katz, "Spatial knowledge via auditory information for blind individuals: Spatial cognition studies and the use of audio-VR," *Sensors*, vol. 22, no. 13, p. 4794, 2022.

[6] D. M. Greenberg, E. Bodner, A. Shrira, and K. R. Fricke, "Decreasing Stress Through a Spatial Audio and Immersive 3D Environment: A Pilot Study With Implications for Clinical and Medical Settings," *Music & Science*, vol. 4, p. 2059204321993992, Jan. 2021, doi: 10.1177/2059204321993992.

[7] "Research ethics and animal testing ethics | Staff Pages." Accessed: Apr. 20, 2026. [Online]. Available: https://www.staff.lu.se/research-and-education/research-support/research-ethics-and-animal-testing-ethics

[8] M. Wallergård, G. Alce, and J. Eriksson, "Virtual reality," 2019.

[9] L. Freina and M. Ott, "A literature review on immersive virtual reality in education: state of the art and perspectives," in *The international scientific conference elearning and software for education*, 2015, pp. 10–1007.

[10] G. Saposnik, M. Levin, and S. O. R. C. (SORCan) W. Group, "Virtual reality in stroke rehabilitation: a meta-analysis and implications for clinicians," *Stroke*, vol. 42, no. 5, pp. 1380–1386, 2011.

[11] G. Plancher, A. Tirard, V. Gyselinck, S. Nicolas, and P. Piolino, "Using virtual reality to characterize episodic memory profiles in amnestic mild cognitive impairment and Alzheimer's disease: influence of active and passive encoding," *Neuropsychologia*, vol. 50, no. 5, pp. 592–602, 2012.

[12] A. Patney *et al.*, “Towards foveated rendering for gaze-tracked virtual reality,” *ACM Transactions On Graphics (TOG)*, vol. 35, no. 6, pp. 1–12, 2016.
[13] R. A. Stevenson and M. T. Wallace, “Multisensory temporal integration: task and stimulus dependencies,” *Experimental brain research*, vol. 227, no. 2, pp. 249–261, 2013.
[14] A. R. Hillock, A. R. Powers, and M. T. Wallace, “Binding of sights and sounds: age-related changes in multisensory temporal processing,” *Neuropsychologia*, vol. 49, no. 3, pp. 461–467, 2011.
[15] N. F. Dixon and L. Spitz, “The detection of auditory visual desynchrony,” *Perception*, vol. 9, no. 6, pp. 719–721, 1980.
[16] “Head motion and latency compensation on localization of 3D sound in virtual reality | Proceedings of the ACM symposium on Virtual reality software and technology.” Accessed: May 12, 2026. [Online]. Available: https://dl.acm.org/doi/abs/10.1145/261135.261140
[17] J. Blauert, *Spatial hearing: the psychophysics of human sound localization*. MIT press, 1997.
[18] F. E. Toole, *Sound reproduction: the acoustics and psychoacoustics of loudspeakers and rooms*. Routledge, 2017.
[19] M. A. Gerzon, “Periphony: With-height sound reproduction,” *J. Audio Eng. Soc*, vol. 21, no. 1, pp. 2–10, 1973.
[20] S. Moreau, J. Daniel, and S. Bertet, “3d sound field recording with higher order ambisonics–objective measurements and validation of a 4th order spherical microphone,” in *120th Convention of the AES*, 2006, pp. 20–23.
[21] F. Zotter and M. Frank, *Ambisonics: A practical 3D audio theory for recording, studio production, sound reinforcement, and virtual reality*. Springer, 2019.
[22] “Dolby Atmos for the Home Theater”.
[23] P. Coleman *et al.*, “An audio-visual system for object-based audio: from recording to listening,” *IEEE Transactions on Multimedia*, vol. 20, no. 8, pp. 1919–1931, 2018.
[24] D. R. Begault and L. J. Trejo, “3-D Sound for Virtual Reality and Multimedia,” NASA/TM-2000-209606, Aug. 2000. Accessed: Apr. 13, 2026. [Online]. Available: https://ntrs.nasa.gov/citations/20010044352
[25] A. Meshram, R. Mehra, H. Yang, E. Dunn, J.-M. Franm, and D. Manocha, “P-HRTF: Efficient personalized HRTF computation for high-fidelity spatial sound,” in *2014 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, Sep. 2014, pp. 53–61. doi: 10.1109/ISMAR.2014.6948409.
[26] S. Riedel, M. Frank, and F. Zotter, “Effect of HRTFs and head motion on auditory-visual localization in real and virtual studio environments,” *Acta Acust.*, vol. 9, p. 21, 2025, doi: 10.1051/aacus/2025005.
[27] X. Zhong and B. Xie, “Head-Related Transfer Functions and Virtual Auditory Display,” in *Soundscape Semiotics - Localization and Categorization*, IntechOpen, 2014. doi: 10.5772/56907.

[28] B. Arons, “A review of the cocktail party effect,” *Journal of the American Voice I/O society*, vol. 12, no. 7, pp. 35–50, 1992.

[29] J. A. Veltman, A. B. Oving, and A. W. Bronkhorst, “3-D Audio in the Fighter Cockpit Improves Task Performance,” *The International Journal of Aviation Psychology*, vol. 14, no. 3, pp. 239–256, Jun. 2004, doi: 10.1207/s15327108ijap1403_2.

[30] Z. Ben-Hur, D. L. Alon, P. W. Robinson, and R. Mehra, “Localization of virtual sounds in dynamic listening using sparse HRTFs,” in *2020 AES International Conference on Audio for Virtual and Augmented Reality, AVAR 2020*, 2020.

[31] H. Wallach, “The role of head movements and vestibular and visual cues in sound localization.,” *Journal of Experimental Psychology*, vol. 27, no. 4, p. 339, 1940.

[32] F. L. Wightman and D. J. Kistler, “Resolution of front-back ambiguity in spatial hearing by listener and source movement,” *J Acoust Soc Am*, vol. 105, no. 5, pp. 2841–2853, May 1999, doi: 10.1121/1.426899.

[33] J. Hartbrich, F. Weidner, C. Kunert, A. Raake, W. Broll, and S. Arévalo Arboleda, “Eye and face tracking in VR: Avatar embodiment and enfacement with realistic and cartoon avatars,” in *Proceedings of the 22nd International Conference on Mobile and Ubiquitous Multimedia*, 2023, pp. 270–278.

[34] L. González-Vides, J. L. Hernández-Verdejo, and P. Cañadas-Suárez, “Eye tracking in optometry: A systematic review,” *Journal of Eye Movement Research*, vol. 16, no. 3, pp. 10–16910, 2023.

[35] N. V. Valtakari, I. T. Hooge, C. Viktorsson, P. Nyström, T. Falck-Ytter, and R. S. Hessels, “Eye tracking in human interaction: Possibilities and limitations,” *Behavior Research Methods*, vol. 53, no. 4, pp. 1592–1608, 2021.

[36] J. R. Pauszek, “An introduction to eye tracking in human factors healthcare research and medical device testing,” *Human Factors in Healthcare*, vol. 3, p. 100031, 2023.

[37] M. Dorr, T. Martinetz, K. R. Gegenfurtner, and E. Barth, “Variability of eye movements when viewing dynamic natural scenes,” *Journal of vision*, vol. 10, no. 10, pp. 28–28, 2010.

[38] J. M. Wolfe and T. S. Horowitz, “What attributes guide the deployment of visual attention and how do they do it?,” *Nature reviews neuroscience*, vol. 5, no. 6, pp. 495–501, 2004.

[39] Y. Matsumura and K. Arai, “Influence of orthodontic appliances on visual attention to smiling faces by eye-tracking evaluation,” *Orthodontic Waves*, vol. 78, no. 4, pp. 135–142, 2019.

[40] J. Goller, A. Mitrovic, and H. Leder, “Effects of liking on visual attention in faces and paintings,” *Acta Psychologica*, vol. 197, pp. 115–123, 2019.

[41] A. Maffei and A. Angrilli, “Spontaneous blink rate as an index of attention and emotion during film clips viewing,” *Physiology & Behavior*, vol. 204, pp. 256–263, 2019.

[42] M. Pedrotti *et al.*, “Automatic stress classification with pupil diameter analysis,” *International Journal of Human-Computer Interaction*, vol. 30, no. 3, pp. 220–236, 2014.

[43] F. Ozkan and B. Ulutas, “Use of an eye-tracker to assess workers in ceramic tile surface defect detection,” in *2016 international conference on control, decision and information technologies (CoDIT)*, IEEE, 2016, pp. 088–091.

[44] C. Spence and S. Squire, “Multisensory integration: maintaining the perception of synchrony,” *Current Biology*, vol. 13, no. 13, pp. R519–R521, 2003.

[45] J. J. McDonald, W. A. Teder-SaÈlejaÈrvi, and S. A. Hillyard, “Involuntary orienting to sound improves visual perception,” *Nature*, vol. 407, no. 6806, pp. 906–908, 2000.

[46] H. McGurk and J. MacDonald, “Hearing lips and seeing voices,” *Nature*, vol. 264, no. 5588, pp. 746–748, 1976.

[47] P. Bertelson, “Ventriloquism: A case of crossmodal perceptual grouping,” in *Advances in psychology*, vol. 129, Elsevier, 1999, pp. 347–362.

[48] K. W. Grant and S. Greenberg, “Speech intelligibility derived from asynchronous processing of auditory-visual information.,” in *AVSP*, 2001, pp. 132–137.

[49] J. Vroomen and M. Keetels, “Perception of intersensory synchrony: a tutorial review,” *Attention, Perception, & Psychophysics*, vol. 72, no. 4, pp. 871–884, 2010.

[50] M. Meehan, S. Razzaque, M. C. Whitton, and F. P. Brooks, “Effect of latency on presence in stressful virtual environments,” in *IEEE Virtual Reality, 2003. Proceedings.*, IEEE, 2003, pp. 141–148.

[51] “Unity Engine: 2D & 3D Development Platform,” Unity. Accessed: Apr. 27, 2026. [Online]. Available: https://unity.com/products/unity-engine

[52] B. Foundation, “About,” Blender. Accessed: Apr. 27, 2026. [Online]. Available: https://www.blender.org/about/

[53] “FFmpeg.” Accessed: May 12, 2026. [Online]. Available: https://www.ffmpeg.org/

[54] M. S. Saccani, G. Contemori, F. Del Popolo Cristaldi, and M. Bonato, “Attentional load impacts multisensory integration, without leading to spatial processing asymmetries,” *Sci Rep*, vol. 15, no. 1, p. 16240, May 2025, doi: 10.1038/s41598-025-95717-0.

[55] S. E. Donohue, J. J. Green, and M. G. Woldorff, “The effects of attention on the temporal integration of multisensory stimuli,” *Front. Integr. Neurosci.*, vol. 9, Apr. 2015, doi: 10.3389/fnint.2015.00032.

[56] C. Li, Y. Dai, G. Chen, J. Liu, P. Li, and H. H. Ip, “Avatar-mediated communication in collaborative virtual environments: A study on users’ attention allocation and perception of social interactions,” *Computers in Human Behavior*, vol. 167, p. 108598, Jun. 2025, doi: 10.1016/j.chb.2025.108598.

A Appendix

A.1 Consent Form for the Experiment

Consent and agreements

I agree to voluntarily participate in a user study regarding audiovisual asynchrony.

I allow the session to capture eye movement, with the understanding that the data will only be used for analysis of the empirical study of my Master thesis.

I am informed that my personal data and test results will be treated confidentially and anonymously. No individual will be identified in the final thesis or any subsequent publications.

I am aware that I can withdraw and cancel my participation at any time without any consequences.

Lund, ____________ [Date]

Signature: _________________________

Name (printed): ____________________

A.2 Post-test Questionnaires for Experiment

Table A. 1 Post-test Immersion Questionnaire

Q1: To what extent did you feel present in the virtual concert hall? (7-point Likert Scale)
Q2: How natural and realistic did the audio environment sound to you? (7-point Likert Scale)
Q3: How confident are you in the synchrony judgments you just made? (7-point Likert Scale)
Q4: How difficult was it to determine if the audio and video were in sync? (7-point Likert Scale)
Q5: How much mental effort was required to perform the task? (7-point Likert Scale)