The effect of distance on audiovisual synchrony perception in virtual reality

Krol, P.

Award date:
2022

Department of Industrial Engineering & Innovation Sciences
Master of Science in Human-Technology Interaction

The effect of distance on audiovisual synchrony perception in virtual reality

by Pelle Krol

in partial fulfilment of the requirements for the degree of

Master of Science
in Human-Technology Interaction

Supervisors:
dr. ir. Raymond H. Cuijpers (IE&IS)
PhD. Candidate Victoria Korshunova (IE&IS)

Eindhoven, September 2021

Abstract

In our natural environment we are continuously confronted with all sorts of events that provide us with both auditory and visual signals. Generally, we perceive these auditory and visual signals as single multisensory events (King, 2005; Spence, 2007). However, it is easy to forget that our experience does not always correspond to physical reality. Due to the physical difference in velocity between light and sound, the sound takes a fraction of a second longer to reach our senses. This means that while *subjective* synchrony of audio-visual signals seems to (mostly) be the rule, *objective* synchrony of audio-visual signals is actually a very unlikely exception. Therefore, our brains must to some extent be tolerant to asynchrony or deploy mechanisms which allow us to consistently perceive synchrony (van Eijk, 2008; Keetels & Vroomen, 2012).

Apart from being relatively tolerant to asynchrony, researchers have hypothesized that the brain ensures a consistent perception of audiovisual synchrony by compensating for the longer travel times of auditory stimuli based on a visual judgement of distance. Sugita & Suzuki (2003) found that observers perceived audio-visual stimuli to be synchronous when the auditory part of the stimulus was increasingly delayed with distance at a rate that was roughly consistent with the speed of sound. This suggests that observers make an 'implicit estimation' of sound-arrival time based on the speed of sound. However, Lewald & Guski (2004) have criticized these findings. In a similar experiment, they found opposing evidence which suggests that observers simply perceive synchrony of audio-visual stimuli when the auditory and visual parts of the stimulus arrive at the observers' senses at approximately the same time. A more recent study by Silva et al. (2013) has proposed a slightly more nuanced view, indicating that the distance compensation mechanism may depend on the amount and quality of visual and auditory depth cues that are available.

For this thesis an experiment was developed using virtual reality, which aimed to measure the perceived synchrony of an audio-visual stimulus at different (virtual) distances. Participants were presented with an audio-visual stimulus at different stimulus onset asynchronies and distances and had to indicate whether the stimulus happened 'audio first', 'video first' or 'synchronous'. From these responses the point of perceived synchrony was calculated for each distance. Results support the idea that the brain simply integrates visual and auditory information that falls within a certain temporal window and showed no evidence of a compensation mechanism for sound transmission delays over distance,

finding that observers perceive synchrony of audio-visual stimuli when the auditory and visual parts of the stimulus arrive at the observers' senses at approximately the same time (supporting previous research by Lewald & Guski (2004)).

Preface

With this thesis, my objective was to find a challenging project where I would not only be applying my pre-existing skills, but also learn new ones. In this project I found a topic that fitted my interests in psychophysiology and also provided ample room for learning about programming, designing and interesting software tools like Unreal Engine and Cycling '74 Max. In the end, I even picked up a bit of Matlab during the data analysis. This thesis is the result of around 10 months of work, and has definitely been a significant challenge for me to complete. At times it was hard to keep up the work, but for the most part I enjoyed spending the days in the lab, building the experiment and learning new things. I think that these experiences will help me figure out my journey after graduation, and I am exited to see what is next.

I would also like to take this opportunity to thank my supervisors Victoria Korshunova and Raymond Cuijpers for meeting with me (almost) every week and being patient with me, letting me go at my own pace. I would like to thank Victoria in particular for always being open to questions and helping me throughout the design process of the experiment. Knowing his busy schedule, I also really appreciated Raymond took the time to work with me on the data analysis as I was completely new to Matlab. I would also like to thank Armin Kohlrausch who was incredibly helpful, helping me to connect my results to the literature in the later stages of the project. Finally, I would of course like to thank my friends, family and partner for all their support and advice. Without your encouragement and the occasional kick in the butt, my thesis would not have ended up how it did.

Contents

1. Introduction

In our natural environment we are continuously confronted with all sorts of events that provide us with both auditory and visual signals. Generally, these auditory and visual signals are integrated into a single multisensory representation (King, 2005; Spence, 2007). Let us take a common example: a friend calls your name from a distance. *'Hey you!'*. You see their lips move, hear the call and instantly recognize your friend called your name. It seems self-evident that we would perceive these auditory and visual signals as a single synchronous event. Your friend just yelled your name, you saw and heard it; what's so special about that?

It is easy to forget that our experience does not always correspond to physical reality. As illustrated above, the *subjective* synchrony of audio-visual signals seems to (mostly) be the rule. However, what happens when we take an objective look at our previous example? A friend calls your name from a distance. *'Hey you!'*. Since your friend is standing some distance away from you, the physical difference in velocity between light and sound causes the light to reach your eyes a fraction of a second earlier than the sound reaches your ears. This is a simple example which shows that, contrary to our *subjective* experience, the *objective* synchrony of audio-visual signals is not the rule, but rather a very unlikely exception.

The aim of audiovisual synchrony perception research is to fill the gap between objective and subjective synchrony. How can we perceive audio-visual signals to be synchronous when physical reality dictates that most audio-visual signals do not reach our senses at the same time? The rest of this introduction will serve to explain how physical and biological factors affect the timing of audio-visual signals, explore what methods researchers use to study audiovisual synchrony perception and most importantly to discuss research on the perceptual mechanisms that may allow us to perceive synchrony of audio-visual signals (with a focus on synchrony perception over distance).

1.1 Visual and auditory perception

To understand the process that leads us to perceive audio-visual synchrony, first we have to understand the basics of both visual and auditory perception. In its most basic form, visual perception refers to our ability to process and translate all the visible light that enters our eyes. Light is reflected by objects in the environment, enters our eyes through the cornea and is focused onto the retina by the lens. The retina is a light-sensitive membrane in the back of our eyes which consists of photoreceptive cells (rods and cones) which detect the particles of light (photons) and translate them into neural impulses through chemical transduction. These neural impulses are transmitted through the optic nerve into the primary visual cortex, which in term leads to the perception we call vision or eyesight.

Auditory perception refers to our ability to perceive sounds by detecting vibrations or periodic changes in the pressure of a surrounding medium (generally the air around us; although sound can be heard through solid, liquid or gaseous matter). Sound waves are picked up by our outer ears and propagate into our ear canal where they cause the eardrum to vibrate following the waveform of the sound. These vibrations are transmitted into the ossicles (three very small oscillating bones) which subsequently transmit the vibrations to the cochlea which is located in the inner ear. In the cochlea, the vibrations are translated into neural impulses through mechanical transduction. These neural impulses are transmitted through the auditory nerve into the auditory cortex, which in term leads to the perception we call hearing.

1.1.1 Intersensory timing

As mentioned before, simple physical factors have an influence on the relative timing between auditory and visual signals. At the physical level, arrival times of auditory and visual signals are affected by the distance between the source of the audiovisual stimulus and the observer. Since light and sound travel through the air at vastly different velocities (300.000.000 m/s for light compared to 343 m/s for sound), auditory and visual signals reach our senses at different times. As the distance increases, auditory signals lag more and more behind visual ones. At large distances these asynchronies can become quite obvious (i.e. lightning and thunder), but at closer distances we are rarely aware of them.

Interestingly, research has shown that the difference in arrival times of auditory and visual signals are partially compensated by biological differences. Because the aforementioned *mechanical* transduction of sound waves at the ear is slightly faster than the *chemical* transduction of light at the retina, there is a difference in response time between auditory and visual neurons of around 40-50 ms (King, 2005; Fain, 2019). In practical terms, this means the difference in travel time of light and sound are cancelled out when a source is around 10-15 meters away from the observer. Researchers have termed this distance

the 'horizon of simultaneity' (Pöppel et al., 1990). Within the limits of this horizon, the auditory signal arrives in our brains first whereas beyond this horizon, the visual signal takes precedence.

It is clear, however, that this biological factor cannot explain how our brains are able to perceive audiovisual synchrony over a large range of distances. The fact remains that as distances varies, so does the asynchrony. Furthermore, this hypothesis rests on the assumption that the physical simultaneity of neural signals in the brain is meaningful to synchrony perception, but this seems unlikely considering brain activity consists of the recurrent firing of neurons. It seems more likely that neural networks in our brains are calibrated to recognize patterns in neural activation as simultaneity, rather than determining simultaneity through exact onset times of said neural activity. It is an inescapable fact that physical factors will almost always cause auditory and visual signals to arrive at our senses at different times. Therefore, our brains must to some extent be tolerant to asynchrony (van Eijk, 2008) or deploy other underlying mechanisms which allow us to consistently perceive synchrony (Keetels & Vroomen, 2012).

1.2 Measuring perceived synchrony

In order to study the underlying mechanisms of audiovisual synchrony perception, researchers have designed several synchrony judgement tasks. All of these tasks aim to measure the point of subjective simultaneity (PSS), which is defined as the relative auditory delay between the components of an audiovisual stimulus at which the perception of synchrony occurs. Typically, researchers measure and compare the PSS in different experimental conditions. By studying how PSS values change in different conditions, researchers can make inferences about the way we perceive synchrony. By convention, positive PSS values indicate that the auditory component is lagging behind the visual component, while negative PSS values indicate that the auditory component is ahead of the visual component. If the PSS is equal to zero, this means that the onset of the components of the audiovisual stimulus was physically synchronous.

The synchrony judgement tasks that are most commonly used to determine the PSS of an audiovisual stimulus are the temporal order judgement task (TOJ), binary synchrony judgement task (SJ-2) and ternary synchrony judgement task (SJ-3). The basic experimental design for each of these judgement tasks is very similar, they only differ in the specific judgements that subjects are asked to make. First, researchers show a short audiovisual stimulus to a subject. After the stimulus is shown, the subject is asked to make a judgement about the relative synchrony of its visual and auditory components. Depending on the exact experimental design, this process is repeated a good deal of times. However, each time the audiovisual stimulus is shown, the onset times of its visual and auditory components (the stimulus onset asynchronies, or SOA's) are slightly changed. SOA val-

ues follow the same convention as PSS values; positive SOA's indicate an audio lag while negative SOA's indicate an audio lead. By performing repeated synchrony judgements at different SOA's, researchers collect data that shows at which relative onset asynchronies the subject perceives synchrony of the audiovisual stimulus.

In a TOJ task, the subject has to indicate whether the audiovisual stimulus happened 'video first' or 'audio first'. After repeated measurements, researchers are able to calculate response curves which show the proportion of 'video first' and 'audio first' responses at different SOA's (see Figure 1.1). The PSS in a TOJ task is subsequently defined as the intersection between the video first and audio first response curves (the 50% point). Since a TOJ task is about judging temporal *order*, it provides a relatively indirect measurement of perceived *synchrony* (perceived synchrony is only inferred at the intersection of audio first and video first responses).

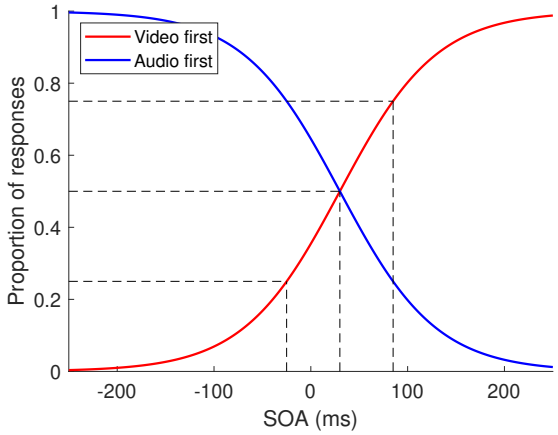

Figure 1.1. An example of TOJ data. Shown are the response curves for 'video first' (red) and 'audio first' (blue) judgements. The dashed lines indicate the PSS at the intersection between the video first and audio first curves (50% point) and the synchrony range between the 25% and 75% response proportions.

The SJ-2 task provides a more direct measurement of perceived *synchrony* by letting the subject judge if the stimulus components were presented 'synchronous' or 'asynchronous'. This produces two response curves which can be seen in Figure 1.2A. The points where synchronous and asynchronous response curves intersect (50% point) are termed the synchrony boundaries, the range between these boundaries is the synchrony range. The PSS is defined as the midpoint of the synchrony range. The SJ-3 task is a variation of the SJ-2 task where the subject has the option to choose either 'video first', 'synchronous' or 'audio first'. An example of the resulting response curves for this method can be found in Figure 1.2B. The synchrony boundaries are defined at the audio first-synchronous and video first-synchronous intersection points. Similarly to the SJ-2 data, the range between these boundaries is termed the synchrony range and the PSS is defined as the midpoint of this range. The SJ-3 task provides a slightly different measurement of perceived synchrony

compared to the SJ-2 task, as it includes the same direct 'synchronous' measurement but splits the 'asynchronous' measurement into the two specific sides of 'audio first' and 'video first'. In that sense, the SJ-3 task is a combination of the TOJ and SJ-2 tasks, which to some extent measures both perceived *order* and perceived *synchrony*.

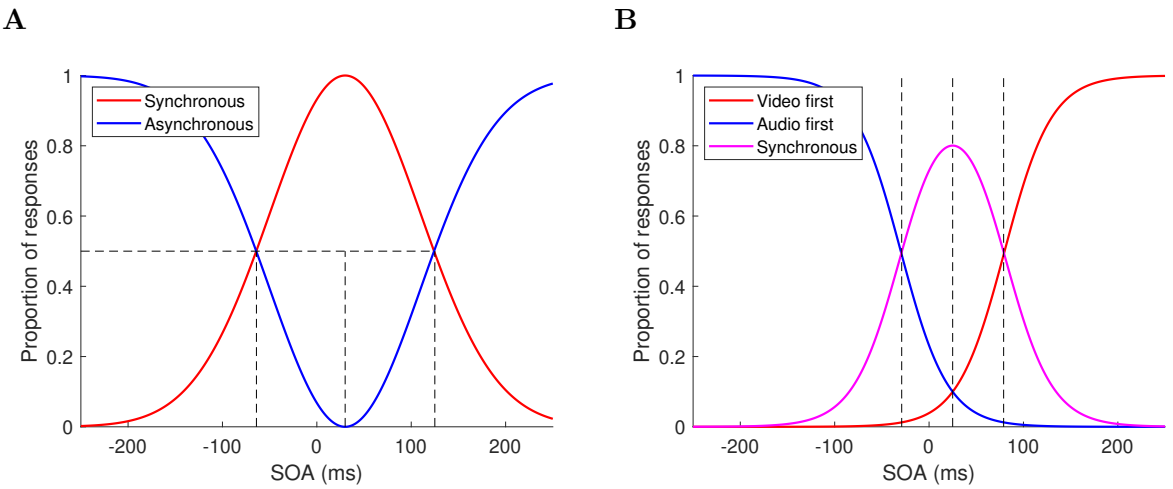

Figure 1.2. An example of SJ-2 and SJ-3 data. (A) The SJ-2 graph contains response curves for 'synchronous' (red) and 'asynchronous' (blue) judgements. Dashed lines show the synchrony boundaries at the intersection points of the synchronous and asynchronous curves and the PSS at the midpoint of this synchrony range. (B) The SJ-3 graph contains response curves for 'video first' (red), 'audio first' (blue) and 'synchronous' (pink) judgements. Same as the SJ-2 data, the dashed lines show the synchrony boundaries at the audio first-synchronous and video first-synchronous intersection points and the PSS at the midpoint of this range.

1.3 Audiovisual synchrony perception

1.3.1 Temporal window of integration

Through experimentation, researchers have found clear evidence that the brain operates with a wide temporal window of integration (TWI; (Spence & Squire, 2003; Vatakis, 2013)). The TWI represents the temporal range within which our brains tolerate the asynchrony of audiovisual stimuli and thus integrates them into a multisensory event which is perceived as synchronous (see Figure 1.3).

Research has found that the range of asynchronies where temporal integration is effective is around 50-100 ms (Lewald & Guski, 2003). For SOA's below 40-50ms, evidence even suggests that people might not be able to accurately judge temporal order at all (Spence et al., 2003). Furthermore, people have been found to be more tolerant to visual leads

compared to auditory leads, which suggests the TWI has a bias towards the more naturally occurring visual leads and auditory lags. In other words, the TWI is characterized by a visual bias. This means that in order for synchrony to be perceived, the auditory stimulus generally has to lag behind the visual stimulus (Lewald & Guski, 2003; Vatakis, 2013). Other parameters like spatial proximity of stimuli, ecological validity of stimuli and availability of depth cues have been found to affect the way audiovisual information is integrated (and therefore affect synchrony perception) (Lewald & Guski, 2003; Kohlrausch et al., 2013; Silva et al., 2013). Since the asynchrony between the visual and auditory parts of a stimulus grows with distance, the aforementioned wide TWI can only explain synchrony perception up to a certain point. Interestingly though, recent research suggests the brain might employ a compensatory mechanism for the difference in arrival time between light and sound which is based on the distance to the source, allowing for the perception of synchrony even at large distances.

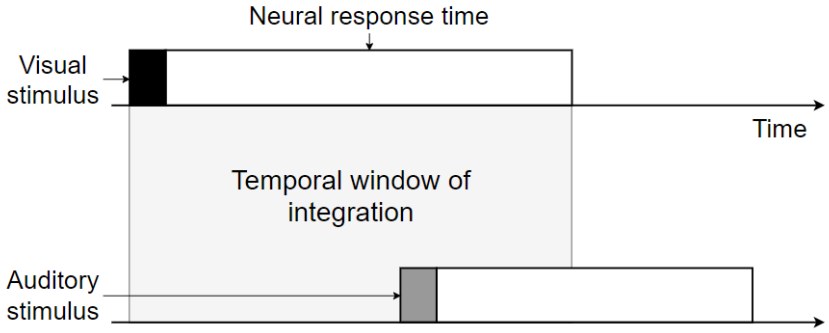

Figure 1.3. A schematic representation of the temporal window of integration, redrawn from Spence & Squire (2003).

1.3.2 Distance compensation

Researchers have hypothesized that the brain ensures a consistent perception of audiovisual synchrony by compensating for the longer travel times of auditory stimuli based on a visual judgement of distance. In a study by Sugita & Suzuki (2003), participants were asked to judge the temporal order (TOJ task) of a visual (light flash) and auditory (white noise burst) stimulus. The light flashes were produced by LED's placed at different distances (1-50 m) from the participant while the white noise bursts were presented through headphones. Importantly, the sound was processed using a head-related transfer function in order to simulate a frontal origin, but not changed over distance. Furthermore, the intensity of the light flash was increased with distance in order to produce a consistent intensity at the eye. SOA's between the light flash and white noise burst were varied randomly *at the observer*, between -125 ms (audio leading) and +175 ms (audio lagging) in step increments of 25 ms. Results showed that the PSS increased with distance in a way that was roughly

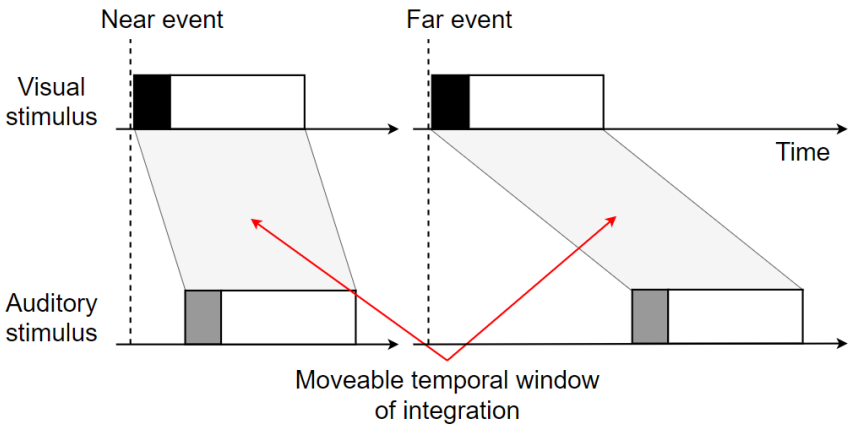

Figure 1.4. A schematic representation of the proposed 'moveable temporal window of integration' (Sugita & Suzuki, 2003), redrawn from Spence & Squire (2003).

consistent with the speed of sound. In other words, participants perceived synchrony at points that were increasingly auditory lagging with distance. This led Sugita & Suzuki to the conclusion that the TWI is moveable due to an 'implicit estimation' of sound-arrival time based on the speed of sound (see Figure 1.4). This implies that we perceive synchrony at increasing audio lags because the brain infers a *synchronous onset at the source*.

Other research by Alais & Carlile (2005) has supported these findings, suggesting that observers were taking account of the distance of the sound source and attempting to compensate for travel time with a subjective estimate of the speed of sound.

The research by Sugita & Suzuki has also been criticized however, with Lewald & Guski (2004) arguing that the experimental procedure might have been problematic since it was inconsistent with the natural occurrence of audiovisual events. In the procedure used by Sugita & Suzuki the intensity of the light and sound were kept constant over distance, while in reality the light and sound intensity should decrease with distance. Therefore, Lewald & Guski argued that the experiment did not reflect physical reality since subjects were not provided with realistic auditory and visual information. This led Lewald & Guski to conduct a similar experiment using a less artificial procedure, attempting to replicate the results of Sugita & Suzuki. This experiment took the same basic approach: subjects had to judge the temporal order of a auditory and visual stimulus at distances ranging from 1-50 meters. To make the experiment more ecologically valid however, the experiment was performed outside on a university lawn and set up so that the auditory and visual stimuli were co-located (a speaker with a LED in front of it). The intensities of the auditory and visual stimuli were kept constant at the source, which meant the intensity at the observer would naturally decrease with distance. SOA's were varied between -200 ms (auditory lagging) and +200 ms (auditory leading; note that a positive value here represents an audio leading stimulus and a negative value an audio lagging stimulus, this is opposite to Sugita & Suzuki). They found that the PSS values increased with distance in a way that

was roughly consistent with the speed of sound. However, due to the flipped PSS values, in this case this means that participants perceived synchrony at points that were increasingly audio leading with distance. Interestingly, these findings were completely opposite to those of Sugita & Suzuki. The results imply that the brain does not compensate for sound velocity in order to infer a synchronous onset at the source, but rather perceives synchrony when stimuli reach the observer at approximately the same time. Therefore, Lewald & Guski concluded that the underlying perceptual mechanism simply works by the integration of stimuli in a wide time window rather than an implicit estimation of sound velocity.

Other research has come to similar conclusions as Lewald & Guski. In a causal attribution task study, Arnold et al. (2005) found that there was no perceptual mechanism to compensate for the difference in velocity between light and sound. Arnold et al. suggest that visual and auditory stimuli that reach the observer at approximately the same time become perceptually bound, even when they might not have originated from the same (distant) event. They added that a compensation mechanism for sound velocity has only been found in the literature when participants were explicitly told to imagine the visual and auditory stimuli had originated from a common source. Therefore, they conclude that the origin of compensation found by Sugita & Suzuki might not be perceptual, but rather cognitive.

1.3.3 Effect of depth cues on distance compensation

A more recent study by Silva et al. (2013) has proposed a slightly more nuanced view, not fully agreeing with either Sugita & Suzuki (2003) or Lewald & Guski (2004). Silva et al. argues that Lewald & Guski did not provide optimal auditory depth information as they conducted the experiment outside in free-field conditions. This meant that one of the most powerful auditory depth cues, the ratio of direct and reflected sounds, was not present which could have caused misjudgement of stimulus distance. Furthermore, Silva et al. argues that by using artificial stimuli (flashes and noise bursts) two other important depth cues were omitted, namely the familiar loudness of the auditory stimulus and familiar size of the visual stimulus. In their own study, Silva et al. (2013) conducted a synchrony judgement experiment where they used a more familiar biological motion stimulus. Subjects had to judge the synchrony of the visual movement of a person making a step and the corresponding step sound. A walking movement was recorded through motion capture and translated into a simple dotted representation of the movement (see Figure 1.5). The visual stimulus was projected onto a wall while the auditory stimulus was presented through headphones. Audiovisual depth cues were simulated through a binaural sound which was pre-recorded in a large room at different distances. Visual depth cues were provided by including perspective depth frames (see Figure 1.5) and cues like familiar size, elevation and angular velocity. These auditory and visual stimuli were presented at SOA's ranging from -240 ms to +300 ms in step increments of 30 ms at distances ranging from 10-35 meters.

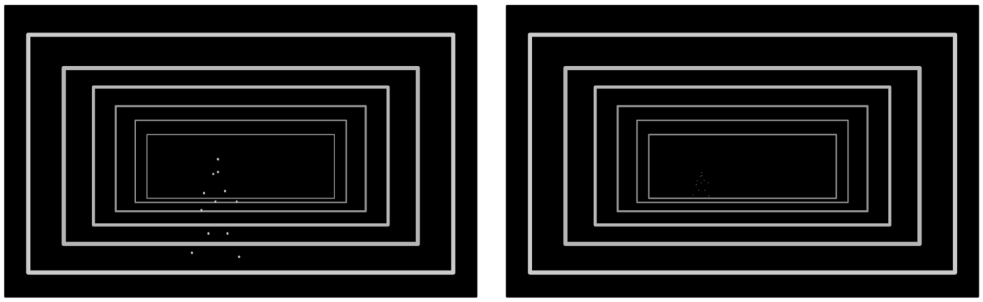

Figure 1.5. The motion capture representation of a person walking with the perspective depth cues (rectangular windows) as seen in Silva et al. (2013).

In order to study the effect of the visual and auditory depth cues, 'audiovisual depth cues', 'visual depth cues' and 'reduced depth cues' conditions were compared. In the 'audiovisual depth cues' condition, both the auditory and visual depth cues were presented to the participant. In the 'visual depth cues' condition, the visual depth cues were presented as before while the auditory depth cues were reduced to a simple free field recording with only directional cues. In the 'reduced depth cues' condition, the perspective depth frames and other visual depth cues were also removed. In general, results showed that the PSS shifted to increasing audio lags with distance. This means that the results supported the existence of a distance compensation mechanism in the brain. Furthermore, Silva et al. found a significant difference between the audiovisual, visual and reduced depth cues conditions. The more cues were present, the more the PSS shifted in the direction of larger audio lags. In the 'audiovisual depth cues' condition, the average shift in PSS was close to the actual physical delay at these distances which supports Sugita & Suzuki's theory that the brain makes an implicit estimation of sound arrival time. However, in the 'reduced depth cues' condition, no such shift in the PSS was found. This led Silva et al. to conclude that the relative weights of the cues for the distance compensation mechanism may depend on the amount and quality of depth cues that are available.

The differences between the methods and subsequent outcomes of the studies by Sugita & Suzuki (2003), Lewald & Guski (2004) and Silva et al. (2013) are intriguing. Given the contradictory results, it is still not clear if and how distance compensation plays an active role in audiovisual synchrony perception. In terms of setup of the stimuli, all three studies took slightly different approaches. Sugita & Suzuki presented the visual stimulus in physical space in front of the subjects and presented the auditory stimulus through headphones. Lewald & Guski set up the experiment so that both the visual and auditory stimulus were co-located in physical space and Silva et al. created a virtual representation of the visual stimulus and presented the auditory stimulus through headphones. Because of these differences in experimental setup, SOA's in Sugita & Suzuki and Silva et al. are defined at the *observer* while SOA's in Lewald & Guski are defined at the *source*.

This means that in the former two studies, a synchronous onset at the source is only implied and the travel time (speed of sound) of the auditory stimulus is not modelled into

the stimulus. While the studies of Sugita & Suzuki and Silva et al. might be criticized in terms of ecological validity compared to Lewald & Guski, apart from a slight difference in calculation these differences should not have led to the change in results. Considering the results of Silva et al. it is more likely a coincidence that both studies that employ this specific method come to results which show a distance compensation mechanism, while the seemingly more ecologically valid method does not. Another difference that Silva et al. notes is that the study by Lewald & Guski does not provide decent auditory depth cues, which is the result of the experiment taking place outside. Since Silva et al. do find a distance compensation mechanism when additional auditory depth cues are present, it could be hypothesized that cue weightings will naturally be different indoors compared to outdoors (due to the different acoustical properties) which causes the difference in results between the studies.

All together these studies give ample reason to further research exactly if and how the distance compensation mechanism works. This thesis will attempt to contribute to the literature by studying distance compensation through virtual reality (VR) to establish whether this is a useful technology for further audiovisual perception research.

1.4 Audiovisual perception in VR

Through the differences between the earlier studies by Sugita & Suzuki (2003) and Lewald & Guski (2004) and the later study of Silva et al. (2013) we can see that the materials that researchers have used in audiovisual synchrony perception research have evolved over the years. In general it can be observed that stimuli are increasingly being presented through virtual means. However, when working with virtual representations care should be taken to ensure that visual and auditory stimuli are representative of real world conditions. Over the years, methods for visualization and auralization of stimuli have drastically improved, already allowing Silva et al. to create accurate and realistic manipulations of both the auditory and visual depth cues that were presented to subjects. Since the publishing of this particular paper the possibilities have expanded even more with the introduction of the first consumer grade VR devices in early 2016, like the Oculus Rift, HTC Vice and PlayStation VR. The popularity of VR has been growing ever since which has opened doors to many commercial, educational and scientific applications. For research purposes specifically, it is extremely compelling to use VR. While the fully natural setup of Lewald & Guski (2004) is great in terms of ecological validity, it would be quite challenging to set up and operate for most researchers. Furthermore, conditions in the real world are hard to control and therefore hard to consistently reproduce. VR offers a perfect 'best of both worlds' option, where a virtual environment can be designed to be both realistic and deterministic, providing more ecological validity while keeping all the control of a laboratory environment. Additionally, the existence of VR itself opens new doors for research since VR is not merely a tool, but also a topic of research in and of itself. To

improve the immersiveness of VR, it is critical to understand the intricacies of (audiovisual) perception since perceived realism in VR is significantly dependent on the availability and congruence of visual and auditory information (Jeon & Jo, 2020; Lindquist et al., 2016).

1.4.1 Distance perception in VR

Research has shown that in VR applications that use head mounted displays (HMD's), subjects consistently under-perceive modelled distance. In a review paper of several VR studies, Renner et al. (2013) found that on average, egocentric distance was underestimated by 27% compared to modelled distance in virtual environments displayed through a HMD. Another study by Shemetova & Bodenheimer (2014) found an underestimation of 13%. In general, real world egocentric distance estimations have been found to follow a power law with different studies finding slight under- or overestimation dependent on the setting and measurement procedure (Teghtsoonian & Teghtsoonian, 1970; Da Silva, 1985). Other studies have found distance estimations followed a logarithmic function, where subjects tend to be accurate or slightly overestimate up close and increasingly underestimate as distances get bigger (Gilinsky, 1951; Foley, 1980; Loomis et al., 1996). However, for limited ranges (like in VR) a linear relationship might be more appropriate. On the whole, egocentric distance perception in VR seems to approximate real world distance perception. Renner et al. also noted that the accuracy of egocentric distance judgements in VR improved with the quality of the visuals and complexity of the environment. This means that as the quality of VR hardware and software improves, the difference between real and virtual world distance perception might shrink. This theory is supported by the analysis and results of Feldstein et al. (2020), who found that on average egocentric distance was only underestimated by 4% in a virtual environment while it was underestimated by 6% in a similar real environment. Analyzing studies over the years, they found that as state of the art VR technology has evolved, the underestimation has become less pronounced.

In terms of auditory perception of distance, little work has been done to test depth perception in virtual environments. However, research has shown that sounds can effectively be spatialized by applying head related transfer functions (HRTF's) Begault et al. (2001) and other dynamic implementations of sound spatialization like Ambisonic systems (Gerzon, 1973). Kearney et al. (2012) found that localization and perception of distance of real sounds was comparable to spatialized audio played through headphones. Auditory distance perception has been shown to be comparable to visual perception in that subjects tend to slightly overestimate up close and increasingly underestimate as distances get bigger (Kearney et al., 2012; Zahorik et al., 2005).

The fact that there might be slight differences between visual and auditory (depth) perception in virtual and natural environments has some interesting implications for audiovisual synchrony perception research in VR. If a distance compensation mechanism (Sugita & Suzuki, 2003; Silva et al., 2013) exists which is dependent on depth cues and

distance perception, synchrony perception might differ significantly between virtual and natural environments which should subsequently be considered when designing virtual environments.

1.5 Research aims

While there is some data on visual and auditory perception in virtual environments, at this moment there is a lack in research on the topic of audiovisual *synchrony* perception in VR. This means there is an opportunity to leverage this new technology to simultaneously research audiovisual synchrony perception and improve virtual environments, as VR offers more independent control over ecological validity, visual and auditory delays and stimulus intensity. To start off, researchers need to understand how audiovisual synchrony experiments in VR measure up to previous research. If experiments in VR turn out to mirror or approximate results found in real life, effects that were observed in virtual environments can be translated to real life and vice versa.

This study will focus on setting up an experiment similar to that of Lewald & Guski (2004) in a virtual environment, measuring the PSS of a co-located audiovisual stimulus at distances ranging from 10 to 50 meters. The experiment will use simple stimuli (light flashes and sound bursts) and sounds will be presented with natural arrival time delays based on the speed of sound. Therefore, SOA's will be defined at the source. Only relative loudness depth cues will be programmed into the sound to approximate the setup of Lewald & Guski. This study is meant to be part of a larger ensemble of experiments that aim to investigate audiovisual synchrony perception in VR in different conditions. The current research will focus on answering the following research question:

RQ: What is the effect of distance on judgements of synchrony of audiovisual events in a virtual outdoor environment?

From the research of Sugita & Suzuki (2003), Lewald & Guski (2004) and Silva et al. (2013), two opposing answers to the research question can be predicted:

Prediction 1: Assuming that observers compensate for the longer travel times of auditory stimuli based on a visual and auditory judgement of distance, the PSS at the observer will shift toward increasing audio delays with distance.

Prediction 2: Assuming that observers do not compensate for distance and thus perceive synchrony when audiovisual signals reach their senses at roughly the same time, the PSS at the observer will stay approximately constant with distance.

Additionally, egocentric distance judgements to the visual stimulus will be measured to determine the relationship between modelled and perceived distance in the virtual environment. Furthermore, if prediction 1 were to be true, it would be expected that these egocentric distance judgements can be used to investigate possible effects of a visual distance estimation on synchrony perception.

2. Method

In a virtual environment, participants were placed at different distances from a tall pole. At the top of this pole, a loudspeaker and light bulb were placed to serve as the sources for a simple audiovisual stimulus (siren burst and light flash). Each trial, participants saw a short flash of light which was preceded or followed by a siren burst sound according to different onset asynchronies. After each trial, participants performed a synchrony judgement task (SJ-3) where they had to indicate whether the audiovisual stimulus had happened audio-first, synchronous or video-first (van Eijk (2008); Kohlrausch et al. (2013). In order to take away uncertainty of when each next trial would take place, participants were given a visual cue providing anticipation time for when the light flash would happen. Each trial the speaker would do a single spin around the pole and when the speaker faced the participant straight on, the light flash would be presented. Since this made the flash predictable, it should have allowed participants to focus better on the flash's (a)synchrony with the siren sound.

2.1 Experimental Design

The experiment used a 5×11 within-subjects design with distance to audiovisual event $\times$ stimulus onset asynchrony as the independent variables. The measured dependent variable was the response to an SJ-3 task which indicated perceived stimulus order expressed in audio-first, synchronous or video-first. All 55 conditions were measured in a fully crossed design and repeated 30 times. This made for a total of 1650 trials which were presented in a completely randomized order. The experiment was split into two sessions with 825 trials each. Furthermore, each session was split into five blocks of 165 trials to allow for small breaks. The five distances were 10, 20, 30, 40 and 50 meters, based on previous research by Sugita & Suzuki (2003) and Lewald & Guski (2004). The stimulus onset asynchronies between the siren sound and flash were varied with values that ranged from -250 ms (audio-first) to $+250$ ms (video-first) in 50 ms step increments. These values were based on previous research and adjusted to cover a wide range of delays while keeping the amount of trials limited (Sugita & Suzuki, 2003; Lewald & Guski, 2004; Kohlrausch et al., 2013; Silva et al., 2013).

2.2 Participants

Participants were recruited using the JFS Participant Database of the Human-Technology Interaction department of the Eindhoven University of Technology and through word of mouth. In total, eight participants took part from which four were male and four were female ($M = 29.13$, $SD = 7.47$). Each participant gave their informed consent before participating in the experiment and was compensated for their time according to university policy. Only participants with (corrected to) normal eyesight and normal hearing participated.

2.3 Materials

The experiment took place in a university lab. A VR setup was placed on a large desk (see Figure 2.1). The VR setup consisted of a PC, a monitor and the first generation Oculus Rift headset with sensors and Oculus Touch controllers. Participants were seated in an office chair facing the desk and were able to point and click on their answers in the experiment by using the Oculus Touch controllers (see Figure 2.2). The PC ran Windows 10 on a Intel® Core™ i9-9900K processor, with the NVIDIA® Titan RTX™ graphics card and a Realtek® LC1220P-VB2 sound card. The virtual environment was created using Unreal Engine (version 4.27.1) and displayed through the Oculus Rift head mounted display (HMD). The stimulus sound was created using Cycling '74 Max (version 8.1.11) and played through the Oculus Rift on-ear headphones.

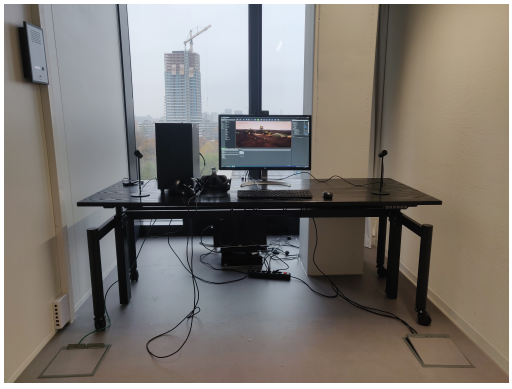

Figure 2.1. VR setup in the lab.

Figure 2.2. HMD, controllers and sensors.

2.3.1 Virtual environment

Participants were placed in a virtual environment with a desert setting. This setting was chosen due to its quiet and uncluttered nature, lacking large trees or other random visual

distance cues which could potentially play as a distractions. In this desert environment, a military encampment was created using assets downloaded from the Unreal Engine Marketplace (*Military Field Camp 3.2*, 2016). Participants were positioned in a large open area within the encampment which was surrounded by tents, watchtowers and other military appliances (see Figure 2.3). The encampment was designed to provide some context and depth cues (as opposed to a flat empty world), without creating an environment which could conceivably have a large acoustic effect, like reverberations from walls or buildings. The time of day was set to be just after sunset so that the used visual stimulus (a flashing light) and its reflection on the ground would be easier to see against a dark background (see Figure 2.4).

Distances used in the experiment were defined as the distance between the audiovisual source and the observer. All stimulus distances were programmed into the virtual environment using Unreal Engine's internal Unreal Units (uu), which are supposed to represent the equivalent of one centimeter (1uu=1cm). However, the distances were not calibrated which means some scaling factors are to be expected. During the experiment the position of the participants was fixated, but participants were able to look around freely in 360 degrees at all times.

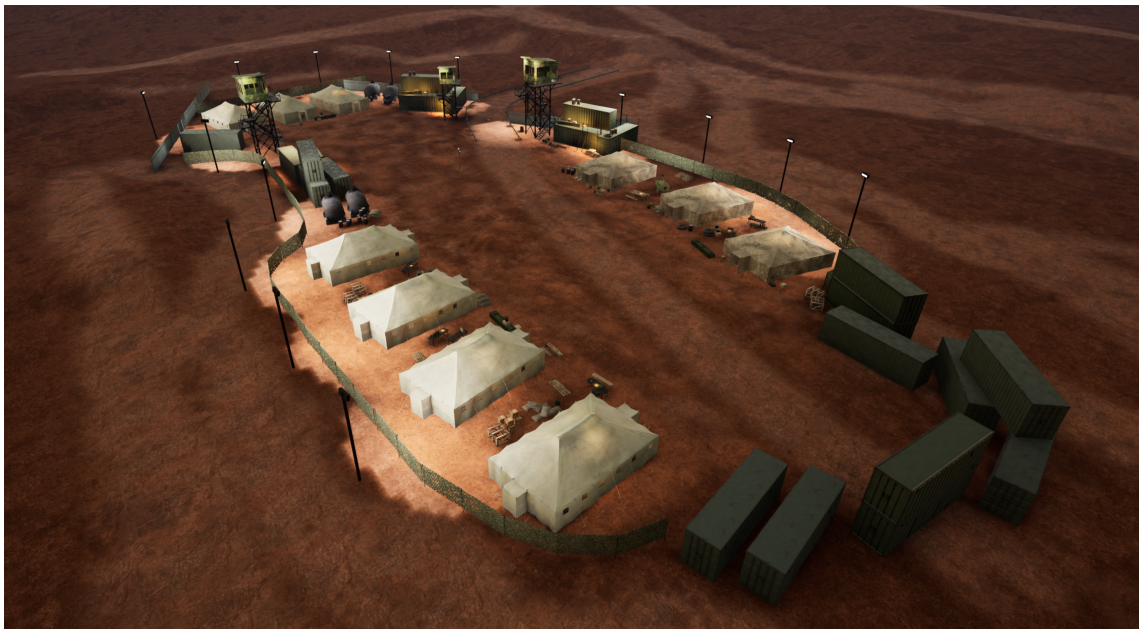

Figure 2.3. Overview of the virtual environment.

Figure 2.4. Participant's POV at 10 meters with the light turned off (left) or on (right).

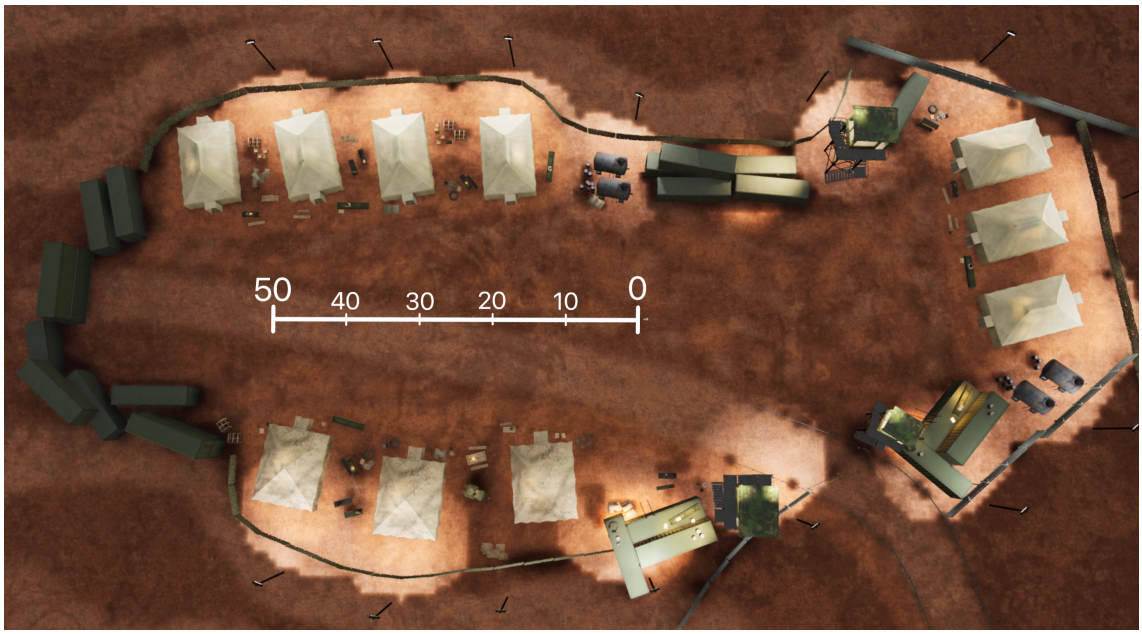

Figure 2.5. Top down view of the virtual environment, including an overview of the position of the observer at each distance condition.

2.4 Stimuli

The visual stimulus was produced in Unreal Engine using a 6500K point light with an intensity of 2000 lumen, which flashed on for 200ms. To produce the auditory stimulus, Unreal Engine was hooked up to a Cycling '74 Max patcher, which generated a 200ms long, exponentially decaying sound burst that resembled a siren sound. The reference sound pressure level was attenuated over distance. The difference in sound pressure level (ΔL, in dB) between distances (r) was calculated using the Inverse Distance Law:

$$\Delta L = 20 \cdot \log_{10} \frac{r_1}{r_2}$$

In practical terms, this resulted in a 6dB reduction in sound pressure level with each doubling of distance. Furthermore, an onset delay was added to the sound to simulate the travel time over each distance (r) assuming a temperature of 40°C (speed of sound $v_{sound} = 355\text{m/s}$):

$$\text{Onset delay} = \frac{r}{v_{sound}}$$

Both the light and sound were triggered through Unreal Engine, which controlled the inputs for each trial. The light was triggered after one full turn of the speaker (360°) when the speaker faced the participant straight on, while the sound was triggered at an offset from the light based on the SOA of the trial. Using the rotations per minute ($RPM = 20$) of the speaker and SOA (in seconds) value of the trial, the offset in degrees ($\Delta\alpha$) where the sound should be triggered was calculated by converting the RPM to degrees of rotation per second and multiplying this value with the SOA (in seconds):

$$\Delta\alpha = 360 \cdot \frac{\text{RPM}}{60} \cdot \text{SOA}$$

To make the trigger resistant to Unreal Engine's internal delays, an error margin of one degree was used for the trigger (equivalent to an error margin of 5 ms). A detailed flowchart of the internal logic of the Unreal code and Max patcher can be found in Appendix C.

2.5 Measurements

The most important measurement in this experiment is the point of subjective simultaneity (PSS). The PSS was measured using a SJ-3 task where the participant gives a judgement

of the simultaneity of the audiovisual stimulus after each trial (audio first, synchronous, video first). PSS values were measured at the source, which means positive values indicate a visual leading stimulus and negative values a audio leading stimulus. The PSS was defined as the midpoint of the range of delays that are predominantly judged to be synchronous (van Eijk, 2008). These PSS values were then compared over distance in order to find if sound delays caused by distance had an effect on perceived synchrony. Furthermore, egocentric distance to the audiovisual stimulus and subjective size of elements of the environment were measured using a verbal judgement task. At the start of the experiment, participants had to give judgements of each of the five distances that were programmed into the virtual environment. At the end, participants were asked to give size judgements of three elements of the environment (pole height, tent width and tower height). This was done to assess the ratio between the perceived and modelled distances and sizes in the virtual environment. Additionally, a short questionnaire was taken at the beginning of the experiment to record age, gender, VR experience, hearing and eyesight.

2.6 Procedure

Participants registered for two sessions with a duration of two and a half ours each. Upon arriving at the lab, participants were sat down at the desk and instructed to carefully read the informed consent forms. After signing the forms they filled in their contact details as required by the university's COVID-19 protocols. To start the experiment, participants filled in a short questionnaire about their age, gender, experience using VR and the condition of their hearing and eyesight. Using a Landolt C chart, a short eye test was conducted by the experimenter to confirm the participant's eyesight was sufficient to do the experiment. After these preliminary steps, the experimenter showed a short video of the experiment to get the participant familiar with the task. The experimenter then proceeded to help the participant put on the VR headset by explaining how to get a comfortable fit, get a sharp image and how the controller worked. When the headset was properly fitted, the experiment was booted up.

First, the experimenter instructed the participant to do a preliminary distance judgement task. This task was intended to get a baseline measurement of perceived distance to the audiovisual stimulus in VR, but also served as a way for the participant to get familiar with the environment. The participant was placed at each of the five stimulus distances in randomized order and asked by the experimenter to give a verbal judgement of the distance between them and the pole. Results for each of the distances were recorded by the experimenter. After completing the distance judgements, participants were instructed to proceed to the main body of the experiment. Participants first performed 30 practice trials of the audiovisual synchrony task, to get acquainted with the stimulus onset asynchronies and answer options (audio-first, synchronous, video first). After the practice trials the experimenter checked in with the participant, gave instructions to start the experimental

trials and left the room.

One session consisted of 825 total trials which were divided into five blocks of 165 trials. After each block the participant was given the option to take a break to prevent fatigue and nausea caused by the VR headset. Upon completing all trials, participants were prompted to take off the headset and make three verbal size judgements of the objects in the virtual environment (pole height, tent width and tower height) based on memory. After the second session participants were debriefed and compensated for their time according to university policy.

3. Results

3.1 PSS data

Response data from all eight participants was initially analyzed separately, first calculating the proportions of 'audio first', 'synchronous' and 'video-first' responses at each of the five distance conditions (see Figure 3.1A). Subsequently, logistic functions were fitted to the proportion data of the 'audio first' and 'video first' responses. The logistic functions were defined as:

$$\frac{L}{1 + e^{-k(x-x_0)}}$$

where L is the maximum value of the curve, k is the logistic growth rate and x_0 is the midpoint of the curve. A function for 'synchronous' responses was calculated by subtracting the sum of the 'audio first' and 'video' first responses from 1. These three functions were fitted as an ensemble using the Nelder-Mead simplex method (Nelder & Mead, 1965) via Matlab's $fminsearch$ function as described by Mareschal et al. (2013) (see Figure 3.1B). The intersection points of the audio first, synchronous and video first curves were taken as the left (L) and right (R) synchrony boundaries. The range of SOA's between these synchrony boundaries was defined as the synchrony range and the PSS was subsequently defined as the midpoint of this range.

All SOA values were corrected for the internal visual and audio latencies of the VR setup (see Appendix A). These latencies were measured after the experiment had already taken place, so were not taken into account in the design process. This means that all SOA's were effectively shifted by approximately +90 ms, resulting in a 'real' SOA range of -160 to +340 ms in stead of the designed -250 to +250 ms. Unfortunately, this meant the data contained a relatively low amount of audio first responses in the 40 and 50 meter distance conditions (because the audio was simply not leading enough to induce an 'audio first' response) which caused problems with model fit (see Appendix B). To solve these problems, the audio first data was assumed to have the exact opposite logistic growth rate (k) to the video first data. This assumption is not recommended by other research (Alcalá-Quintana & García-Pérez, 2013) as both curves tend to have slightly different characteristics, but was a necessary compromise to ensure a more consistent fit across distance conditions.

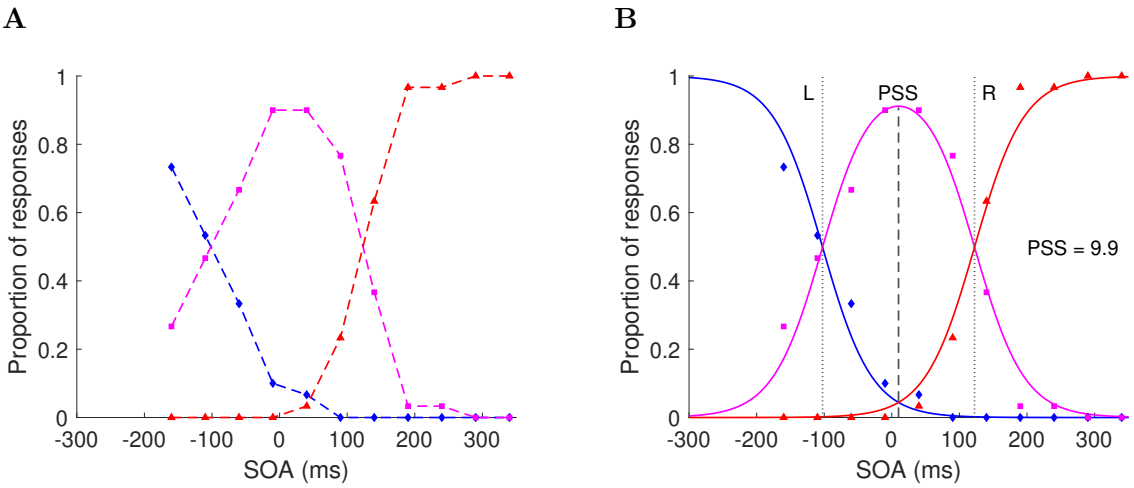

Figure 3.1. (A) Example response data at a distance of 10 meters. Shown are the audio first (◆), synchronous (■) and video first (▲) response proportions for each of the measured SOA's. (B) The logistic model fit to this data. Solid lines show the fitted audio first (—), synchronous (—) and video first (—) response curves. Intersection points L and R represent the left and right synchrony boundaries, which form the synchrony range. The dashed line indicates the PSS at the midpoint of this range.

3.1.1 Regression analysis

PSS values were determined for each participant at each distance. Cases where a low amount of audio first responses caused problems with the fit were excluded from analysis. An overview of participants' PSS values at each distance can be seen in Table 3.1. While none of the measured PSS values were more than 3 standard deviations away from the mean, a initial ANOVA analysis did indicate potential outliers. In a subsequent linear regression analysis with PSS as the dependent variable and distance as the sole predictor, four potential outliers were identified (values with a Cook's distance more than three times the mean; Cook (1977)). This included one of the datapoints in the 50 m condition. As a consequence, the decision was made not to include the 50 m distance condition as there was only a single remaining reliable datapoint.

Analyzing this PSS data, a one-way ANOVA showed a significant difference in PSS between distance conditions ($F(3, 19) = 12.08, p < 0.001$). A regression analysis showed that distance negatively predicts PSS ($R^2 = 0.62$, $F(1, 22) = 34.31$, $p < 0.001$; see Figure 3.2A) with a slope of -2.67 ms/m and a shift of 30.36 ms on the y-intercept. This slope approximates the speed differential of light and sound (-2.82 ms/m with $v_{sound} = 355$ m/s). This means that as the sound transmission times *at the observer* increased with distance, the PSS values measured *at the source* decreased at an approximately equal rate (shifting toward audio leads). As these two cancel out, this means that observers consistently perceive audiovisual synchrony when the light and sound arrive at their senses at roughly

the same time. This is thus in line with the results of Lewald & Guski (2004), and not with the perceptual compensation of audiovisual asynchrony over distance proposed by Sugita & Suzuki (2003) and Silva et al. (2013) (see dashed line in Figure 3.2A). The shift of 30.36 on the y-intercept could be a manifestation of the 'horizon of simultaneity' at around 10-15 meters (PSS = 0 at 11.37 meters) where the difference in speed of light and sound is cancelled out by the difference in response time between auditory and visual neurons (Pöppel et al., 1990).

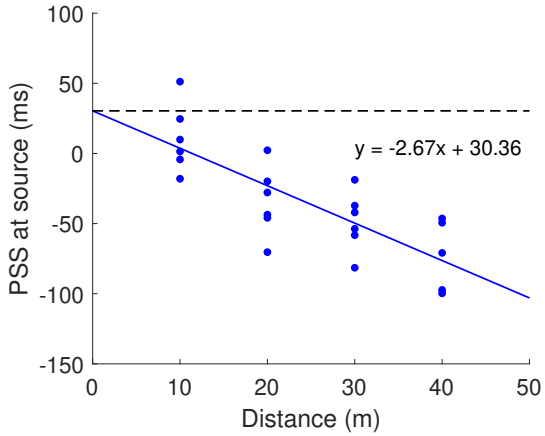

Figure 3.2. The linear regression model of PSS values as displayed at the source, plotted as a function of modelled distance (—). The horizontal dashed line represents the theoretical plot of PSS if observers were able to perceptually compensate for asynchrony over distance.

Table 3.1. PSS values in milliseconds for each distance. Missing values represent cases where the response data was not sufficient to fit an accurate logistic model. PSS values marked with and asterisk were found to be outliers in the linear regression analysis.

	PSS (ms)				
Participant	10 m	20 m	30 m	40 m	50 m
1	9.9	-45.9	-42.1	-70.9	-79.9
2	-	-	-	-	-
3	81.9*	91.6*	46.3*	-99.7	-
4	1.4	-19.9	-58.3	-49.4	-
5	24.6	-70.4	-37.2	-	-
6	-18.1	-27.8	-81.5	-	-
7	51.2	2.3	-18.8	-46.4	-142.2*
8	-4.3	-43.7	-53.7	-97.3	-
Mean	21.0	-16.2	-35.0	-72.7	-111.1
SE	13.2	19.9	15.4	9.6	16.6

3.2 Perceived distance in VR

At the beginning of the experiment, each participant made a single egocentric distance judgement (distance to the loudspeaker pole) for each of the five distance conditions. An overview of the egocentric distance judgements of each participant at each distance can be seen in Table 3.2. A t-test showed that there was a statistically significant difference between the modelled and perceived distance ($t(34) = -10.31, p < 0.001$) as on average, participants underestimated egocentric distance by 41%. Interestingly though, participant 7 overestimated most of the distances. It seems like shorter distances are underestimated slightly more than longer distances, but this is likely due to the influence of the data of participant 7. A one-way ANOVA showed no significant difference in the ratio of perceived/modelled distance between distance conditions ($F(4, 26) = 0.58, p = 0.682$). In addition, a t-test with the most extreme conditions (10 and 50 meters) also did not show a significant difference in the ratio of perceived and modelled distance between conditions ($t(7) = -1.76, p = 0.122$). In general, the underestimation of distance is stronger than in the research by Renner et al. (2013) (41% vs. 27%) and there is definitely more underestimation than the recent research by Feldstein et al. (2020) (41% vs. 4%). This could be explained by differences between the HMD's that were used, software behind the virtual environment or might have to do with participant's inexperience with VR. Additionally, at the end of the experiment participants judged the height of the speaker pole (5 m), width of a tent (8.8 m) and height of a tower (14.6 m). An overview of these size judgments can be seen in Table 3.3. On average, the height of the pole and tower were overestimated by 16% ($M = 5.69, SD = 2.84$) and 22% ($M = 17.83, SD = 8.86$), while the width of the tent was only very slightly underestimated by 3% ($M = 8.50, SD = 4.80$). These judgements seem more consistent with modelled size, which could be the case because they are more reliant on size recall of real objects.

Table 3.2. Egocentric distance judgements for each distance. The average ratio between perceived distance and modelled distance is given for each participant and distance.

Participant	10 m	20 m	30 m	40 m	50 m	Avg. ratio
1	8	17	20	25	30	0.71
2	1	3	6	12	15	0.21
3	4.5	14	18	24.5	30.5	0.59
4	4	8	10	12	14	0.34
5	3	10	15	25	30	0.51
6	3	8	10	25	35	0.47
7	5	25	50	60	75	1.28
8	5	10	25	20	30	0.59
Mean	4.2	11.9	19.3	25.4	32.4	0.59
SE	0.72	2.39	4.90	5.32	6.66	0.11
Avg. ratio	0.42	0.59	0.64	0.61	0.64	

Table 3.3. Size judgements for each object (pole height, tent width and tower height). The ratio between judged size and modelled size is given for each participant and object.

Participant	Pole (5 m)	Tent (8.8 m)	Tower (14.6 m)	Avg. ratio
1	10	15	30	1.92
2	10	15	30	1.92
3	4.5	4.5	7.6	0.64
4	4	5	8	0.64
5	5	4	12	0.76
6	3	8	15	0.85
7	3	12	20	1.11
8	6	4,5	20	1.03
Mean	5.7	8.5	17.8	1.11
SE	1.00	1.70	3.13	0.19
Avg. ratio	1.14	0.97	1.22	

4. Discussion

4.1 Main findings

This study aimed to test the effect of distance on judgements of synchrony of audiovisual events in a virtual outdoor environment. Based on previous research by Lewald & Guski (2004) and Silva et al. (2013), it was hypothesized that observers would not take distance into account in a virtual outdoor environment. To test this hypothesis, a VR experiment was conducted which aimed to measure PSS values at five different distances (10, 20, 30, 40 and 50 meters) in a virtual outdoor environment. Results of this experiment support the hypothesis that observers do not take distance into account in audiovisual synchrony judgements. Data showed that PSS values *at the observer* remained approximately constant over distance. The results suggest that synchrony is perceived when auditory and visual stimuli reach the observer at approximately the same time, which is in agreement with results from Lewald & Guski (2004), but diametrically opposed with the results of Sugita & Suzuki (2003) and (to some extent) Silva et al. (2013). Analyzing the results and differences between the current study and studies like Sugita & Suzuki, Lewald & Guski and Silva et al. is therefore critical to understanding if and how a compensation mechanism for the transmission time of sound exists in audiovisual synchrony perception.

It is interesting to see that (1) the results of this VR study replicate earlier results of a similar 'real' study and (2) that there is a clear difference in results between the outcomes of the current study & Lewald & Guski (2004) and the studies by Sugita & Suzuki (2003) & Silva et al. (2013). The fact that there is now mounting evidence on both sides, suggests that a difference in experimental setup or rendering of the auditory and visual stimuli is leading to these conflicting results.

One of these differences could be the indoor vs. outdoor nature of the experiments. In an outdoor setting, conditions approximate free field acoustics, which means there is exclusively direct sound and no reflections. This means that one of the most powerful auditory depth cues, the ratio of direct to reflected sound, is not present in outdoor conditions. Furthermore, the current study used a relatively artificial visual and auditory stimulus combination (flashing light and siren sound) which means that visual and auditory depth cues due to familiar size and loudness were lacking. This difference in availability of visual

and auditory depth cues between the studies is likely a factor that affects their outcomes, as was demonstrated by Silva et al. (2013). However, in the study by Silva et al. the data suggests that in their 'visual depth cues' condition, when only visual depth cues and directional auditory cues were present, there is a slight effect of distance compensation. It could be argued though, that in the current study and the study by Lewald & Guski, the amount of visual and auditory depth cues that was available was similar if not better than in this condition, so what is causing Silva et al. to still find this effect? Furthermore, as mentioned by Lewald & Guski, an argument can be made that Sugita & Suzuki also did not provide participants with realistic visual and auditory depth cues. If the distance compensation mechanism would be reliant on the amount and quality of depth cues available, it seems inconsistent that Sugita & Suzuki would find the result they did.

Taking a closer look at the difference between the stimuli and experimental procedures that were used, we see that the current study used a relatively artificial flash-beep stimulus setup while Silva et al. used a biological motion stimulus. The usage of a realistic biological motion stimulus by Silva et al. likely enhanced the perceptual coherence and perceived co-localization of the visual and auditory stimuli. Furthermore, participants in the study of Sugita & Suzuki were explicitly instructed to 'imagine' the visual and auditory stimuli were spatially co-located. Since a synchronous onset at the source of the audiovisual stimulus was thus more explicitly inferred or 'to be imagined', this could have led observers to employ biased cognitive strategies, as was argued by Arnold et al. (2005). This means that the distance compensation mechanism might specifically be observed in studies like those of Sugita & Suzuki and Silva et al. because participants are able to effectively use knowledge from daily experience, allowing them to recall previously learned audiovisual timing. In the case of Silva et al., this could explain the importance the auditory and visual depth cues as accurate depth cues should optimally calibrate participants to recall previously learned audiovisual timings and therefore enhance their 'predictive capabilities' (i.e. distance compensation). Importantly though, this does not mean that observers actually used any 'implicit knowledge' of absolute distance or the speed of sound. This conclusion would fit with research by Vroomen et al. (2004) and Heron et al. (2007), who found that after brief phases of exposure to a natural sound lag, participants shifted lag expectations (temporal recalibration). Heron et al. performed a similar experiment to Arnold et al. and Lewald & Guski and found exactly the same results, namely that PSS values shifted to be more audio leading with distance (at a gradient which approximated the speed differential between light and sound). However, after a short adaptation to a audio lagged stimulus, results showed that PSS values shifted toward significantly smaller audio leads. This means that judgements of synchrony shifted more towards the physical timing at the source ('distance compensation'). These temporal recalibration studies show that perceived synchrony is most likely judged relative to a previously established baseline, which is constantly updated by new experiences. It is however, hard to definitively explain the difference in results with Silva et al. with the current data, which means further research is needed.

In addition to the synchrony judgements, perceived distance in the virtual environment was measured and compared with modelled distance through egocentric distance judgements. Results showed that modelled distance was on average underperceived by 41%. An underestimation of egocentric distance in VR is consistent with previous research by Renner et al. (2013), although they found less underestimation at 27%. Furthermore, recent research by Feldstein et al. (2020) found only 4% underestimation. The quite severe underestimation found in the current study is hard to explain, but is most likely a consequence of different calibration (or lack thereof) in the hardware and software that was used in the current study. Another reason for the larger underestimation of modelled distance could be the usage of verbal estimates for the distance judgements, as this method has the disadvantage that possible cognitive influences might bias the measurement. Additionally, studies using verbal estimates have on average resulted in lower estimates (Feldstein et al., 2020). Nonetheless, VR has proven to be a valid tool to research audiovisual synchrony perception as data from all participants showed great internal consistency and alignment between modelled and perceived distance, it should just be taken into account that there will be a certain level of gain in egocentric distance estimation.

4.2 Limitations

The current experiment showed that modelled distances were on average underestimated by about 41%. This perceptual difference should be considered during the design process of future VR experiments as a severe underestimation of distance combined with natural audio delays could cause some incongruence in presentation between visual and auditory stimuli. Furthermore, the sound had little acoustical qualities except for the built in travel time and attenuation. This is a very basic approximation of what a free field sound would be like in real life and certainly carries less auditory information. For instance, the speaker was not always directly facing the observer when the sound was played at different SOA's, however this directionality not present in the sound and might have impacted synchrony judgements. It would most likely not make a difference to the overall trends, but could aid observers in more accurately making the distinction between audio first, synchronous or video first judgements. In general, because the set up and results of this study are largely similar to real world experimentation by Lewald & Guski (2004), it suffers from the same flaws. As Silva et al. (2013) has demonstrated, auditory depth cues and stimulus familiarity can be especially important in the proposed distance compensation mechanism. In the current study, the participant is expressly focusing on a small flashing light and a concurrent simple auditory siren sound which is presented at slightly different delays to simulate sound transmission time over distance. There are some visual depth cues in the environment but the extremely simple nature of the sound used in the experiment is arguably not very 'realistic'. This means that the promise of using VR as a tool to simulate a completely ecologically valid environment and stimuli was not entirely fulfilled in the current research and therefore such an experiment might not be representative of

real world 3D visual and auditory perception.

As for the specific stimuli and procedure that was chosen, some participants reported that the moving speaker was distracting which might have had an influence on performance. While it was meant to provide anticipation for the next light flash through visual movement (the light flash always happened when the speaker was facing the observer straight on), it might have been an unnecessary addition to the experiment that ended up causing more distraction than it helped participants to prepare for the next stimulus. Furthermore, there is a chance that some participants did not have as clear of an understanding of the experiment as others. Two participants were familiar with the general purpose of the experiment beforehand (1 and 8). Interestingly, the synchrony windows for these participants were smaller and model fits were visually better in line with data points (see Appendix B). This suggests they might have been able to discern between answer options better than most of the other participants. Proportion data of the 'audio first' answer option of participants 4, 5 and 6 show unexpectedly high values at some of the most audio lagging SOA's, which might indicate mistaken answers, inattention or a general misunderstanding of the meaning of the answer options. While none of the participants specifically reported not understanding the answer options, future research might consider breaking convention and changing the 'audio first' and 'video first' answer options to wording which is simpler and more clearly related to the scene ('sound first', 'light first') or expressing the answer option in a single modality ('sound before', 'sound after'). The data was split into a first and second half to check for indicators of improved performance through experience with the experiment, but synchrony windows and unexpected answers were not significantly different between the start and end of the experiment. The general trend towards increasing audio leads with distance was also consistent across all data.

In terms of the data analysis, it is important to mention that the low amount of 'audio first' responses in the data caused the analysis to be less robust than it would have been with a more complete dataset. The low amount of 'audio first' responses was brought on by a failure to adjust the SOA's for the visual and audio latency of the VR setup, causing the effective SOA range to shift by +90 ms. As a consequence the calculation of most PSS values relies on a significant amount of extrapolated data (which is based on just a few datapoints) as the left synchrony boundary was often not within the scope of the SOA range. Ideally, the experiment would be repeated with a slightly larger range of SOA's that are properly adjusted to the visual and audio latencies of the system. This should produce a more reliable dataset, meaning that the assumption of inverse slopes for the 'video first' and 'audio first' curves would not have to be made (Alcalá-Quintana & García-Pérez, 2013).

5. Conclusion

5.1 Conclusion

This study aimed to test the effect of distance on judgements of synchrony of audiovisual events in a virtual outdoor environment. Results showed that as the sound transmission time *at the observer* increased with distance, the PSS values measured *at the source* decrease at an approximately equal rate (shifting toward audio leads). These results suggest that synchrony is simply perceived when auditory and visual stimuli reach the *observer* at approximately the same time, since they become perceptually bound. This means that this study did not find any evidence that people use an implicit estimation of sound transmission time when judging the synchrony of an audiovisual stimulus. These findings are in line with previous research by Lewald & Guski (2004), Arnold et al. (2005) and Heron et al. (2007) but directly oppose the results of Sugita & Suzuki (2003) and to some extent Silva et al. (2013). Additionally, results of egocentric distance judgements showed that modelled distance was on average underperceived by 41%. An underestimation of egocentric distance in VR is consistent with previous research by Renner et al. (2013), although they found less underestimation at 27%. Additionally, recent research by Feldstein et al. (2020) found only 4% underestimation. The quite severe underestimation found in the current study is hard to explain, but is most likely a consequence of different calibration (or lack thereof) in the hardware and software that was used in the current study.

5.2 Future work

In future work, more efforts should be made to compare the influence of different stimulus properties, cues and settings on synchrony perception. Such experiments could reveal that the existence of the distance compensation mechanism in audiovisual synchrony perception is not straight forward, but rather dependent on these factors. Interesting avenues to keep exploring are the indoor vs. outdoor experiments, effects of co-location and adding or removing natural audio delays from the sound. Furthermore, the current and future research would benefit from more sophisticated auditory modelling. VR seems to be a

promising technology for audiovisual synchrony perception research, but care should still be taken in the future to present realistic stimuli and correct for possible perceptual differences. The fact that the results of ecologically valid studies (Lewald & Guski, 2004; Heron et al., 2007) were accurately replicated in a relatively simple VR environment shows the potential for using VR as a research tool in this field.

References

Alais, D. & Carlile, S. (2005). Synchronizing to real events: Subjective audiovisual alignment scales with perceived auditory depth and speed of sound. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(6), 2244–2247. doi: 10.1073/pnas.0407034102

Alcalá-Quintana, R. & García-Pérez, M. A. (2013). Fitting model-based psychometric functions to simultaneity and temporal-order judgment data: MATLAB and R routines. *Behavior Research Methods*, *45*(4), 972–998. doi: 10.3758/s13428-013-0325-2

Arnold, D. H., Johnston, A. & Nishida, S. (2005). Timing sight and sound. *Vision Research*, *45*(10), 1275–1284. doi: 10.1016/j.visres.2004.11.014

Begault, D. R., Wenzel, E. M. & Anderson, M. R. (2001). Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source. *AES: Journal of the Audio Engineering Society*, *49*(10), 904–916.

Cook, R. D. (1977). Detection of Influential Observation in Linear Regression. *Technometrics*, *19*(1), 15–18. doi: 10.1080/00401706.1977.10489493

Da Silva, J. A. (1985). Scales for perceived egocentric distance in a large open field: comparison of three psychophysical methods. *The American journal of psychology*, *98*(1), 119–144. doi: 10.2307/1422771

Fain, G. L. (2019). *Sensory Transduction*. Oxford University Press. Retrieved from https://oxford.universitypressscholarship.com/view/10.1093/oso/9780198835028.001.0001/oso-9780198835028 doi: 10.1093/oso/9780198835028.001.0001

Feldstein, I. T., Kölsch, F. M. & Konrad, R. (2020). Egocentric Distance Perception: A Comparative Study Investigating Differences Between Real and Virtual Environments. *Perception*, *49*(9), 940–967. doi: 10.1177/0301006620951997

Foley, J. M. (1980). Binocular distance perception. *Psychological Review*, *87*(5), 411–434. doi: 10.1037/0033-295X.87.5.411

Gerzon, M. (1973). Periphony: With-height sound reproduction. *Journal of the Audio Engineering Society*, *21*(1), 2–10. Retrieved from http://www.aes.org/e-lib/browse.cfm?elib=2012

Gilinsky, A. S. (1951). *Perceived Range and Distance in Visual Space* (Vol. 58) (No. 6). Retrieved from https://psycnet.apa.org/record/1952-05323-001

Heron, J., Whitaker, D., McGraw, P. V. & Horoshenkov, K. V. (2007). Adaptation minimizes distance-related audiovisual delays. *Journal of Vision*, *7*(13), 1–8. doi: 10.1167/7.13.5

Jeon, J. Y. & Jo, H. I. (2020). Effects of audio-visual interactions on soundscape and landscape perception and their influence on satisfaction with the urban environment. *Building and Environment*, *169*(November 2019), 106544. Retrieved from https://doi.org/10.1016/j.buildenv.2019.106544 doi: 10.1016/j.buildenv.2019.106544

Kearney, G., Gorzel, M., Rice, H. & Boland, F. (2012). Distance perception in interactive virtual acoustic environments using first and higher order ambisonic sound fields. *Acta Acustica united with Acustica*, *98*(1), 61–71. doi: 10.3813/AAA.918492

Keetels, M. & Vroomen, J. (2012). Perception of synchrony between the senses. In M. Murray & M. Wallace (Eds.), *Frontiers in the neural bases of multisensory processes* (pp. 147–177). London: Taylor and Francis.

King, A. J. (2005, may). Multisensory Integration: Strategies for Synchronization. *Current Biology*, *15*(9), R339–R341. Retrieved from https://linkinghub.elsevier.com/retrieve/pii/S0960982205004227 doi: 10.1016/j.cub.2005.04.022

Kohlrausch, A., van Eijk, R., Juola, J. F., Brandt, I. & van de Par, S. (2013). Apparent causality affects perceived simultaneity. *Attention, Perception, and Psychophysics*, *75*, 1366–1373. doi: 10.3758/s13414-013-0531-0

Lewald, J. & Guski, R. (2003). Cross-modal perceptual integration of spatially and temporally disparate auditory and visual stimuli. *Cognitive Brain Research*, *16*(3), 468–478. doi: 10.1016/S0926-6410(03)00074-0

Lewald, J. & Guski, R. (2004, mar). Auditory-visual temporal integration as a function of distance: no compensation for sound-transmission time in human perception. *Neuroscience Letters*, *357*(2), 119–122. Retrieved from https://linkinghub.elsevier.com/retrieve/pii/S0304394003014691 doi: 10.1016/j.neulet.2003.12.045

Lindquist, M., Lange, E. & Kang, J. (2016). From 3D landscape visualization to environmental simulation: The contribution of sound to the perception of virtual environments. *Landscape and Urban Planning*, *148*, 216–231. Retrieved from http://dx.doi.org/10.1016/j.landurbplan.2015.12.017 doi: 10.1016/j.landurbplan.2015.12.017

Loomis, J. M., Da Silva, J. A., Philbeck, J. W. & Fukusima, S. S. (1996). Visual perception of location and distance. *Current Directions in Psychological Science*, *5*(3), 72–77. doi: 10.1111/1467-8721.ep10772783

Mareschal, I., Calder, A. J., Dadds, M. R. & Clifford, C. W. (2013). Gaze categorization under uncertainty: Psychophysics and modeling. *Journal of Vision*, *13*(5), 1–10. doi: 10.1167/13.5.18

Military Field Camp 3.2. (2016). Retrieved from https://www.unrealengine.com/marketplace/en-US/product/military-field-camp

Nelder, J. A. & Mead, R. (1965). A Simplex Method for Function Minimization. *The Computer Journal*, *7*(4), 308–313. doi: 10.1093/comjnl/7.4.308

Pöppel, E., Schill, K. & von Steinbüchel, N. (1990, feb). Sensory integration within temporally neutral systems states: A hypothesis. *Naturwissenschaften*, *77*, 89–91. Retrieved from https://doi.org/10.1007/BF01131783http://link.springer.com/10.1007/BF01131783 doi: 10.1007/BF01131783

Renner, R. S., Velichkovsky, B. M. & Helmert, J. R. (2013). The perception of egocentric distances in virtual environments - A review. *ACM Computing Surveys*, *46*(2), 1–40. doi: 10.1145/2543581.2543590

Shemetova, E. & Bodenheimer, B. (2014). Egocentric distance estimation on a discontinuous ground surface in the virtual environment. *Proceedings of the ACM Symposium on Applied Perception, SAP 2014*, 131. doi: 10.1145/2628257.2628358

Silva, C. C., Mendonça, C., Mouta, S., Silva, R., Campos, J. C. & Santos, J. (2013, nov). Depth Cues and Perceived Audiovisual Synchrony of Biological Motion. *PLoS ONE*, *8*(11). Retrieved from https://dx.plos.org/10.1371/journal.pone.0080096 doi: 10.1371/journal.pone.0080096

Spence, C. (2007). Audiovisual multisensory integration. *Acoustical Science and Technology*, *28*(2), 61–70. doi: 10.1250/ast.28.61

Spence, C., Baddeley, R., Zampini, M., James, R. & Shore, D. I. (2003). Multisensory temporal order judgments: When two locations are better than one. *Perception and Psychophysics*, *65*(2), 318–328. doi: 10.3758/BF03194803

Spence, C. & Squire, S. (2003). *Multisensory integration: Maintaining the perception of synchrony* (Vol. 13) (No. 13). Cell Press. doi: 10.1016/S0960-9822(03)00445-7

Sugita, Y. & Suzuki, Y. (2003, feb). Implicit estimation of sound-arrival time. *Nature*, *421*(6926), 911. Retrieved from http://www.nature.com/articles/421911a doi: 10.1038/421911a

Teghtsoonian, M. & Teghtsoonian, R. (1970). Scaling apparent distance in natural indoor settings. *Psychonomic Science*, *20*(6), 281–283.

van Eijk, R. L. (2008). *Audio-visual synchrony perception* (Doctoral dissertation, Technische Universiteit Eindhoven). doi: 10.6100/IR634898

Vatakis, A. (2013, mar). The Role of Stimulus Properties and Cognitive Processes in the Quality of the Multisensory Perception of Synchrony. In *Handbook of experimental phenomenology* (pp. 243–263). Chichester, UK: John Wiley Sons, Ltd. Retrieved from http://doi.wiley.com/10.1002/9781118329016.ch10 doi: 10.1002/9781118329016.ch10

Vroomen, J., Keetels, M., De Gelder, B. & Bertelson, P. (2004). Recalibration of temporal order perception by exposure to audio-visual asynchrony. *Cognitive Brain Research*, *22*(1), 32–35. doi: 10.1016/j.cogbrainres.2004.07.003

Zahorik, P., Brungart, D. S. & Bronkhorst, A. W. (2005). Auditory distance perception in humans: A summary of past and present research. *Acta Acustica united with Acustica*, *91*(3), 409–420.

Appendix A

Visual and audio latency

Table 1. Visual latency for Unreal Engine on the PC used in the current study. Visual latency was defined as the latency between a trigger in Unreal Engine and subsequent light flash on the Oculus Rift head mounted display (averaged over 10 measurements).

VR Display	VR Engine	V-Sync	Min (ms)	Max (ms)	M (ms)
Oculus Rift	Unreal	On	21	32	29

Table 2. Audio latency for Unreal Engine (4.27.1) on the PC used in the current study. Audio latency was defined as the latency between a trigger in Unreal Engine and subsequent sound burst on the headphones of the Oculus Rift (averaged over 10 measurements). In this case, the Unreal Engine trigger was routed through Cycling '74's Max software (8.1.11) which generated the sound that was sent to the headphones.

VR Display	VR Engine	V-Sync	Min (ms)	Max (ms)	M (ms)
Oculus Rift	Unreal	On	101	136	119

Appendix B

Fitted logistic models

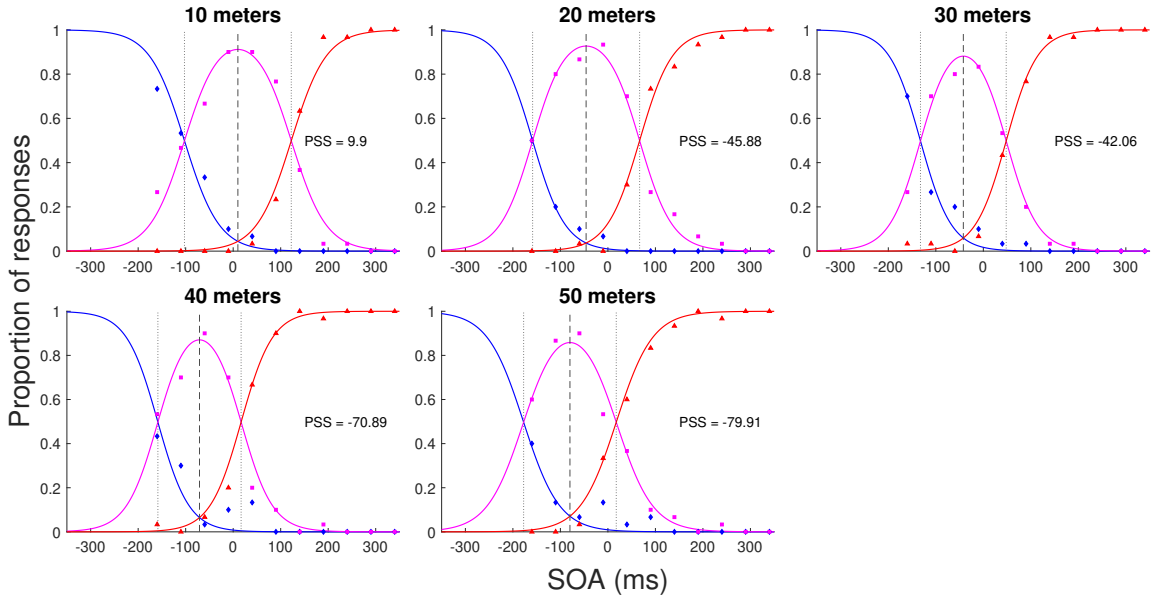

Figure 1. Fitted models of participant 1.

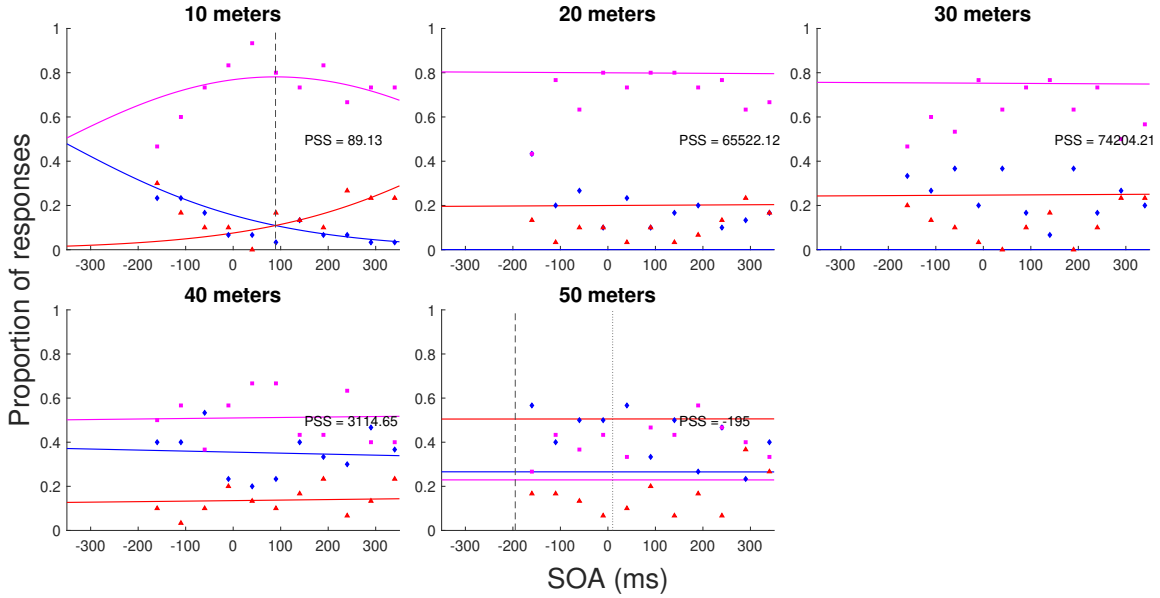

Figure 2. Fitted models of participant 2. Bad fit for all of the distance conditions.

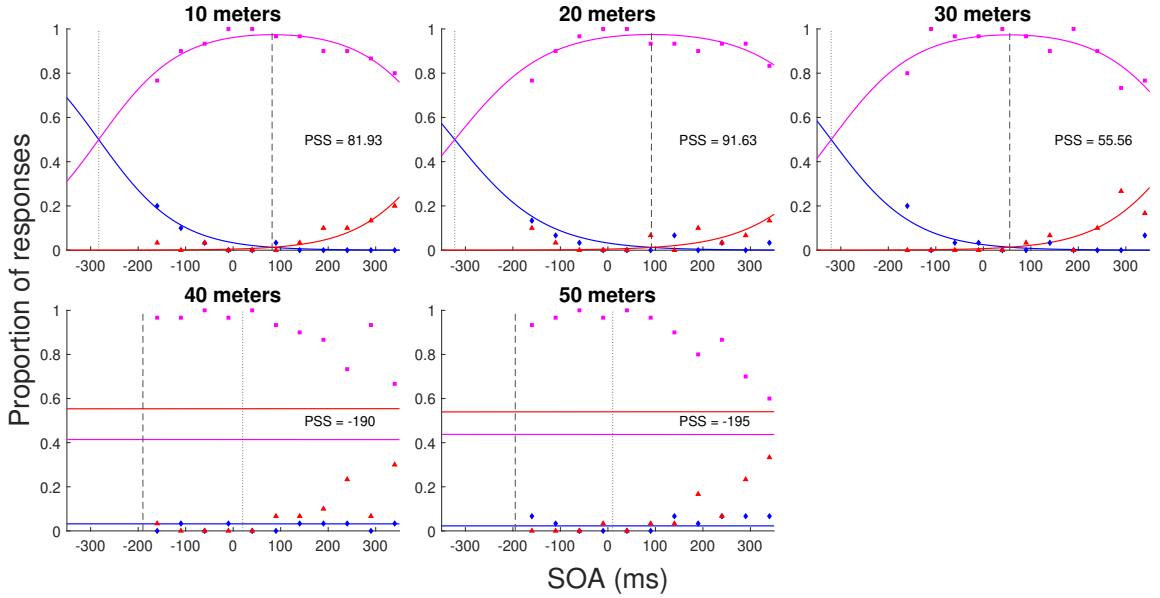

Figure 3. Fitted models of participant 3. High percentage of 'synchronous' responses in all distance conditions. Bad fit at 40 and 50 meters.

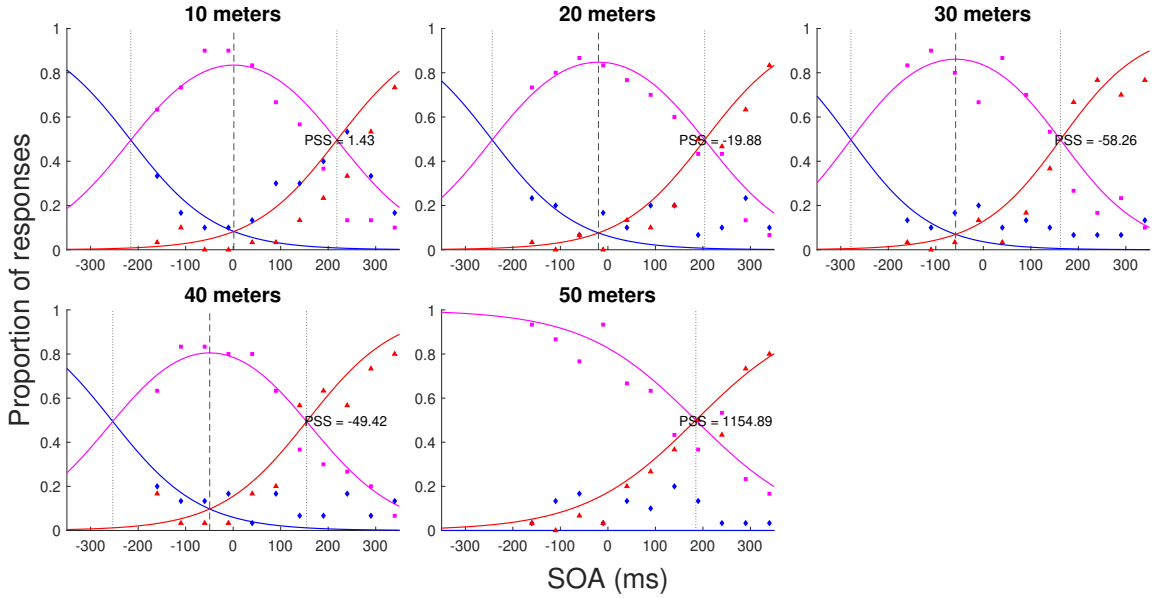

Figure 4. Fitted models of participant 4. Bad fit at 50 meters.

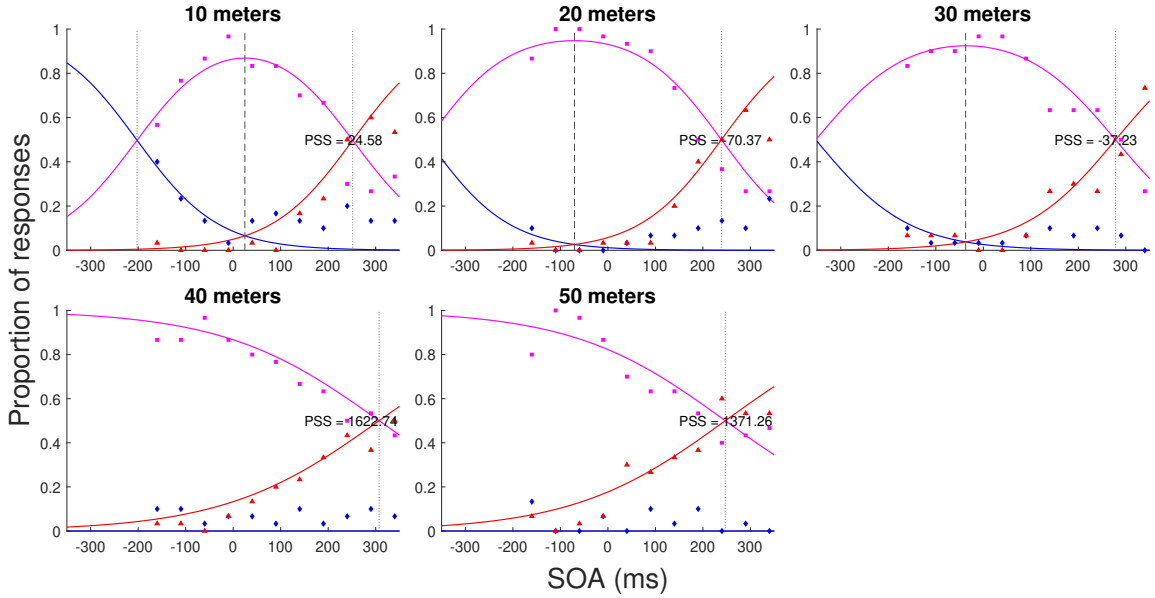

Figure 5. Fitted models of participant 5. Bad fit at 40 and 50 meters.

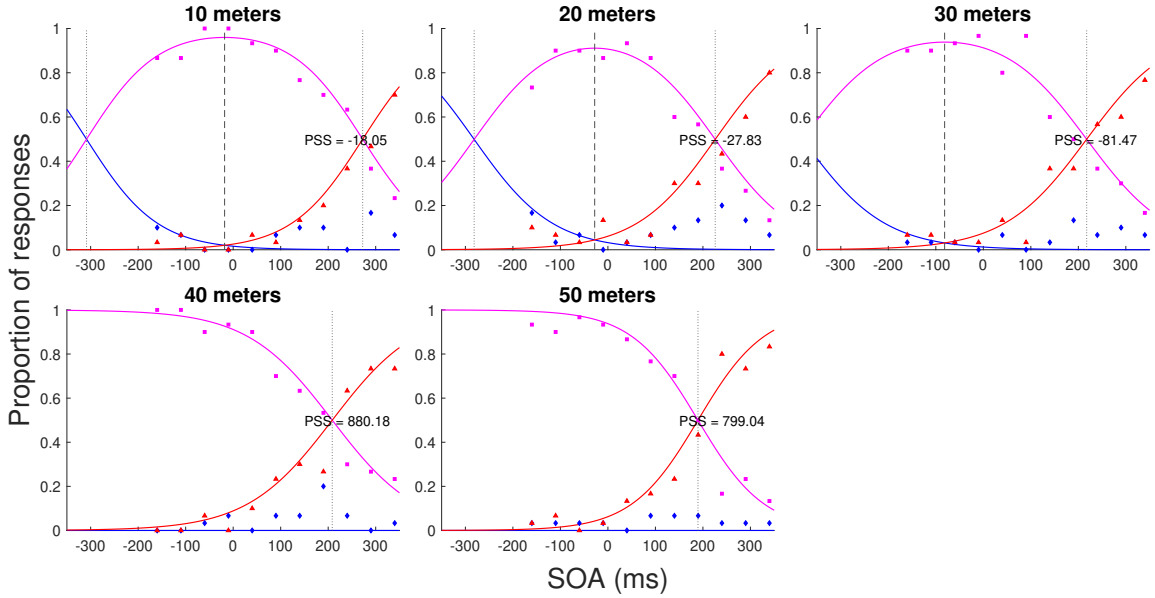

Figure 6. Fitted models of participant 6. Bad fit at 40 and 50 meters.

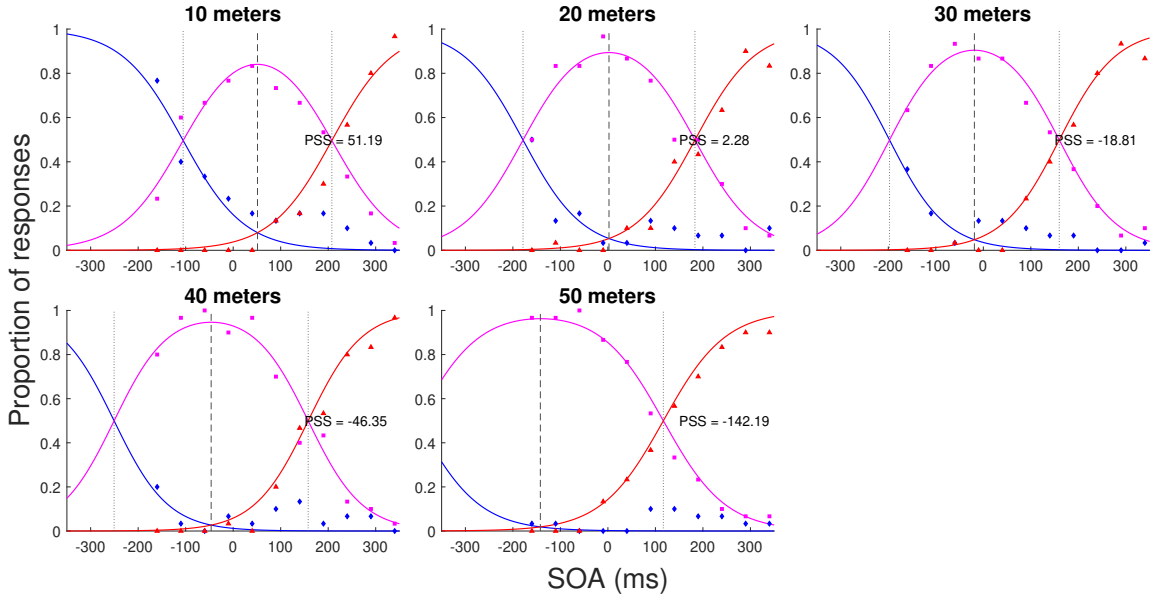

Figure 7. Fitted models of participant 7.

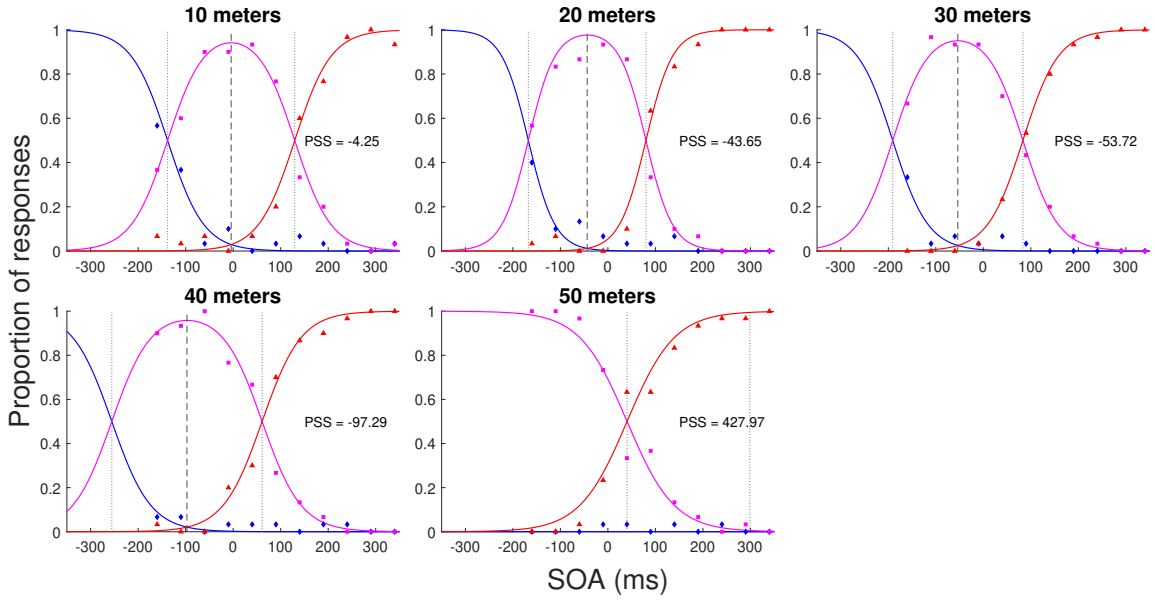

Figure 8. Fitted models of participant 8. Bad fit at 50 meters.

Appendix C

Flowchart Unreal code