FAKULTÄT FÜR !NFORMATIK

Audio-Visual Perception in Interactive Virtual Environments

Diplomarbeit

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Computergraphik & Digitale Bildverarbeitung

ausgeführt von

Karl Grosse
Matrikelnummer 0225662

am:
Institut für Computergraphik und Algorithmen

Betreuung:
Betreuer: Associate Prof. Dipl.-Ing. Dipl.-Ing. Dr.techn. Michael Wimmer

Wien, 07. 12. 2009

(Unterschrift Verfasser) (Unterschrift Betreuer)

Abstract

Interactive virtual environments (VEs) are gaining more and more fidelity. Their high quality stimuli undoubtedly increase the feeling of presence and immersion as ''being in the world'', but maybe they also affect user's performance on specific tasks. Vision and spatial hearing are the main contributors of our perception. Sight dominates clearly and has been in the focus of research for a long time, but maybe it is the audio-visual combination which facilitates the user in his decision making and in completing a task.

Mere identification of the task is not enough. Of course one could find dozens of problems where spatial sound reproduction has a practical relevance. More interesting are those which reside on a high cognitive level. Tasks that combine visual stimuli and auditive perception with movement provide a wide field of activity like for example crossing a busy road, an every day task that contains a high information density and demands fast processing by the brain. But how does hearing have an impact on this? Does spatial audio lead to better performance? Can one adjust naturalistic, spatialized hearing virtually? This diploma thesis asseses the effect of spatial sound reproduction compared to conventional stereo sound or no sound at all. Within the scope of the practical part, a simulator was implemented to produce a virtual street crossing experiment. It was later used to perform a study with volunteer participants.
The results give evidence that there is a statistically significant difference between spatialized sound rendering compared to stereo sound or no sound. In the future this can not be used solely to boost the naturalistic fidelity and authenticity of a virtual environment but also as a user supportive measure.

Zusammenfassung

Interaktive virtuelle Umgebungen werden immer realitätsgetreuer. Ihre qualitativ hochwertigen Stimuli erhöhen ohne Zweifel das Gefühl der Präsenz und Immersion ''mittendrin in der Welt zu sein''. Sehen und räumliches Hören machen den größten Teil unserer Wahrnehmung aus, der visuelle Bereich dominiert deutlich und ist daher schon seit längerem im Blickpunkt der Forschung, aber vielleicht ist es gerade ihre Kombination, die den Anwender in seiner Entscheidungsfindung und damit in der Bewältigung einer Aufgabe unterstützt.

Mit der Identifikation der Aufgabe alleine ist es nicht getan. Natürlich könnte man sich Dutzende einfallen lassen bei denen räumliche Tonwiedergabe einen praktischen Sinn macht. Interessant sind hier solche, die kognitiv auf einem ziemlich hohen Level angesiedelt sind. Aufgaben die sowohl zu einem hohen Maß aus visuellem Reiz und auditiver Wahrnehmung in Kombination mit Bewegung gestaltet sind, bieten hier ein breites Betätigungsfeld. Ein Beispiel ist das Überqueren einer befahrenen Straße, eine alltägliche Aufgabe, die für unser Gehirn eine hohe Informationsdichte beinhaltet und schnelle Verarbeitung erfordert. Wie allerdings wirkt sich dabei unser Hören aus? Bringt räumlicher Ton eine gesteigerte Leistung? Kann man dieses natürliche, räumliche Hören virtuell nachstellen? Hier setzt diese Diplomarbeit an und untersucht die Auswirkung von räumlicher Tonwiedergabe verglichen mit konventionellem Stereoton oder gar keinem Ton. Im Rahmen des praktischen Teils ist ein Simulator entstanden mit dessen Hilfe ein Straßenquerungsexperiment nachgestellt wird. Dieses wird anschließend im Rahmen einer Studie mit freiwilligen Probanden durchgeführt.
Die Ergebnisse geben einen Hinweis, dass es einen statistisch signifikanten Unterschied zwischen räumlichem Ton, Stereoton und keinem Ton gibt. Zukünftig könnte dies nicht nur die natürliche Qualität und damit die Glaubwürdigkeit der virtuellen Umgebung steigern, sondern auch die Anwender selbst unterstützen.

Acknowledgements

I would like to gratefully acknowledge the supervision of Michael Wimmer during this work.

I wish to acknowledge Matthias Bernhard for the numerous stimulating discussions, his help with the experimental setup and general advice.

I would like to thank the people at INRIA for providing the audio library used for the experiment.

I thank Andrea Fübi and Anita Mayerhofer who were a big administrative help to me and the conducted study.

My parents receive my deepest gratitude and love for their dedication and the many years of support during my studies that provided the foundation for this work. It is them that this thesis is dedicated to.

And to all the friends and fellows who kept asking me about my progress: It's done!

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

An important application of computer graphics is the creation of sufficiently realistic simulations that provide the ability to analyze and learn about real-world situations that are otherwise difficult and expensive to reproduce. One of these situations involves understanding pedestrian behavior at street crossings. In 2008, pedestrians injured had a share of 11% (4.233 of 50.521) among traffic accidents in Austria with a 15% share (102 of 679) of lethal outcome [Acc]. Crossing a street is a commonly encountered situation, but as accident statistics show pedestrians are making poor decisions about when to cross. Therefore every year many pedestrians, especially children and old people are killed crossing a street as they either do not see an oncoming car, or fail to correctly estimate its time to impact. Making the decision to cross the road is a highly complex task which requires efficient perceptual and cognitive processes. One has first to detect the vehicles and safety installations on the street and integrate information from the various parts of the road in order to form a mental representation of the situation, which then must be frequently updated to keep up with environmental changes.

With a computer-generated model one can visualize and analyze structural projects in advance. There is thus a need to provide realistic simulators that enable urban planners to judge whether a pedestrian crossing is safe and whether additional building measures (marks on the street, posts that make parking near the crossing impossible etc.) are needed to improve the safety. Together with new approaches in the field of town development, VEs can be used to teach safe traffic situation handling to protect especially the youngest road users from danger.

In order to realize such systems, the question arises, which factors are to be considered. ''Factor'' here means perceptual influences on the human cognition, and whether or not they have a mutual effect on each other, and if they provide an augmentation of the human sensory system. It has been demonstrated that by exploiting crossmodal interactions in visual attention due to the limitations of the human visual system audio can compensate the perceptibility of visual defects [MDCT05]. That does not mean that vision and audition are limited by a common attentional resource, it is more the combination and integration of different sensory modalities [EB04, AMB06] for a robust and reliable perception. So maybe it is possible to achieve a better sensory reliability by boosting the auditory modality in the simulator?

This thesis picks up the question whether increasing the quality of audio will significantly affect a person's ability to correctly judge the safety of crossing a road as these conclusions are important in the context of the design and operation of such simulators. As part of the inter-university project called CROSSMOD [Xmd], we describe our experiment based on neuro-science literature study in order to examine a set of crossmodal phenomena that can be potentially exploited to improve quality and efficiency of VEs.

3D sound can contribute to the sense of immersion in a 3D-environment [RVSP09, HB95]. But we want to explore if one might be able to work more efficiently in a VE with authentic 3D sound, which emits from its proper location like in reality. To investigate this requires a close-to-reality scenario with a task from daily life. Therefore the proposed scenario is a pedestrian simulator that requires users to correctly determine, in the face of an oncoming car, whether it is safe to cross a road in a given virtual environment or not.

1.2 Outline of the document

First we will present some background in Chapter 2 about human seeing (Chapter 2.1) and hearing (Chapter 2.2) to form a knowledge base necessary for audio-visual perception described in Chapter 2.4. After the psychophysical background is set we continue to highlight well established rendering techniques for both modalities in Chapter 3. In Chapter 4, we present existing literature and its implications regarding the researched area. Chapter 5 describes our methodology and implementation of an interactive virtual environment application, after which Chapter 6 explains the study which has been carried out. Results are shown in Chapter 6.7 and discussed in Chapter 6.8. The thesis is concluded with Chapter 7, which dissects our findings in 7.1 and gives an outlook on future work in Chapter 7.2.

Chapter 2

Seeing and Hearing

For the acquisition of environmental information all sensory systems normally work together in such a way that the contribution of one individual system can be seen hardly independently of one another. Nevertheless we will first examine them separately for better clarity and then try to explain the merging of senses. We limit this chapter to seeing and hearing and discuss some anatomical and physiological aspects.

2.1 Seeing

Several anatomical factors have substantial meaning for our seeing [Gus00]:

1. we have two eyes
2. these are horizontal next to each other arranged with a distance between 5–8 cm from each other (inter-ocular distance, horizontal disparity)
3. the spatial resolution of the eye decreases with increasing distance on the retina from the fovea to the periphery
4. they are above all remaining parts of the body except the brain
5. the eyes are mobile in the head
6. the head is mobile relative to the body
7. the body is mobile in the horizontal level relative to the ground

The information supplied by our two eyes differs slightly in the horizontal level from each other. This horizontal disparity leads to the sensation of depth from the two slightly different projections of the world onto the retinas of the two eyes. Six eye muscles, innervated by 3 nerves, permit arbitrary and automatic eye movements, which run binocularly coordinated. We do not look at a scene in fixed steadiness; instead, the eyes move around, locating interesting parts of the scene to build up a mental 'map' corresponding to the scene. One reason for these so-called saccadic movement of the human eye is that the central part of the retina, known as the fovea, plays a critical role in resolving objects. By moving the eye so that small parts of a scene can be sensed with greater resolution, resources can be used more efficiently.

The optics of the eye consists of a compound lens system with components which possess different refraction and illustration quality. Light falls on the transparent cornea, through the anterior chamber filled with water and the iris behind it, passes the lens, which is variable in its curvature, afterwards travels through the ciliary body filled with a clear gel and finally reaches the retina, which covers the rear surface of the eye inside. At its center is the fovea, a pit that is most sensitive to light and is responsible for our sharp central vision, not far away from the optic disc (blind spot), where the nerve fibers step out of the eye [EB].

The so-called field of view (FOV) is the angular extent of the observable world that is seen at any given moment with stationary eyes and motionless head. It is almost elliptical for binocular vision with 180° horizontal size [Sch57]. In our visual field we can see only a relatively small area sharply. This part falls onto the fovea and is about 1–2 visual angles [O'S91]. A visual angle is the solid angle of an object on the half-spherical fovea. The angular resolution of the eye decreases with increasing distance from the fovea to the periphery (Figure 2.1). If an object is at the edge of our visual field, then we will hardly notice it, as long as it is stationary. As soon as it moves, we will not be able to recognize it immediately in detail, but we will discover that something moves. This points out that discovering movements can take place also independently of noticing the form and recognizing objects [Gus00].

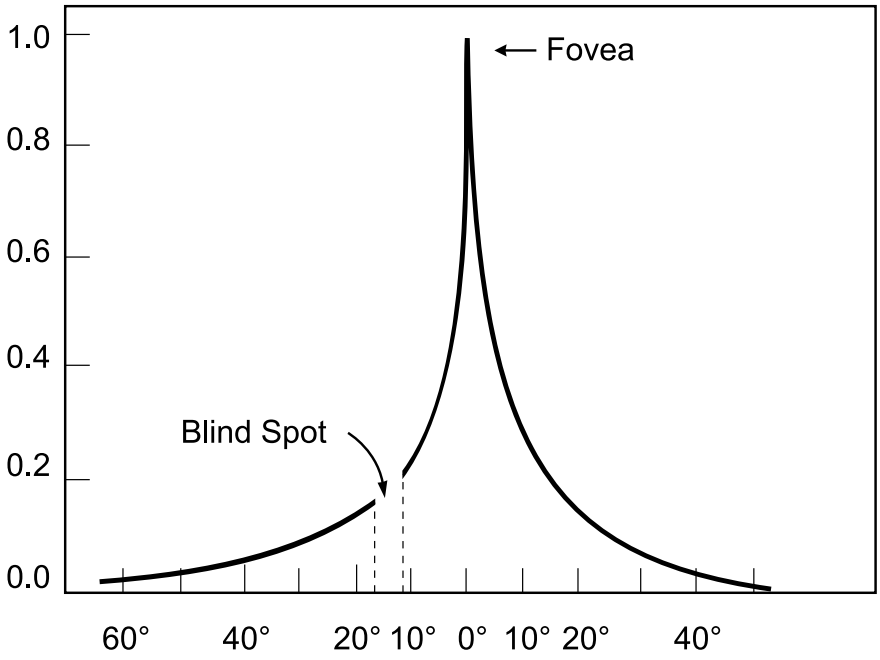


Figure 2.1: Acuity function of the eye; resolution decreases with increasing distance from the fovea to the periphery [Wik]

2.2 Hearing

The *hearing* is the sense responsible for perceiving *sound* [EB]. It has evolved in the course of time to a very sensitive and complex organ. Before we discuss the human auditory system, we first approach the topic of sound to form a basis for the terms used later on.

2.2.1 Sound

Sound is an oscillation of pressure transmitted as longitudinal traveling wave through gases, plasma and liquids or as transverse wave through solids, composed of frequencies within the range of hearing and a level sufficiently strong to be heard. These longitudinal waves are alternating pressure deviations from the equilibrium pressure causing local regions of compression and rarefaction[1], while transverse waves are waves of alternating shear stress at right angle to the direction of propagation [Ols67].

[1] Rarefaction is the reduction of a medium's density, or the opposite of compression [Wik]

Sound waves are characterized by the properties of waves which are frequency, wavelength, period, amplitude, direction, intensity and speed. They can depend on the physical properties of the medium as well as on the type of sound waves. The fundamental perceived frequency of a sound is called the pitch. Whenever the pitch changes, so does the distance between the sound wave maxima, resulting in a change of frequency (e.g. Doppler effect). When the loudness of the sound wave changes, the amount of compression in the air the wave is traveling through also changes, which can be defined as amplitude.

A sound wave generates a tone, its frequency determines the tone's pitch [Ols67]. Sound is produced by superimposing different sound waves. Each pitch must be a multiple of the root tone. Otherwise the overlay of non periodic waves is called noise. The human hearing detects such noise and tries to analyze the parts in order to extract single tones.

2.2.2 Organs

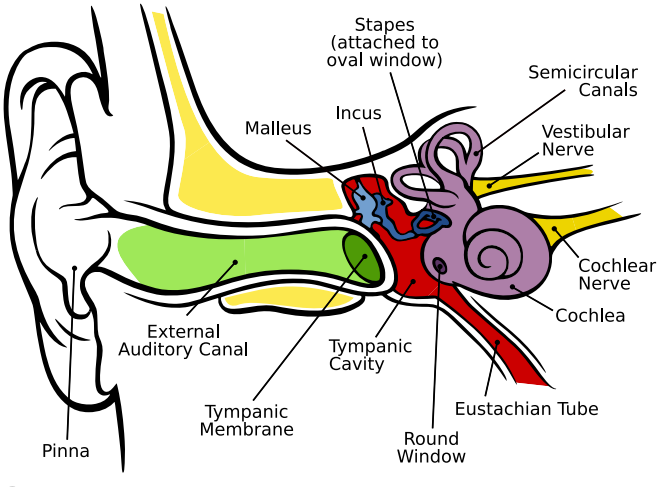


Figure 2.2: Schematic picture of the human ear, showing the three parts outer-, middle- and inner ear [Wik]

The ear is the sense organ that recognizes sound. The hearing range describes the range of sound frequencies that can be heard and lies between 16Hz and 20000Hz. The upper boundary is dependent on the age and tends to lower about 1000 Hz every ten years. The ear is formed by three parts, the outer-, middle- and inner ear.

The outer ear

The outer ear collects the sound, it includes the earlap (pinna) which is the only outside visible part, the ear canal and the most superficial layer of the ear drum which separates the outer ear from the middle ear. The hollows and elevations of the pinna form a resonating body that amplifies sound coming from certain directions. The resulting pattern of minima and maxima of frequencies is used to determine the direction of the sound source [Gus00]. Through the outer ear sound is relayed to the next part.

The middle ear

The middle ear, an air filled cavity, is confined by the ear drum (tympanic membrane) and includes the three ear bones (ossicles): the malleus (or hammer), incus (or anvil) and stapes (or stirrup). The opening of the Eustachian tube, which is responsible for pressure equalization between the middle ear and the atmosphere, is also located there. The malleus is directly connected to the ear drum, the incus is the bridge between the malleus and stapes. Sound pressure arriving at the middle ear causes movement of the tympanic membrane, which causes movement of the malleus, which causes movement of the incus, which causes movement of the stapes. The force at the ear drum forms a lever with the hammer, as well as the anvil forms a lever with the stapes plate. The lever length of the stapes is shorter and therefore leads to an amplification of force by a factor of 1.3 [EB]. This together with the amplification caused by the surface area difference of the tympanic membrane (approx. $55mm^2$) and the stapes footplate (approx. $3.2mm^2$), results in a 22-times gain and leads to another important task of the middle ear, the protection of the inner ear from too intense sound waves and the adaption of the working sound level and the environmental sound level. This is accomplished by muscles at the tympanic membrane and the stapes that alter the sensitivity of our hearing through contraction. Since the reaction time bears a certain delay of $100ms$, the inner ear can suffer severe damage through sudden changes in loudness (acoustic shock).

The inner ear

The inner ear includes both the organ of hearing (cochlea) and a sense organ that is attuned to the effects of gravity and motion (labyrinth or vestibular apparatus), the latter has nothing to do with perceiving sound. The inner ear is encased in the hardest bone and the hardest material after teeth of the body. The cochlea has three fluid filled spaces: the tympanic canal, the vestibular canal and the middle canal. The inner ear

is responsible for transforming mechanical impulses to nerve signals. When sound reaches the ear drum, the movement is transferred to the footplate of the stapes, which presses in the fluid filled ducts of the cochlea. The fluid inside is moved flowing against tiny little hair cells called receptor cells from the oval window until they reach the end of the cochlea (helicotrema). For very low frequencies (below 20Hz), the waves propagate along the complete route of the narrowing cochlea, higher frequencies do not propagate as far to the helicotrema. Every frequency relates to a certain area of the cochlea, which fires the most at this frequency. The hair cells release nerve impulses that are transmitted by the acoustic nerve to the brain. This corresponds to a transformation of mechanical to electrical signals. The brain knows the components of the noise, the tones, as each neuron is activated only by certain frequencies.

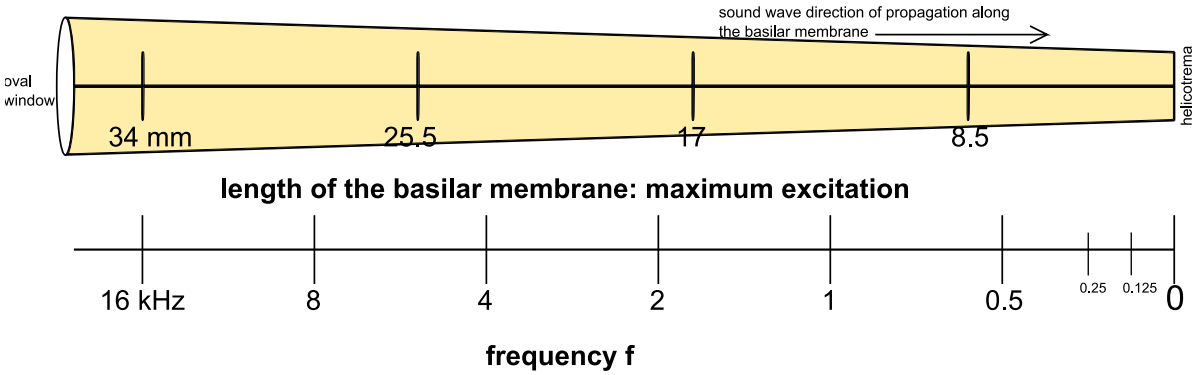


Figure 2.3: Propagation of sound waves and their frequencies along the narrowing basilar membrane; cochlea is assumed to be stretched out for depiction [Wik]

2.3 Binaural Hearing and Sound Localization

As the word binaural indicates, both ears are involved in the perception of a sound event and this principle enables humans to determine the direction of origin of sound.
Let us recall: A wavefront reaches the outer ear, works its way through the ear canal, reaches the ear drum and after passing the inner ear the signal is processed by a complex apparatus in the human brain. As an abstraction one could think of two microphones attached to a computer, placed away at a certain distance from each other. The brain is responsible for interpreting different temporal and frequency-dependent cues to reassemble a three-dimensional image of the complete sound field. The two most important cues are
[Wik]

- Interaural time difference (ITD)

  The ITD corresponds to the temporal difference in arrival time of sound between our two ears and is most important for spatial perception of sound.

- Interaural level difference (ILD)

  The ILD is the difference of the sound pressure level arriving at the two ears.

Apart from the ILD and ITD for localization, there are some other cues used for perceiving distance and movement of sound events as well as implications of the environment

- Doppler Frequency Shift

  The change in frequency of sound waves while the source and the observer approach each other or move away from each other.

- Distance Attenuation

  Describes the reduction in sound volume based on the distance to the listener.

- Monaural Spectral Cues

  These cues are the result of the direction-dependent filtering of incoming sound waves accomplished by the pinna.

- Reverberation

  Reverberation is the persistence of sound in a particular space after the original sound is removed.

2.4 Audio-Visual Perception

With the proper physiological background set, this chapter is about the cooperation and interaction between the different perception systems, particularly the mutual influence of seeing and hearing. These two systems cooperate in everyday life in the sense that they can support themselves mutually and draw together more usable information from the situation than one system alone. A typical example is: If we walk on a close road and a car approaches from the rear, then our hearing notices this first; it alarms the entire body, and if sufficient time up to the threatening collision remains, we turn, in order to take the car into inspection and prepare further actions. If there is not sufficient time left for further visual inspection, then we jump immediately to the side. We can call this a co-operation of the sense systems to survival protection in a way that the auditive system functions as ''early warning system'' over the presence of a potential threat, it supplies information about the distance of the threatening object and the remaining time for possible preventive measures. If distance and time are large enough, then a transfer of the auditory to the visual subsystem takes place for controlling the motor function. The visual system can then explore the kind of the object, its position, direction of motion and speed more exactly.

The above mentioned example highlights a substantial functional difference between seeing and hearing: the differences in spatial and temporal selectivity [Gus00].

Spatial Selectivity

The visual system has a central field of view (fovea), where acuity and color distinction are very good. With increasing distance from this center to the periphery both become worse, but movements can be recognized more easily. The visual field is however altogether limited to approximately half of the environment which lies in front of our nose, and with good visibilities we can see several kilometers far. In addition we steer the eyes arbitrarily around so that they can examine fields of the surrounding area more exactly. There is no comparable organization in any other sensory system. When hearing we perceive spacious information from the environment which surrounds our body, when smelling something similar happens at shorter distance, pressure and temperature are felt at close range to the whole body surface (with regional different resolving power). We can achieve a higher spatial selectivity only by moving closer to a sound

source, but we cannot achieve the close spatial selectivity which is present on larger distance in the visual system in any of the other systems.

Temporal Selectivity

Regarding temporal selectivity, there are differences as well: We can close our eyes, which temporarily locks the visual information channel, however we cannot seal the ears, the nose and our skin. Hearing, smelling and feeling take place all the time - also whilst sleeping. The temporarily complete suppression of information other than visual is not possible. The visual system can on the one hand catch fine details from the enormous information offer, on the other hand it misses all the available information which lies outside the field of view. A condition for the survival of an organism equipped with so different systems is the close coalition and mutual control between the subsystems.

Implications

In our everyday life we usually have multimodal information about the place of an object or an event: If telephones ring, cars drive, humans speak, dogs bark, then the visible place of the sound agrees with the audible, and we are informed both by the visual and the auditive system about the place. The senses cooperate here usually in the way that they notice corresponding information. While objects are usually visible and touchable, we can see and hear, sometimes also smell or feel events. If we want to identify objects, our auditive system can only contribute if the object is part of an actual event, e.g. a human speaks, a telephone rings, an engine runs, a pencil falls on the ground etc. We can often identify events correctly if information from only one modality is available (e.g. when telephoning), however we feel even safer if we get suitable information from a further modality [Gus00].

Contrary to our everyday life observation that intermodal noticing and identifying of events represent the rule and unimodal noticing and identifying the exception, intermodal investigations and studies are very rare - and unimodal very common. A reason for this might be the specialization of nearly all researchers in the cognition and perception domain since it is difficult enough to grasp. Another reason might be the variety of

the examinable intermodal combinations and their range of unimodal variation in each case.

As [EB04] state, the key to robust perception is the combination and integration of multiple sources of sensory information and they suggest that humans combine the available information following two strategies:

Sensory combination

This strategy tries to maximize information delivered from the different sensory modalities. The incoming information streams of the environment are processed by the human brain to reconstruct and update a mental image of the scene. Sometimes the human brain is confronted with ambiguous situations: for example, we sit in a train looking out of the window at a neighboring train. If the other train starts moving, we are presented with an ambiguous situation: is it our or the other train that is moving? This leads the brain to a – right or wrong – answer, if the brain is wrong the illusory self-motion is noticed either when looking out of an opposite window or when a different sensory modality such as the vestibular system disambiguates the situation. That means, it collects more and more information about the perceptual event to finally resolve the ambiguity. If a single modality is not enough, information from several modalities are combined [SM93]. These collected signals are not redundant, they may be in different units, coordinate systems or about complementary aspects of the same environmental property. The brain at any given moment picks a single solution from all the possibilities rather than delaying an uncertain decision.

Sensory integration

This strategy describes interactions between redundant signals. If there is more than one sensory estimate available for perceiving some environmental property, the information has to be integrated so that a coherent multisensory percept is formed. To come up with the most reliable (meaning unbiased) estimate, the variance of the final estimate should be as low as possible. If the system made 10 estimates of the same environmental property, all 10 would be slightly different due to the fact that every sensory signal is noisy. The estimate with the lowest variance is the Maximum Likelihood Estimate [SM93] and according to this, the integrated estimate $\hat{s}$ is the weighted sum of the individual estimates with weights ω_i proportional to their inverse variances σ_i^2 (i referring to the different sensory signals):

$$\hat{s} = \sum_i \omega_i \hat{s_i} \qquad \sum_i \omega_i = 1$$

$$\omega_j = \frac{1/\sigma_j^2}{\sum_{1...,j,...N} 1/\sigma_i^2}$$

The estimate's weight should take the quality of information into account. The sensor's quality can vary according to its reliability (depending on the noise level). Reliability r is defined as the inverse variance of the estimates:

$$r_i = 1/\sigma_i^2$$

Then the reliability of the integrated estimate is simply the sum of the individual estimates

$$r = \sum_i r_i$$

where the reliability of the integrated estimates is increased and yields the most reliable unbiased estimate possible.

The integration can be exemplified by the ventriloquism effect [AB04]: For centuries people are fascinated by so-called ''belly talkers'' (ventriloquists), who are able to bring out words without moving their mouth. By pulling together the palate, withdrawing the tongue and narrowing the laryngeal entrance they decrease the resonance in the throat area. This produces the imagination that the voice comes from the belly. They ''throw'' their voice to appear somewhere else. But this is still no genuine illusion, because the main energy of their voice really develops below the throat area. If the ventriloquist additionally moves a doll with the hands suggesting a dialogue between him and the doll, then the second voice does not seem to come from his belly but from the doll. We locate this acoustic source at the nearest place towards which our attention is directed. This place does not have to be under any circumstances plausible for sound formation. Reliability for the puppet r_i is high for $i = visual$, leading to the most reliable result.
The brain's attempt to minimize uncertainties, is one of the key factors for the positive impact of spatialized sound in our main experiment (see Chapter 6).

Chapter 3

Rendering Virtual Environments

VEs offer many advantages over real-world setups due to their controllability, configurability, flexibility, repeatability, and the ability to imitate, or implement situations and locations which are either dangerous, costly or difficult to reproduce in reality. As the creation of convincing VEs is used by us to explore audio-visual perception, this chapter is about rendering VEs. First we will briefly describe graphical rendering, followed by stereo sound rendering, and finally binaural synthesis.

3.1 3D-Graphics

Rendering of 3D graphics is the process of generating an image from a model by means of computer programs. The model is a description of three-dimensional objects in a strictly defined language or data structure. Rendering contains geometry, viewpoint, texture, lighting, and shading information. In a real-time application such as games or simulators, rendering of interactive media is calculated and done in real-time at frame rates of approximately 20 to 120 frames per second (FPS). At 20 FPS the mind sees movement as motion rather than flashing images. At 60 FPS the system reevaluates and updates the necessary outputs 60 times per second under all circumstances, we perceive smooth animations without any ghosting artifacts. Rates of 120 FPS are necessary if one wants to produce two different images used for stereoscopic real-time rendering, one for the left and one for the right eye, each one with a single rate of 60 FPS. Keep in mind that the designed frame rates of real-time systems vary depending on the equipment. Even when 75 FPS are computed for a monitor running 60 Hz refresh, no more than 60 FPS can actually be displayed on screen.

The goal is primarily speed in terms of smooth animations and not photo-realism. In fact, optimizations are made in the way the eye ''perceives'' the world, and as a result the final image presented is not necessarily that of the real-world, but one close enough for the human eye to tolerate. Real-time rendering is often polygonal and aided by the computer's graphics processing unit (GPU), which is a specialized parallel processing unit, very efficient for manipulating computer graphics and calculating floating point operations. The rendering is done by a rendering engine, which makes use of the hardware acceleration on today's graphic cards via an Application Programming Interface (API). We trust that readers will appreciate that this subject cannot be covered in great detail here, and refer to [AMHH08, Real-Time Rendering 3^{rd} edition].

Direct3D, OpenGL (API)

A 3D API provides a software abstraction of the GPU or video card, hence it contains many commands for 3D rendering and exposes the advanced graphics capabilities of 3D graphics hardware, including z-buffering, anti-aliasing, alpha blending, mipmapping, atmospheric effects, perspective-correct texture mapping and the possibility to execute shader programs on the GPU.

Rendering Engine (Middleware)

A rendering engine can be seen as a package that provides convenient functions for complex and frequently used processes such as loading textures, geometry, or scene data, with the main purpose to ease programming, and to avoid having to reinvent the wheel for each new application. Some of the noteworthy engines are OGRE, Shark 3D, Unreal Engine, OpenSceneGraph, CryEngine and id Tech.

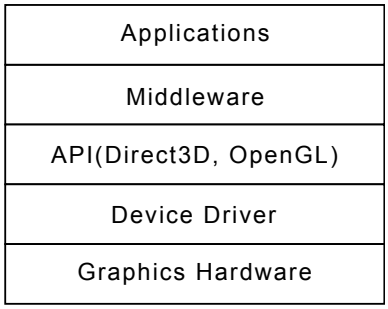


Figure 3.1: Abstract Layer model for real-time computer graphics

3.2 Audio Rendering

Audio rendering means to place a signal, e.g. the prerecorded soundfile, at a defined position in the VR scene, and use the various cues (ILD, Doppler Effect, Distance Attenuation, see Chapter 2.3) to create the impression that the virtual sound truly emanates from its position in the virtual scene. As our work is about the impact of spatialized audio rendering compared to stereo, this chapter deals with the process of stereo sound rendering, and binaural synthesis or spatial audio rendering.

3.2.1 Stereo Sound

Stereophonic sound, commonly called stereo, is the reproduction of sound using two (or more) independent audio channels through a symmetrical configuration of loudspeakers in such a way as to create the impression of sound heard from various directions, as in natural hearing. The reproduction uses a psycho-acoustic phenomenon, namely that humans can discriminate the direction of a sound source in terms of left-/right due to the interaural level difference between the two channels (see Chapter 3.2.2 for details). This is done by calculating the angle between the normalized position of the sound source and the orientation of the listener's head. The value determines the amount of sound heard by the left and the right ear. A schematic depiction can be seen in Figure 3.2. The two values for the left ear l and the right ear r are computed as follows:

$$l = ((|\vec{a}| \cdot |\vec{c}|) * (-0.5) + 0.5)$$

$$r = 1.0 - l$$

φ	l	r	φ	l	r
0°	1.00	0.00	180°	0.00	1.00
45°	0.85	0.15	225°	0.15	0.85
90°	0.50	0.50	270°	0.50	0.50
135°	0.15	0.85	315°	0.85	0.15

Table 3.1: Stereo sound levels for l(eft) and r(ight) ear; φ in clockwise direction

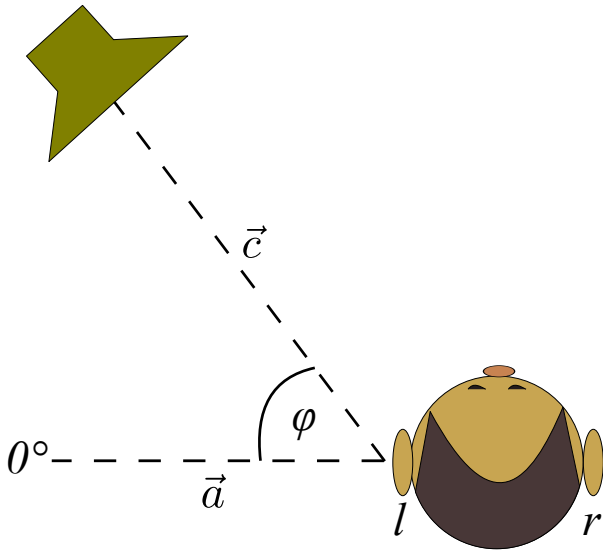


Figure 3.2: Stereo sound level determination; a is polar axis, φ is measured in clockwise direction

3.2.2 Binaural Synthesis to Spatialize Sound

The simplest way to produce 3D sound is to position loudspeakers at many different locations in space. However, this multi-channel approach is both cumbersome and expensive. Fortunately, because we have only two ears (binaural), it is possible to generate full 3D sound using only two channels. This can be achieved by using certain binaural transfer functions. Generally speaking, a transfer function is a mathematical representation, in terms of spatial or temporal frequency, of the relation between the input and output of a (linear time-invariant) system. The transfer function characterizes the direction dependent filtering of sound waves by our two ears.

Furthermore head, torso, hair and cloths have an influence on the binaural transfer function, as they diffract, reflect and shadow certain frequencies [Len08]. The shadowing of the head has an influence at high frequencies, whereas diffraction influences the sound field in the low frequency range. These characteristics are direction-dependent and enable us to perceive the origin of a sound event. As the last step of the transmission to the ear drums, the ear canal brings a rise of the transfer function at approximately 3 kHz. The directional and non-directional components of the transfer function and the affected frequency ranges are depicted in Figure 3.3, ordered by importance from top to bottom [Gie92].

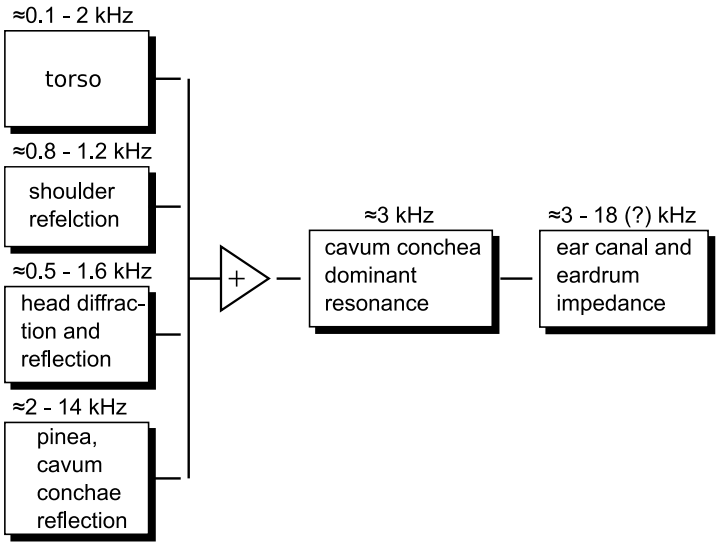
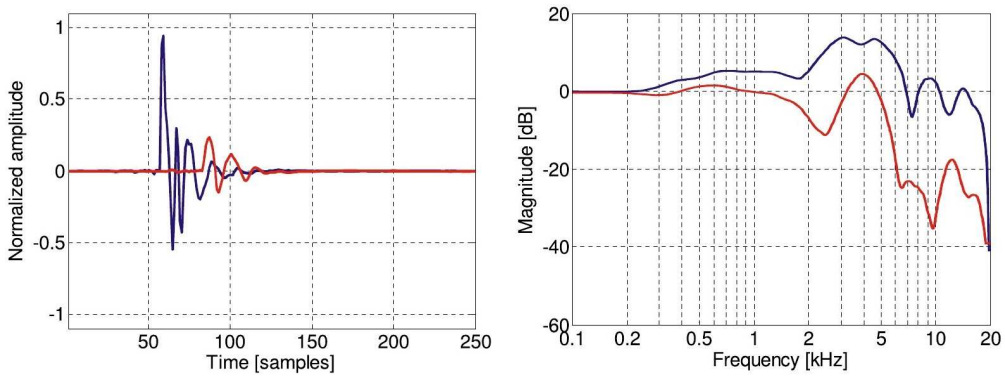


Figure 3.3: Gierlich's description of directional and non-directional components of HRTFs; the range of frequencies most likely affected by each stage is indicated

From HRIR to HRTF

When a representation in time domain is used, the transfer function is called Head-Related Impulse Response Function (HRIR). A sound source recorded by two microphones placed in the ear canals is represented in the HRIR and carries information about the interaural time difference (ITD). The equivalent function in frequency domain is called Head-Related Transfer Function (HRTF) and is obtained by applying the Fourier Transform to the HRIR, revealing the interaural level difference (ILD) [Len08]. Both functions are depicted in Figure 3.4.

Later the HRTF, which is the key to 3D-sound reproduction for imitating real-world hearing conditions, is used during audio rendering a process called binaural synthesis. The simplified pipeline, as shown in Figure 3.5, takes a mono audio signal as input in frequency domain, multiplies it with the values of the transfer function for each ear, performs an optional crosstalk cancellation and outputs the sound. The listener experiences the synthesized sound as a virtual sound source emitting from the desired location in space. We do not intend to use a multi speaker setup, because of its disadvantages compared to headphones, as discussed later in this chapter, are too grave (Table 3.2).



(a) HRIR in time domain (b) HRTF in frequency domain

Figure 3.4: Recorded sound event at the (near) left ear (blue line) and at the (far) right ear (red line) for a single sound source at a specific position [Len08]

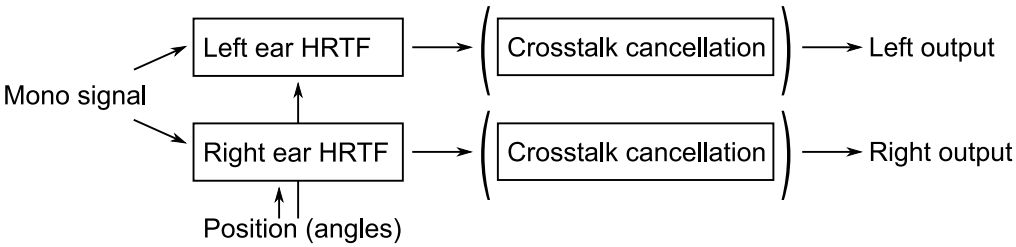


Figure 3.5: Simplified binaural synthesis pipeline; crosstalk cancellation is optional as it is only necessary for loudspeaker setups

Obtaining HRTFs

The HRTF serves as a kind of unique spectral fingerprint [Beg00], which characterizes the filtering of our ears. As would be expected for an anatomical property like the HRTF, every person has differences in the shape of the head, ear and physical characteristics. Therefore measurements have to be done either for each one individually (personalized HRTF) or with a dummy head (non-personalized HRTF), by playing an analytic signal at a desired position at a certain distance in an anechoic chamber and recordind the response function with two small probe microphones placed at the entrance of the ear canals, since all direction-dependent filtering is applied to the signal at this point (see Figure 3.3). The usage of an anechoic chamber for optimal measurement is costly, but strongly advised to minimize unwanted environmental reverberation in the HRTF. Reverberation serves as a sound cue but, as it is different for every kind of room (and almost non existent outdoors), it is not the intention to record this characteristic. Reverberation is typically applied during rendering as one of the last steps.

Measuring Personalized HRTFs

For measurement, the positions and orientations of the sound source are defined in a coordinate system relative to the listener. This is closely related to a horizontal coordinate system, which is called celestial coordinate system in astronomy, and uses the observer's local horizon as the fundamental plane to express point coordinates through two angles. One is the azimuth angle in a horizontal plane and the other is the elevation angle in a median plane as displayed in Figure 3.6. The HRTF depends on those two variables, but is additionally constrained by frequency and distance between source and listener. To compensate the latter, a distance greater than one meter is used and at every required position the measurement has to be repeated with specific angle increments typically between 15° to 30° according to the desired precision [Len08]. This is accomplished by moving the analytic sound source or the measured head in the corresponding direction or by using a multi-speaker setup. Either way the measurement is a rather expensive process both in terms of money and time, and researchers are still exploring new equipment and techniques to minimize measurement errors and variability [Beg00].

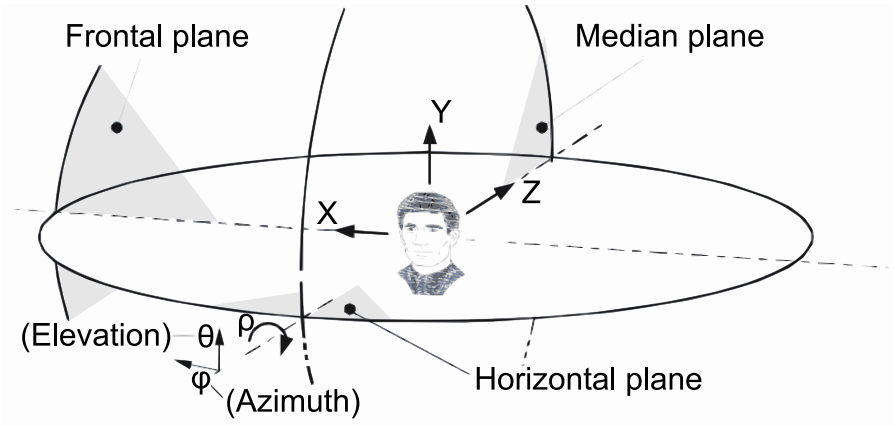


Figure 3.6: Azimuthal coordinate system relative to the listener; positions of sound sources are described by two angles [Len08]

Other Ways to obtain an HRTF

In many cases measuring an individual HRTF for various subjects will be impracticable. Even with the proper hardware available, the procedure takes very long depending on the desired precision and is susceptible to errors, resulting from minor

head movements during that time, or to calibration mistakes. However, the use of non-personalized HRTFs is strongly discouraged due to the poor localization accuracy, although for many applications a generalized HRTF is preferred. One possibility to achieve a higher accuracy is averaging of HRTFs by examining spectral features in frequency domain [ST74], or to statistically evaluate them with principal components analysis (PCA) to isolate spectral features that change as a function of movement and those that remain more constant.

But averaging does not have to rely on spectral features alone. HRTFs could also be averaged by analyzing physical features. Such features, called anthropometric features, are for example the location of the entrance of the ear canal, the pinna, or the size of the cavum cochlea of different people. To increase accuracy of a personalized HRTF, [ZHDD03] propose a test and strategy based on matching such anthropometric ear parameters with an HRTF database where each measured HRTF is categorized according to physiological features. Their results show that localization and subjective perception of the virtual auditory scene was improved. Similar to them [HLJ+06] present a method to customize an HRTF by applying multiple linear regression analysis on the function together with correlation analysis of anthropometric parameters, which leads to better localization accuracy of personalized HRTFs compared to non-personalized HRTFs. A very recent approach to obtain a personalized HRTF is stated by [MTNK08]. Instead of recording the impulse response function for each subject, they measure the head and ear morphology by magnetic resonance imaging (MRI) and then use the 3D data in a computer sound wave propagation simulation by the Finite Difference Time Domain method [XL03].

To use an MRI is almost as costly as the unechotic chamber. [DPT+08] proposed a system to reconstruct head models from photographs. Starting with five photos and some key-points indicated by the user, their system extracts information from the images and deforms a 3D dummy head to represent user head features. They compared their method with laser scanned 3D head models and performed a virtual sound scattering. Results indicate that their approach is robust, fast and reliable enough for personalized 3D-audio processing and HRTF generation.

A different way to find a good fitting HRTF for a subject is described by [MBT+07], who take a set of exemplary HRTFs and let participants perform a ''point and click'' pretest in order to find the most suitable one for each candidate. The subset should be chosen wisely. If it is too large, the process will be too long and exhausting, if it is too sparse, no

speaker setup	**headphones**
+ Applicable for many listeners	+ Predictability
+ Cheap for crowds	+ Less overspeaking from surrounding
+ Can combine surrounding acoustics	+ Can use personalized HRTF
	+ No crosstalk
- Complex crosstalk cancellation	+ Controlled playback situation
- Need to use universal HRTF	
- Sound experience position dependent	- Typically single user setup
- Unwanted reverberation artifacts	- Expensive and complex for crowds
- Must be adapted to room properties	- Need to wear headphones
	- (acoustical) isolation

Table 3.2: Pros and cons of multi speaker setup compared to headphones [Beg00]

fitting HRTF will be found. This demands for easily searchable databases where different feature requests lead to satisfying results, but the few sparsely available collections on the Internet contain few HRTFs, are not regularly updated and not searchable.

Remaining Parts of Binaural Synthesis

As the HRTF is the key for realistic spatial sound synthesis, the rest of the pipeline such as computing distance attenuation, applying Doppler effects etc. is straight forward and we will not go into detail about it here. Considering the actual output of the signal, headphones provide an ideal single user setup due to their predictability of sound. For the advantages and disadvantages of binaural synthesis via headphones compared to multi-channel loudspeaker playback (i.e. surround sound) see Table 3.2.

We mentioned above that HRTF generation is still an important field of research as successfully synthesizing and imitating real hearing depends on its quality, and the solution to the individual difference problem of each person's physical properties. Perhaps one day we shall all routinely visit the HRTF-metrist who will fit us with individual transfer functions which can then be used in all our every day task facilitation and entertainment devices.

Chapter 4

The Role of Audio in Virtual Environments

Until now, we regarded sound as a source of information for our sensory system. We discussed the various aspects of our auditory perception and the two concepts of sound rendering. Right now it is time to focus on previous work studying the use of spatialized audio to support the feeling of presence, the implications on task-performance, and the auditory task facilitation. Finally we address experiments and VE applications concerned with pedestrian safety and elaborate their connection with our work.

4.1 High-Quality Sound Increases Presence

Presence is the term used to describe the sense of ''being in the world'' in a computer generated model, and is strongly influenced by the immersion of the VE and the fidelity of the presented stimuli. [HB95] state that one's feeling of presence significantly increases by adding spatialized sound to a stereoscopic display, but the addition of spatialized sound did not increase the overall realism of that environment. As they discussed one explanation could be that their questionnaire used the term of ''realism'' with some semantic load, which would imply visual-realism to the user's mind. Spatialized sound does not change the visual representation, hence the user's visual perception of the scene does not change and their feeling of realism for the scene remains the same. However, participants acted more seriously in a VE of high fidelity. With photo-realistic visual scenes at hand, [RVSP09] describe the influence of auditory

cues on the visually-induced self-motion illusion in VEs, also known as ''circular vection''. As their study demonstrates, adding HRTF-based spatialized sound rendering can be used to improve one's feeling of presence even when the visual representation is already of high realism. This concurs with [SZ00], who describe the quality of realism in a VE as a function of both auditory and visual display fidelities inclusive of each other rather than realism being a function of both modalities mutually exclusive of each other by evaluating combinations of high-quality (HQ) and low-quality (LQ) visual/auditory displays.

4.2 High-Quality Sound Increases Task Performance

To investigate the effect of audio rendering on user performance, [DSP+99] carried out a study in a recall and recognition task of different objects in a VE. Their results give evidence that task-performance can depend on the audio quality. Furthermore, LQ visual displays can be compensated by higher quality audio rendering. This was again picked up by [LVK02], who performed a similar experiment where no significant effect on task-performance was found, instead subjects again reported a higher degree of presence with the HQ auralization[1] VE: they were more concentrated on the task and enjoyed the VE more than participants assigned in the low-quality auralization VE and as a conclusion the synergy between task-performance and presence strongly depends on the task itself.

4.3 Auditory Task Facilitation and HCI

Another domain that shows the advantages of using the auditory modality would be task facilitation. For example, this is applied in its simplest way by using the audio channel to give certain feedback (e.g. a peep tone) to aid users in their task. Many different fields such as Human-Computer Interaction (HCI), UI feedback agents for handicapped people, visual displays, key finders, electronic parking aids, electronically

[1] Auralization is the process of rendering audible, by physical or mathematical modeling, the sound field of a source in a space, in such a way as to simulate the binaural listening experience at a given position in the modeled space

guided walking sticks for the blind [DHJA01] etc. benefit from this principle. To evaluate the impact of audio feedback compared to visual feedback on task-performance, [ZF03] described a VR setup with a system to track user actions in a VE. With this system they later demonstrated that compared to visual feedback, audio feedback increases performance in an assembly task [ZFXT06]. To preserve attentional resources for visually monitoring task relevant objects, auditory feedback can be perceived in parallel, allowing people to be aware of background information [Ale04]. For cars equipped with auditory displays using spatialized sound cues to represent menu items, [SDTB08] reported a significantly better driving performance compared to standard visual displays. Auditory displays can also compensate for visual impairments of computer users, especially when icon sounds are spatialized [BJH07]. Advantages have also been reported for spatialized audio cues in Augmented Reality applications [STG+06, SWC+03]. Applications of auditory displays re-synthesizing the external audio scene of a vehicle have been engineered for airplanes [Del00], mostly in order to improve combat performance through audio cues to detect the direction of threats and targets [ADS03].

4.4 Bimodal perception

Since the trend in neuro-science is towards a multimodal understanding of perception, accounting for intermodal interactions of senses (e.g. [SM93, EB04]) and the real world linkage between sound and visual stimuli, recent hypotheses state that task facilitation may also result from an audio-visual perceptual integration. A bimodal task facilitation is assumed due to the existence of bimodal cells which process spatio-temporally correlated stimuli from different modalities in a unified manner [KP85]. A bimodal integration can resolve unimodal ambiguities at an early stage and enhance localization and orientation behaviors. Concerning effects of bimodal integration, HQ spatialization of sound can be critical to convey a sufficient spatial congruence between auditory and visual stimuli [MWRZ05] (see Chapter 2.4).

4.5 Pedestrian Safety and Behavior

The area of perceptive processes of pedestrian behavior was the aim of several interdisciplinary studies such as psychology [MD01, TD04], accidentology (e.g. [CA05]), transportation planning (e.g. [Fru71]) and traffic simulation (e.g. [YDW+06, BKSZ01, WR04]). Much research is conducted in the area of pedestrians' road-crossing with regard to age related differences, gap acceptance, crossing duration or waiting time [LYM84, CCPI98, TVVdKBS05, TLO08, PKC07]. [CCPI98] concluded that children at the age of 12 selected more safe gaps than younger children, who easily tended to overestimate their potential and speed.

A study that incorporates the performance of children and adults was carried out by [TVVdKBS05, PKC07], [TVVdKBS05] designed an application in a VE with a bike approaching the pedestrian at different velocities and distances. The performance of groups concerning accuracy and safety did not differ significantly, although younger children tended to be more cautious and all groups adjusted their crossing time to the time to contact (TTC) [TLO08] of the bike. [LYM84] made an experiment with children between the age of 5 and 9 where they set up a ''pretend road'' in parallel to an actual road and used the real road's vehicles for the task (i.e. to cross the ''pretend road'' gap before the real road vehicle crosses their way). Although the kids were cautious in general, they accepted some too short gaps.
As a conclusion, one factor about road-crossing in connection with accident prevention that can be identified is the pedestrian. To lower the accident rate, [BSM07] picked up the pretend road technique of [LYM84] to design a training for children to ensure their pedestrian behavior in a real world setup, as they are one of the most threatened group in urban traffic due to several risk factors [ST96]. [BSM07] reported better performance comparing results before and after training. The usage for pedestrian training issues found great reception [OCW+08, MMP02, BKWJ06] not only for teaching children, but elderly people too [CLDV09, TKK04, NKW00].

Perceptual studies about TTC and point of collision were conducted by [SO90]. They examined the TTC estimation of men, women and blind people by showing movies of approaching vehicles in three different conditions namely, visual stimulus only, audio stimulus only, and audio-visual stimuli together. Interrupted abruptly by a blue screen, the participants had to estimate the time to collision. Their results show two fundamental things. First, people are capable of estimating point of collision and time of collision

just by hearing as long as point of collision is not too far in the future and second, sighted people are more cautious in their judgment and tend to underestimate.

The next section tries to give an insight about past and current studies of pedestrian behavior and is meant to be the bridge to our work.

Previous gap experiments

This section gives an overview about studies that are closely related to our work. Summarizing this field is not an easy task as most of the studies are analyzing gap crossing behavior of children compared to adults, adults compared to elder people or people who suffer a certain disease compared to healthy ones. None of them regard sound as the important part of their work and this aspect therefore not mentioned in detail or neglected. Nevertheless we picked up interesting findings and used some of the proposed methods and principles for our application (compare Chapter 5.3.3).

The studies can be categorized in two main parts:

- estimation tasks
- gap crossing tasks

Estimation tasks of time to contact (TTC) are conducted since the 1970s [SD79], when they showed prerecorded video material of oncoming cars. With the computational increase of today's hardware and lower prices, interactive gap crossing tasks become more and more attractive as they give insights about more behavioral variables of the pedestrian before and whilst crossing the street. That does not mean that estimation tasks are inferior. Which one to use depends mainly on the application-requirements and domain.

[SAB07] can be seen as a classical example of the estimation task category. They carried out a study investigating the effect of long range TTC vs. short range TTC, vehicle model, observer's viewpoint and immersive view on TTC estimation. A desktop environment was used with an enclosure to limit participants' field of view on task relevant regions of the screen (see Figure 4.1) and a head mounted display for the immersive viewing condition. According to the experimental setup, no audio was used or at least not mentioned. The results are vital for our study as TTC is necessary to judge whether it is safe to cross or not. Concerning the vehicle model, no significant difference could

be observed as well as for the different viewpoint (on the street vs. on the sidewalk). We picked up on these results to confirm our decision to use only one vehicle model for our experiment design.

Another representative study of the category estimation task in conjunction with pedestrians is [TLO08]. They used prerecorded road environment videos (with vehicle speeds of 40, 60, 80 km/h and different time gaps of 3s, 5s, 7s) to analyze the effect of age on road crossing behavior. Three 32" LCD monitors were used to show a frontal and two lateral images of the scene. For each participant two types of walking speeds (fast and slow) were recorded in order to determine a safe crossing. The subject had to press a button on the keyboard indicating if it is safe to cross, the outcome was calculated afterwards. Auditory representation was not explained. Their results show evidence that pedestrians walking across the road base their decisions on the distance between him/her and the position of the vehicle, which might cause the pedestrian to overestimate the distance when the vehicle speed is increased. Our interactive application shares the large display region to generate a high degree of immersion, the binary user input (start walking), the idea of gaps between the vehicles expressed in seconds, and a moderate driving speed of 50 km/h.

Concerning the gap crossing tasks, [BKWJ06] explored the effect of VE for training children safe road crossing and verified the effect in real crossings afterwards. Their simulation consists of nine successively graded stages of difficulty that provide users with an opportunity to decide if it is safe to cross a virtual street while their avatar is standing on the sidewalk. The scene is displayed in third person view. Stages differ in the number of cars appearing and their driving speed. Keyboard buttons are used to turn the head from the left to the right and to cross the street. After a success, the user proceeds to the next stage. In case of an accident there is a braking sound and the user is reset to the beginning of the same stage. Although our study is not about training, their model of the VE and task design has influenced our application. The consecutively driving vehicles with different gap sizes in between, the feedback and reset mechanism as well as the idea of an avatar is similar to our design. Apart from that, the effect of their training on real-world crossing conditions show us the efficiency of VE applications on gap crossing tasks. As in [SAB07, TLO08], there is no hint about their used auditory representation.

Most of our terminology like safe/unsafe gap, safety margin, safe crossing window is from [CRO06]. In an immersive virtual environment containing a straight, flat section

of road, a street light, a tree, sky, roadside grass and vehicles displayed with a Virtual Research Systems V8 head-mounted display (HMD), participants encountered a line of 11 oncoming vans of different colors in the 1^{st} person perspective. The HMD is equiped with a 48° FOV and Sennheiser HD25 headphones, but they only mention sound as a feedback technique for an unsafe crossing. Different chassis colors were found to be a good method to introduce believability to the VE and therefore adapted by us.

Using a 3D sound-rendition system, [CLDV09] examined the effect of age, and the approaching vehicle's speed on crossing behavior in an interactive street crossing simulation. Participants were asked to cross between two approaching cars if they judged crossing possible. Speed (40 to 60 km/h) and time gap between cars (from 1 to 7s) were varied. Cars were moving from the left to the right in reference to the participant positioned on the edge of the sidewalk facing the experimental road. Participants' view was turned to the left on the simulated road environment and the oncoming vehicles. The near position of our avatar to the road and the adjusted viewport to the left in the direction of the consecutively driving vehicles were mainly inspired by their study.

(a) (b) (c)

Figure 4.1: (a) pedestrian experiment screenshot from [BKWJ06], (b) experiment setup of [SAB07], (c) immersive screen setup by [PKC07]

This is only a subset of important studies our work is based upon. As one can see, the design of an interactive VE gap crossing experiment can be done in different ways depending on the focus of the work. Pedestrian behavior and safety are an active field of research. This thesis presents the first experiment that investigates the influence of HRTF filtered audio spatialization in a combination of audio-visual stimuli on task-performance, which involves complex multi-modal perception such as arise for situations like traffic in a large screen immersive VE.

Chapter 5

From Reality to Virtuality

Before we explain our gap crossing experiment, we would like to describe the required steps from the application description of the target protocol to the final simulator. This is supposed to make a deeper view of the work possible and shall furthermore help readers to understand the various difficulties we ran into, and had to deal with.

5.1 Original Idea

As stated before, our application's goal, as part of the CROSSMOD project [Xmd] (see Chapter 1.1), are behavioral studies and psychophysics experiments, based on a neuro-science literature study, conducted in order to examine a set of crossmodal phenomena that can be potentially exploited to improve quality and efficiency of VEs. Furthermore the idea is to assess the crossmodal knowledge and technologies developed during the project in the context of selected use cases exemplifying realistic applications, and to demonstrate the potential of high-fidelity VEs in order to overcome user interaction limitations such as ''precomputed only'' walkthroughs for future projects.
The scenario for our application is described by the target document as follows:

Target Document

''The scenario will comprise an outdoor urban scene, consisting of a road, a variety of oncoming traffic and a number of buildings and (potentially) other pedestrians. The sounds present will be the ambient sound of the environment, at different levels (i.e.

Figure 5.1: Car approaching at 40 km/h at 1sec intervals

according to the density of traffic and surrounding human activity), plus the sound of the oncoming cars. The interface will be the same as a first person shooter game. The user will approach the edge of the road and look left (or right) at the oncoming car and then push a button to indicate ''cross'' or ''don't cross''. The oncoming car will start the scenario at a predetermined place and approach the point of crossing at a specified speed. The parameter to be considered is the speed of the car: 40, 50, 60, 70 and 80 km/h, with the user deciding whether it is:

- Safe to cross
- Not safe to cross

If the gap between the car and the observer is safe to cross we refer to it as a safe gap. A gap not safe to cross is called unsafe gap. Hence crossing the first can result in a safe (or successful) crossing, crossing the later certainly results in an unsafe crossing (or an accident).

As Figure 5.1 shows, even at the relatively low speed of 40 km/h, the car reaches the point of crossing in under 3 seconds. In the first two images, it can be considered safe to cross, but not in the third. The sound of the car will be spatialized and will correspond to its speed.''

5.2 Implementation

The implementation of the pedestrian simulator application was part of my internship at the Institute of Computer Graphics. Since the 1960s, object-oriented programming languages are favored for simulator design [DN66]. The language used is C++, which is a popular object-oriented language with compilers for various platforms. Furthermore C++ code achieves high-performance compared to JAVA code, as it does not need a Virtual Machine for interpretation at runtime. Visual representation is done by the Object-Oriented Graphics Rendering Engine (OGRE) [Ogr]. OGRE is a powerful and flexible open-source scene graph 3D rendering engine. It provides an easy to use object-oriented interface designed to minimize the effort required to render 3D scenes, and to be independent of 3D implementation like Direct3D or OpenGL, which introduces multi-platform usability. OGRE has a material declaration language which allows material asset outside the code with the support for vertex and fragment programs written in arbitrary languages like Cg, HLSL or GLSL. As it is only a 3D rendering engine, keyboard and mouse inputs are handled by the Object-Oriented Input System (OIS).

Besides keyboard and mouse we use the Nintendo Wii-Remote console controller, as it provides a simple button layout, fits left-handed as well as right-handed people and through being wireless can be used whilst standing. The implementation of the controller is handled by the native C++ WiiYourself library ([Wii]).

The Open Dynamics Engine (ODE), an open C/C++ library, is used for simulating the dynamic interactions between bodies in space. ODE is a fast, flexible and robust engine which is not dependent on any particular graphics package. All objects have an underlying physics shape which is used for collision detection and movement.

For sound rendering, we use the XModSoundLib by INRIA [INR, Xmd]. It is a sound engine which allows perceptual scalable premixing of multiple sound sources [MBT+07] and spatialization through HRTF filtering in real-time. With some modifications, the library is also capable of stereo sound mixing using the same pipeline (see Chapter 3.2.1).

5.3 Application Modes and Prearrangements

To mimic a real-life situation, a virtual urban environment containing buildings, skyscrapers, roads, crosswalks, sidewalks, parks, trees and a sky-blue background at bright daylight was modeled. Several background sounds are placed in the city to emphasize natural surroundings. Cars are traveling on the road with sound sources attached to their engine and tires. We render the scene without shadows, which could provide important depth cues, but for our outside scene, they would be time-of-day dependent.

5.3.1 Simulator Mode

Figure 5.2: In-game simulator screenshot showing street-nodes as white columns

Apart from the target document description, we thought about the simulator and its purpose. First, we created a so-called "Simulator Mode". It provides a game where the user has to follow an overhead cursor pointing to certain locations in the VE with a certain game-time limit. If the street is crossed at a crosswalk, the user gets a better score and some bonus time, otherwise he will lose some playing time. Invisible waypoints, so-called street-nodes, are placed on the street to provide a track for the cars as can be seen in Figure 5.2. The artificial intelligence (AI) knows each successor node, and a car is passed on from street-node to street-node as it moves. At crossroads the car can decide which path to follow. The Shared Space traffic engineering concept developed in the '90s is used to avoid accidents between the cars. The main idea behind it is the

removal of traditional road priority management devices such as curbs, lines, signs or signals and manage traffic by using simple rules like giving way to the right. Cars at crossroads check the street-node map in order to determine if there is an incoming car from the right and stop. For the rare case of four cars waiting at a 4-way crossing, one of them is randomly given way and the traffic jam vanishes. The cars react to the players' presence, and an accident is penalized with play time decrement. After the time is up, the game is over. The achieved score is saved and the application terminates.

Discussions and evaluations showed that this mode is funny to play, but impossible to utilize statistically. Nevertheless it provides a good basis to work on.

5.3.2 Discrimination Experiment, a pilot

Now that we had the foundation, we decided to re-implement the time to contact (TTC) discrimination experiment design by [SAB07]. This pilot should prove our application as an experimental testbed. Compared to [SAB07] we use our simulator instead of prerecorded, randomly interleaved video material. We want to determine the ability of people to discriminate and estimate TTC for approaching vehicles under certain conditions. In this experiment, we examine the effect of sound source clustering. This gives us the opportunity to create the ventriloquism effect (see Chapter 2.4).

Our hypothesis is that sound quality, especially the quality of spatialization, influences TTC judgments. In order to study this, we alter the quality between full spatialization using each sound source as a cluster, and no spatialization by assigning all sound sources to a single cluster [MBT+07].

Participants

All participants were recruited from the university campus. We tested 10 participants in total, 5 for each condition. All participants had normal or corrected to normal vision, reported normal hearing and were naive to the experiment.

Method and Apparatus

Our desktop environment consisted of an Intel Core2Duo CPU T9300 @ 2.5 GHz with 2 GB RAM, an NVIDIA GeForce 8600M GT and a 15.4‘‘ screen at a 1920x1200 resolution. For audio playback we used KOSS Porta Pro headphones.

Procedure

Participants were instructed for the experiment. Each trial consisted of two experiment runs. Participants were presented with a pair of sequences, one for each run. Each sequence showed a vehicle for $2s$ followed by a $3s$ black screen. After a trial participants were asked to determine which vehicle would have reached them first by pressing one of two buttons on the keyboard. They were not allowed to view a trial twice and were not provided with any feedback about their decision.

Velocity of vehicles was randomly selected from the 5 defined ones. Our reference TTC was $4s$. Each pair of sequences contained one vehicle with the reference TTC and one with a TTC altered by a certain threshold. The initial and at the same time maximum threshold was $2s$. Therefore the initial trial would show the first car with $4s$ TTC at a random vehicle speed and the second car with a TTC at either $2s$ or $6s$ at a random vehicle speed. The order was chosen randomly so that participants would not learn whether the first vehicle was the one with the reference TTC or not. For each participant the threshold procedure was a staircase in which the difference between TTCs in a trial was decreased (made more difficult) after two consecutive correct trial outcomes, and increased (made less difficult) after an incorrect trial. The step for increasing and decreasing TTC was performed in $0.25s$ increments.

The experiment ended when a subject reached 8 reversals or gave seven correct answers in a row for the lowest TTC difference $0.25s$. A reversal consisted of an incorrect answer after a prior sequence of two consecutive correct answers, or two consecutive correct answers after a sequence of incorrect and single correct answers (correct answers followed or preceded by an incorrect answers).

TTC and random vehicle speed together resulted in the distance of the observer to the vehicle to vary. This should discourage the use of final vehicle-image size in a sequence for TTC discrimination in a trial.

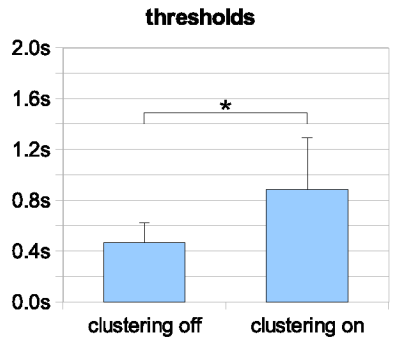


Figure 5.3: TTC threshold of participants in seconds at last four reversals; significance is denoted by "$*$" ($p < 0.05$)

Results

The results show that TTC discrimination is more accurate with full clustering. Full clustering means, that the virtual sound source corresponds with its position in the scene. This minimizes the ventriloquism effect. Compared to clustering all sounds into one single cluster, the reliability is high for both modalities (auditive and visual), which leads to a better performance at TTC discrimination.

As we do not intend to use hundreds of sound sources, the FPS increase by the clustering algorithm is negligible. Therefore we decided to disable it for further experiments. Then the spatial alignment of the sound source with the scene would be more accurate. The pilot shows that sound affects the user-performance in such tasks as TTC estimation. Moreover, we proved that the application can be used as a testbed for crossmodal experiments.

Though we instructed all participants, many reported to be unsure about the task. Some accidentally pressed the wrong button. They were exhausted by the rather boring scenario, not always concentrated and forgot about the previously seen car during a trial.

To make it more attractive, we decided to focus on a street crossing experiment which we denote *"gap experiment"*. It is more interactive, the task is known by everyone, the interaction is not susceptible to input errors, and it is more challenging for the participants.

5.3.3 Gap Experiment Preparations

The simulator mode and the discrimination experiment can be seen as a point of entry for the gap experiment, where we put the focus on a more realistic task with an increased traffic volume, but the scenario, as proposed by the target document (see Chapter 5.1), raises some problems. The following list represents the most relevant points we discussed and adapted:

- Start position of the user
  Because of the concept of placing the user's virtual scene representation at a certain start position, moving towards the street, looking left and right and then decide whether to cross the street or not, people can move to the street and look left (or right) quicker or slower. This affects the distance the car has driven so far as more time has passed and could make it impossible to cross safely because it has already advanced too much and the time to contact (TTC, see Appendix B.1), which was long enough at the beginning of the trial, became too short. The task definition showed that in case of accidents it was not possible to determine if the TTC was judged incorrectly or if the time added to the crossing duration (CD, see Appendix B.1) was too long. The possibility to walk towards the street and look left (or right) adds the time it takes to the CD and therefore differs from trial to trial and part of the original TTC of the vehicle, provided by the experiment design, has already elapsed. This is one of the most crucial latent variables which biases the statistical results in an unresolvable way. Compensation could be using a long TTC, but this would make the decision whether to cross or not obviously easy whereas a too short TTC would lead to poor task performance.
  Therefore we removed looking left (or right), as it was not applicable by untrained people without any computer knowledge and the required time varied too much. Furthermore motivated by [BKWJ06], we fixed the start position on the sidewalk next to the street as seen in Figure 5.5.
- Start position of the car
  The start position is described by the TTC and therefore different for each distance and speed combination, but it would be very easy for the observer to use landmarks and certain environmental features to distinguish between a priori safe and unsafe crossings as shown in Figure 5.1. This together with using only one approaching car could lead to unwanted but successful crossing strategies where

it is not necessary to judge the distance correctly but only to press each button quickly.

- Number of cars
  A big concern about the concept was the usage of only one car for each trial. As stated above it would be very difficult to find an appropriate TTC. A high TTC would lead to fast button pressing as soon as a trial starts without concerning the task, as the car would always be far away and crossing the street would be manageable in any case. A low TTC may not leave enough time to judge the situation as too much time has passed and the attempt would lead to an accident anyway. For the first issue we thought about some sort of ''red traffic light'' which would disable the controls as long as it does not turn green, but the problem of reducing the task to a simple quick button pressing would remain and would only be delayed from the time a trial begins to the moment the light turns green. To deal with this, we decided to alter the concept so that not a single car would approach in each trial but a chain of many cars driving consecutively down the road with gaps in between them passing the observer's position similar to [CRO06]. Now one must judge the situation regarding the gaps between the cars, i.e. whether or not it is possible to cross the street without having to hurry, as the cars are spawned endlessly until a trial is over. To determine whether a gap between two vehicles is safe to cross or not, perceivers must judge the temporal size of a gap in relation to the time it takes them to cross. Therefore, both overestimation of gap size or underestimation of crossing duration can contribute to errors in judging if the gap is sufficiently large.

- Vehicle velocity
  Using different vehicle velocities, as stated by the target document, was found to be very disturbing during the pilot studies, as candidates were not aware of it. They could not suit themselves well to the task and the results were very biased. A vehicle speed of 50 km/h was chosen for the final experiment run, as this corresponds to normal driving speed in cities. As we use more than one vehicle, the gaps in between them, called inter-vehicle distance (IVD, see Figure 6.5), correspond to the proposed TTC. The idea is that pedestrians do not base their decision on whether to cross or not on the time of arrival from the initial start position (spawn point) of the vehicles, but primarily on the distance in between two vehicles and compare those inter-vehicle distances with each other [SJR03, OIF+05]. This is applicable in our scenario as the $safetyratio$ (see Appendix B.1) is cal-

culated by the distance between the nearest vehicle to the observer's position. Figure 6.5 shows three consecutively driving cars with the intended TTC being their inter-vehicle distance as the gap in between them.

- Vehicle model
  A question was whether to use different vehicle models or not. One advantage would be to discourage participants from using vehicle's proportions compared to landmarks as a hint for judging the situation, another to introduce believability to the scenario as there are different cars in reality. When testing the application with three different car models, participants reported some strange crossing strategies. Some said that they were frightened by a particular car model and they would not cross the street with that car approaching them, although all cars had the same length and width and were driving at the same speed. As a result only a single model was used for the experiment, a model of a 1992 Nissan Primera P10 with different textures to vary colors, and sound sources attached for the engine and driving noise.

- Feedback mechanism
  The original concept did not state any feedback mechanism for the user, so it would be just a decision making task without knowing the outcome. This would minimize the learning curve as one could not see whether the crossing was safe or not and this is not needed for the statistics as we know the outcome by computing the $safetyratio$, but tests showed that this misled most of the pilot study participants as they were not sure if they performed correctly and they were puzzled about the whole concept. Using the sound of a honking horn in case of an accident with a car or a charming little beep tone in case of a successful crossing would help, but to make feedback more significant we thought about actually moving the observer across the road in a first person view after pushing a button to start to cross the street. Now the experiment runs would take longer but would satisfy participants with a clear feedback about their decision.

- 3^{rd} person view
  The 1^{st} person perspective for observing the crossing after triggering the walking procedure was finally replaced by a 3^{rd} person view showing an animated avatar walking across the road which can be seen in Figure 5.4, as this would make collisions between the bounding box of the car and the avatar more obvious and would lower participants' fear of the car approaching them directly.

Figure 5.4: In-game screenshot showing the avatar crossing the road

After many cycles between concept evaluation, programming and pilot studies, we managed to implement the target application and stick to the requirements as close as possible, adding some improvements and new ideas. The relation between the final application and the concept can be seen in Figure 5.5, which compares an in-game experiment screenshot with a concept photography.

(a) (b)

Figure 5.5: (a) In-game screenshot of approaching cars, (b) concept photography

Chapter 6

Experiment: Gap crossing

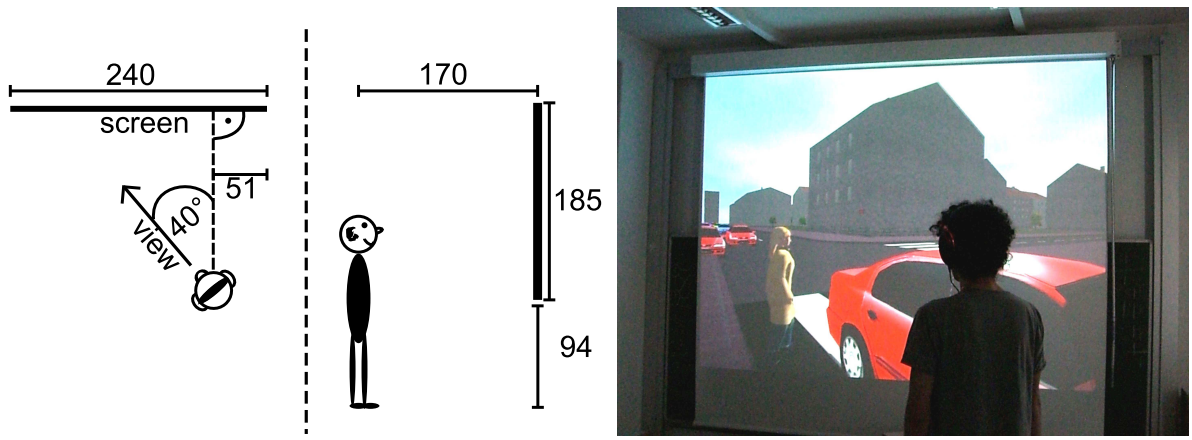
This chapter describes the main experiment. We decided to use a task from daily life, because this realistic scenario is obviously known to all. Furthermore, crossing the street requires no a priori knowledge about the experiment design or the procedure. This enables us to fully assess the effect of binaural sound rendering.

6.1 Participants

All participants were recruited from the university campus or by an advertisement. They were paid a participation fee. Ages ranged from 19 to 31 (mean 24.3). We tested 48 participants in total (24 male, 24 female), 16 for each condition with a 50% fraction of female participants. All participants had normal or corrected to normal vision, reported normal hearing and good health condition and were naive to the experiment.

6.2 Method and Apparatus

To provide a high degree of immersion, the experiments were carried out with a large projection screen (240cm by 185cm) setup. The application was run on a laptop computer equipped with an Intel Core 2 Duo T9300 running at 2.5 GHz, 2 GB RAM and an NVIDIA Geforce 8600M GT GPU, providing sufficient performance to run both audio and video, with minimum video frame rates of 60 FPS and minimum audio frame rates of 90 FPS. For visual rendering, we used the open source engine OGRE [Ogr].



(a) (b)

Figure 6.1: (a) a schematic picture of the experiment setup (measure in cm), (b) a picture of the actual experiment with a candidate performing the experiment

For spatialized audio, we used a custom sound library for rendering and Sennheiser HD650 headphones for playback. Spatialization of sound was achieved with binaural rendering [Bla99] using individual Head-Related Transfer Functions (HRTFs) which were chosen for each participant from a database of example HRTFs. Besides HRTF-based filtering of sound signals, the sound library performs distance attenuation and simulates the Doppler effect for sound sources in motion.
For stereo, the same library with the same settings was used, but HRTF rendering was disabled and directional information was conveyed only by binaural frequency independent level differences, following a simple cosine function for the angle between the listener's orientation and the sound source's position. To avoid the necessity for head tracking, participants had a fixed viewpoint, they were instructed to stand in a fixed position, to look in a fixed direction and to avoid head motions. The viewpoint was chosen during pilot studies in such a manner that the participants had the best overview of the task-relevant screen regions (see Figure 6.1a), while maintaining a high degree of immersion and avoiding occlusion caused by the participant's shadow and the projector. Head position, view direction and the coordinates of the screen corners were used to configure the engine so that camera and listener position of the VE align with the spatial layout of the setup. A 120° vertical FOV was chosen for the camera to match reality. As input device we used the Nintendo Wii-Remote console controller. It has a simple button layout, which is favorable when conducting the experiment with users without computer experience. A picture of the final setup can be seen in Figure 6.1b.

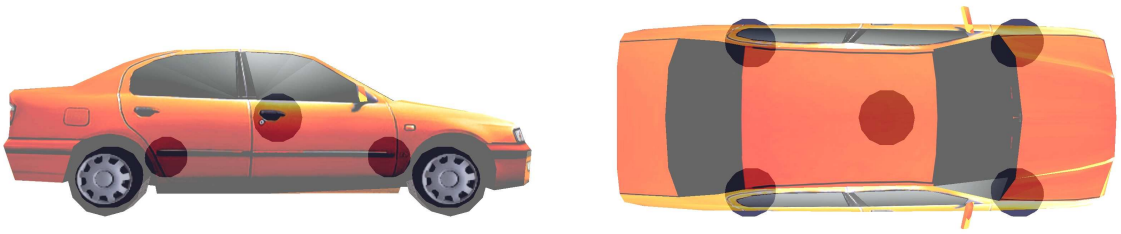


Figure 6.2: Model of a 1992 Nissan Primera with sound sources attached

cue	mute	stereo	spatial
Visual Depth & Motion Cues	+	+	+
Doppler Frequency Shift	-	+	+
Distance Attenuation	-	+	+
Interaural Level Difference	-	+	+
Interaural Time Difference	-	-	+
Monaural Spectral Cues	-	-	+
Reverberation	-	-	-

Table 6.1: Important visual and auditory cues; "+" denotes whether the cue is rendered in the respective condition.

The model of the car is a 1992 Nissan Primera with sound sources to simulate engine- and driving-noise.

6.3 Conditions

The goal of this study is to evaluate the impact of spatialized audio rendering on task performance. We compare the results against a unimodal control condition (only video display) and conventional stereo audio rendering as bimodal control condition. For comparison, in Table 6.1 we listed the most important cues for distance and motion perception and whether they are present or absent in the respective conditions.
Note that reverberation does not appear as a cue in the stereo- nor in the spatial con-

dition. It is not implemented in the engine, as it would mainly be an auditory cue for scenes inside buildings and not for outdoor scenes where almost no reverberation occurs. Furthermore correct reverberation, which would take scene geometry and materials into account following a ray casting approach for sound waves' propagation, is difficult to implement and does not satisfy any cost/performance ratio between complexity and benefits for our scenario.

Each participant performed the experiment in one of the three conditions, which we denote as follows:

- *Spatial*: High-quality spatialized sound rendering with HRTF filtering
- *Stereo*: Conventional stereo sound rendering with low-quality spatialization
- *Mute*: Unimodal baseline condition

The idea of varying between conditions, although it would give us the possibility to perform within-subject analysis, was skipped for the final experiment, as pilot study participants were not comfortable with changing sound conditions and reported, that they adapt their perceptional strategy towards a particular condition, and continued to use this strategy. For instance, when starting with the *Mute* condition, a participant adapts his skills to rely on pure visual properties. This can result in the tendency to ignore audio information provided in the next condition being tested, since attention remains focused on visual stimuli as trained in the previous block.

6.4 Hypotheses

We expect a task facilitation resulting from the perceptual utilization of auditory cues provided by high-quality spatialized sound rendering. The following hypotheses shall be tested to verify this expectation:

- H_1: Better performance in *Spatial* than *Mute* (comparison to unimodal control condition)
- H_2: Better performance in *Spatial* than *Stereo* (comparison to bimodal control condition)
- H_3: Better performance in *Stereo* than *Mute*

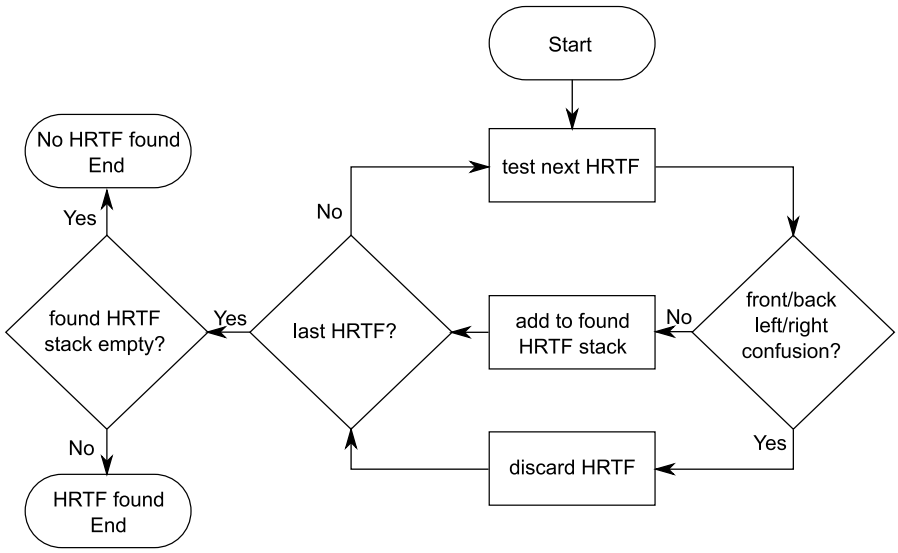


Figure 6.3: HRTF selection procedure flowchart

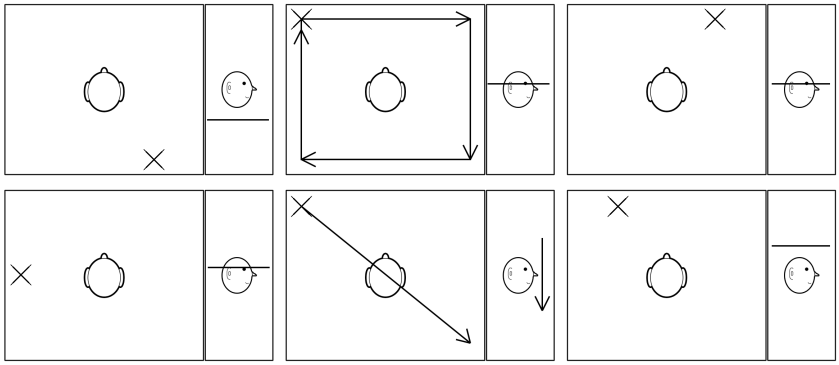
6.5 HRTF Selection

For the Spatial condition, an HRTF is required for the spatialized audio rendering, which should ideally be created for each participant individually. However, measuring an individual HRTF is a costly and time consuming process, requiring expensive equipment and a setup in an anechoic chamber. To avoid this costly method, but not at the expense of inaccuracies due to non-individualized HRTFs, we used a selection of 6 exemplary individual HRTFs selected to be representative from the LISTEN HRTF DATABASE [Lis], and determined the most appropriate for each participant.

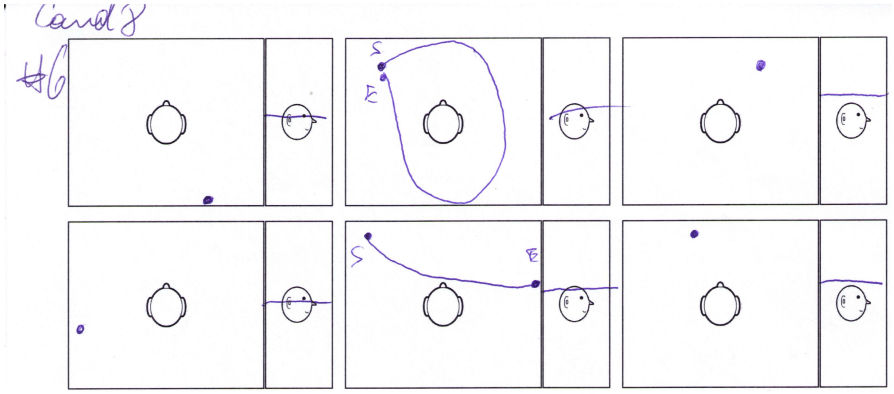
To select an appropriate HRTF, [MBT+07] used an application which let candidates choose an HRTF performing a ''point and click'' pretest. However, after trying this approach we were afraid of subjective mistakes caused by distraction and ventriloquism effects resulting from stimuli on a visual display. Leting candidates decide which HRTF would match them best by moving themselves or the sound source in the scene was found to be unsuitable and time consuming as well. They reported that they could not hear a difference between various HRTFs and we had no possibility to judge their subjective impression. So we designed a formalized selection procedure which requires no display and would work without confusing the participant. Using six candi-

date HRTFs, the method was efficient enough not to exceed the participants' patience and not to exhaust them. For each HRTF candidate, a sequence of six test scenarios was presented in a randomized order. Each scenario contained one sound source which was rendered either on a particular static position (4 scenes) or on a particular trajectory (2 scenes). While listening to each scenario, the participant had to sketch the perceived position (or trajectory) relative to his/her head into a transversal plane for the azimuthal angle and into a sagittal plane for the elevation (see Figure 6.4). Each test scenario was presented until the participant reported confidence about his/her perception, assuring that each participant had enough time to decide carefully. The average time needed for one scenario was about 20s. In the selection we accounted more for deviations in the azimuthal angle than the elevation, since the main movement in the traffic simulation is in the transversal plane. The HRTF with the best fit between actual and perceived angles was chosen. Participants who had at least one clear front-to-back or left-to-right confusion in any of the HRTF sheets were skipped. This strategy was necessary to avoid negative effects resulting from spatial misalignments of the auditory scene and the visual scene.

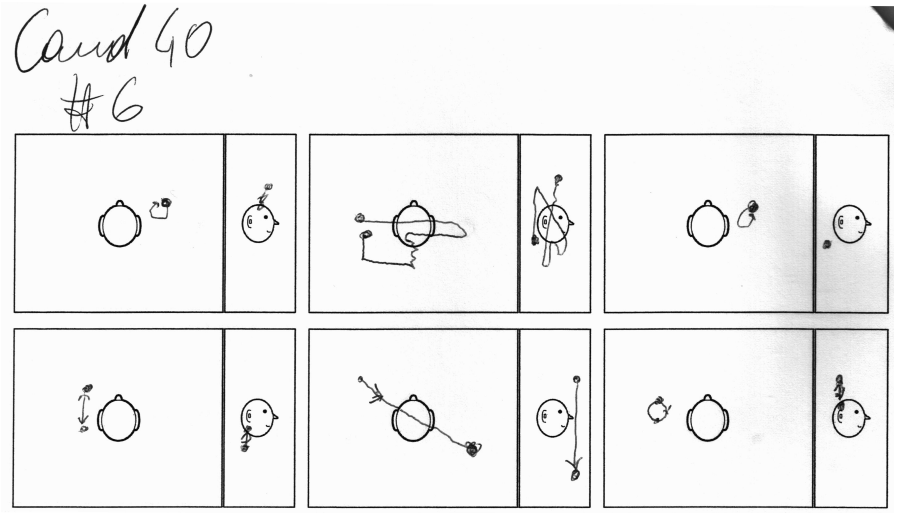
A flowchart explaining the HRTF evaluation is depicted in Figure 6.3. As previously mentioned, we put our main focus on deviations in the azimuthal angle, therefore there is no separate elevation evaluation stage in the flowchart. In the rare case that two equally good HRTFs remain, both are tested again. This time the source is placed at arbitrary positions and the candidate person draws its position into a separate sheet for interpretation.



(a)



(b)



(c)

Figure 6.4: The test sheets used to score the applicability of a particular HRTF: (a) reference sheet, (b) example sheets for a selected HRTF candidate, (c) a rejected HRTF candidate

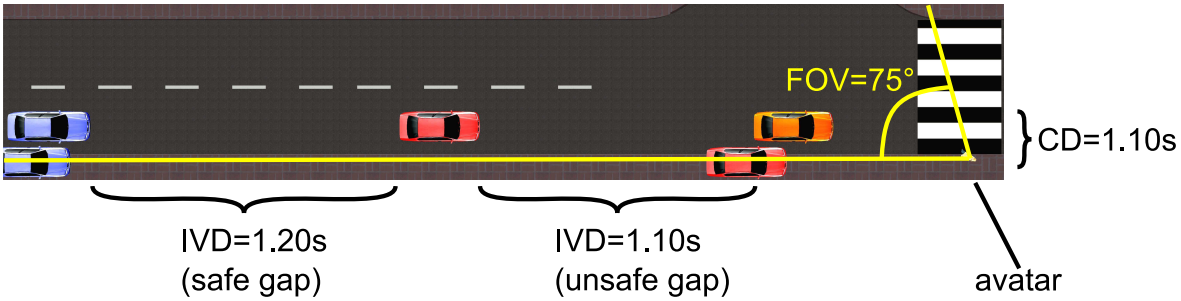


Figure 6.5: Top view with TTC as the IVD between the cars driving at 50 km/h; crossing duration (CD) is based on a walking speed of 7 km/h; the field of view (FOV) is marked by yellow lines; parked cars on the sidewalk serve as occluders

6.6 Procedure

The observer stands on the sidewalk facing towards the double-lane road in a first-person view. Cars come from the left side at specified intervals and a constant speed of 50 km/h. The cars are constantly spawned and do not react to the presence of the pedestrian or any aids such as road signs or crosswalks. If the test participant thinks it is safe to cross, he/she presses a button to start a non-interactive forward movement. After the button is pressed, the camera switches to a third-person view to provide good visual feedback, and the participant can watch an animated avatar crossing the street (see Figure 6.6). The goal is to cross the road several times without being hit by the oncoming cars. After each trial, the screen turns black, a status report about the current progress and a success rate is displayed with a comment complimenting the participant to reward good performance and increase his/her motivation. Note that in contrast to previous gap choice experiments [CRO06], we had no motion tracking system and the movement speed is not under the participant's control.

Thus a short training phase is required to accustom the participant to virtual walking conditions. However, using a constant walking speed reduces the amount of latent variables and we also observe that a change in movement speed during a crossing implies that the participant has chosen a gap which turns out to be unsafe after all. We consider this a "negative outcome" of the trial since the participant decides to walk faster/slower in order to avoid getting hit by a car due to a misjudgment of the situation.

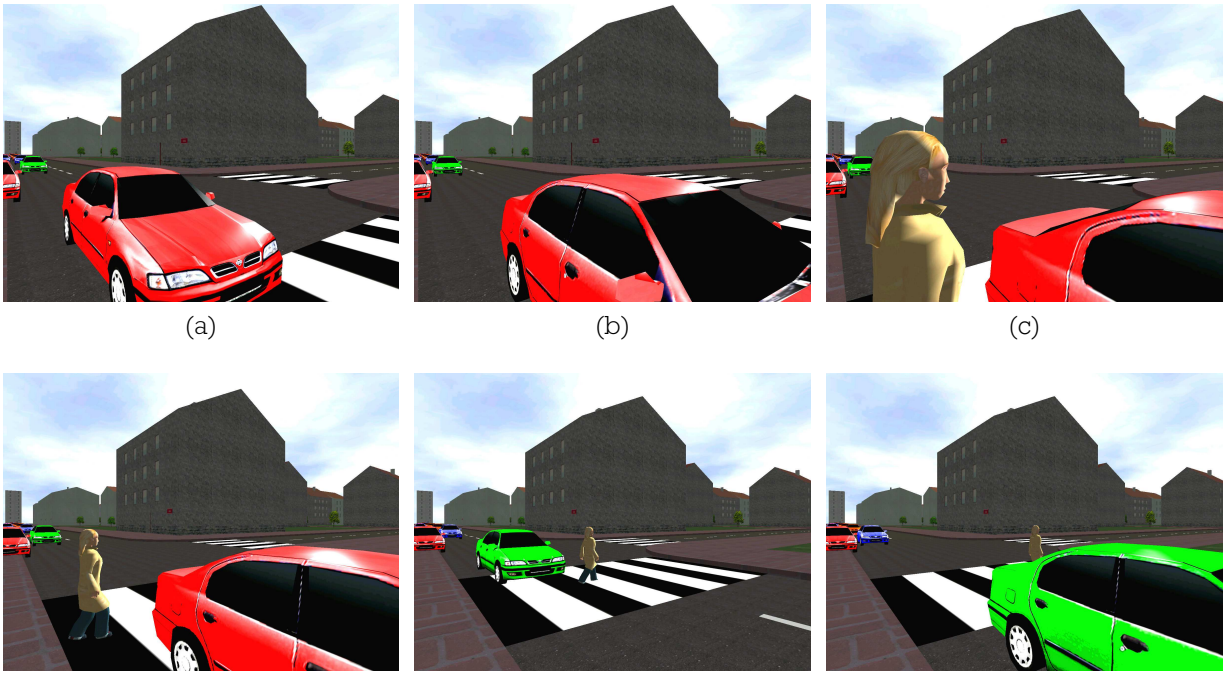


Figure 6.6: Procedure: (a) first person view of car approaching, (b) starting to cross the street, (c) camera switches to third person view, (d) avatar is walking across the street, (e) avatar passes outer bound of car, (f) avatar crossed the road safely

Pilot studies revealed that the effects of the conditions to be evaluated are subtle compared to latent factors introduced by the variety of possible strategies to carry out a gap-crossing task. Hence, the experiment was modified until participants were not able to develop their individual strategy. For this we used only two different gap sizes between cars, both appearing randomly with equal probability. One is assumed to be primarily safe and the other is assumed to be certainly unsafe. Participants were not informed about the distribution or the number of different gaps. To provoke the participants to decide whether to cross the street in a rather intuitive manner, the difference between safe gaps and unsafe gaps was made small enough to be indistinguishable on a conscious level. Pilot testing revealed that a difference of 100*ms* at a car velocity of 50 km/h was subtle enough to emphasize a pure intuitive strategy. Though participants reported that they had the impression not to be able to discriminate safe and unsafe gaps, there is a clear evidence that intuition allowed them to judge above chance level (see Table 6.2). Whether a gap is assessed as safe or unsafe depends on the participant's

	k $(avg\ \#\ of\ chosen\ safe\ g.)$	$B_{N,0.5}(x \geq k)$
mute	36	0.08
stereo	39	0.01
spatial	42	0.001

Table 6.2: Probabilities to observe a result of k or more correctly chosen gaps computed with the Cumulative Binomial Distribution, assuming that participants operate on chance level ($p = 0.5$); k was obtained by averaging the number of correctly identified *safe* gaps over all participants in one group

personal cognitive capabilities and reaction time. A participant familiar with computer games, which demand fast reaction, such as ego shooters, could have a lower safe/unsafe gap threshold than a person without these skills. Therefore, we used a first block of trials for each participant to find the threshold when a gap becomes unsafe to cross. We define a gap as *unsafe* if the probability to cross this gap without being hit by a car is below 10%. During this first block called training phase, participants were presented with a scene where all gaps between cars were equal and they had to cross the street avoiding an accident. Starting with a gap size of 2000ms, size was lowered by 100ms for the next trial in case of a success. At a gap size of 1500ms we reduced the decrement to 10ms for better fidelity. If a car had hit them, the gap size remained constant and the error counter was incremented. The termination condition was that a participant consecutively failed 22 times to cross a gap of one size. To avoid side effects due to frustration resulting from too many failures, two dummy trials with an easy to cross gap size were run after every eight consecutive failures (without the participant's notice) to re-encourage his/her motivation. Statistically speaking, terminating after 22 consecutive failures yields a probability of $B_{22,0.1}(X \leq 0) < 0.1$ (Binomial distribution) that we observe this pattern of failures under the assumption that the actual success rate is $P(success) \geq 0.1$. In other words, we can reject the null-hypothesis $P(success) \geq 0.1$ with a significance level of 10%. This training phase was completed in an average time of 5 minutes. Since participants experienced subsequently more difficult gap sizes, this block of trials already served to sufficiently familiarize them with virtual gap crossing and walking conditions. Finally, we decided to not use an individual unsafe gap size for each participant, but used a general threshold of 1.43s, which was the prime value participants reached during pilot studies. This allowed a better balance between the conditions and better comparability.

	Task A	**Task B**	**Task A∘B**.
H_1(Spatial > Mute)	$t = 3.14, p = 0.002$	$t = 2.06, p = 0.02$	$t = 3.53, p = 0.0007$
H_2(Spatial > Stereo)	$t = 1.81, p = 0.04$	$t = 2.35, p = 0.01$	$t = 3.31, p = 0.001$
H_3(Stereo > Mute)	$t = 1.13, p = 0.13$	$t = 0.12, p = 0.45$	$t = 0.65, p = 0.26$
ANOVA	$F = 4.67, p = 0.01$	$F = 3.09, p = 0.06$	$F = 8.61, p = 0.0006$

Table 6.3: Significance tests (one-sided t-test $df = 30$, one-way ANOVA $df = 2$ for group and $df = 45$ for sample size)

The main stage comprised 60 trials. As stated above participants were not informed that there are only two different gap sizes, where one is impossible to cross without getting hit by a car (*unsafe*) and the other is possible to cross (*safe*). The size of the *unsafe* gap was $1.43s$. The *safe* gap was created by adding $100ms$ to the *unsafe* gap. Participants were instructed to intuitively choose a gap they find safe to cross. For each trial we recorded whether the chosen gap was *safe* and whether the participant was able to reach the other side of the street without an accident.

6.7 Results

To examine the gap crossing experiment, two sub-tasks, both reflected by different variables, were identified.

The first task is accomplished before crossing the street, as a participant has to discriminate *safe* gaps from *unsafe* ones (*Task A*). The performance for this can be measured according to the number of *safe* gaps chosen out of all trials (Figure 6.8a).

The second task is, after a *safe* gap was chosen, to find the best timing to launch the avatar to cross the street (*Task B*). For this purpose the performance was measured by dividing the number of trials where a *safe* gap was selected and the street was crossed successfully through the number of correctly selected *safe* gaps (Figure 6.8b).

The joint performance of both sub-tasks, which we further denote as *Task A∘B*, is scored with the number of trials when a *safe* gap was selected and crossed successfully divided by the total number of trials (Figure 6.8c). Note that a few of the *unsafe* gaps were also crossed successfully. We call them ''lucky accidents''. The explanation is that some candidates accidentally managed to cross the low but still crossable unsafe gap

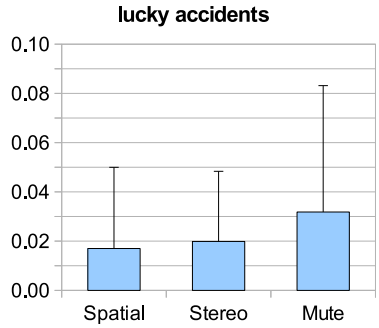


Figure 6.7: Sometimes participants managed to cross the street although the chosen gap was intended to be unsafe; this is caused by mere luck rather than condition

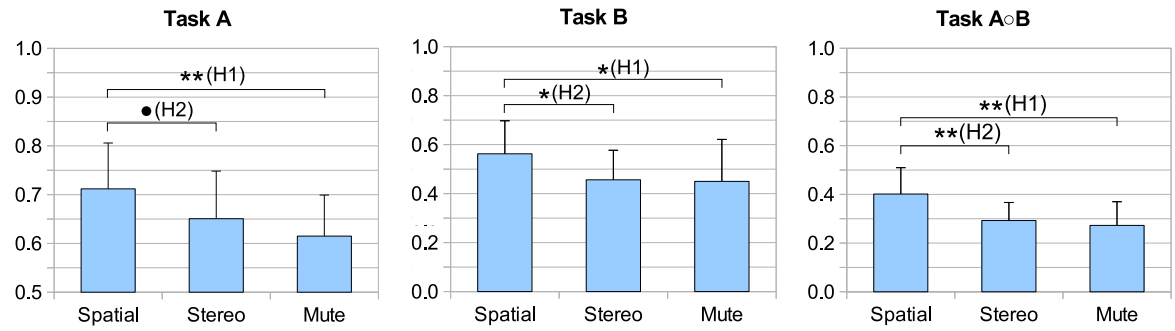


Figure 6.8: (a) fraction of safe gaps chosen, (b) rate when safe gaps were crossed successfully, (c) success rate among all trials; significance of a certain hypotheses is denoted by "●" ($p < 0.1$),"$*$" ($p < 0.05$) or "$**$" ($p < 0.01$)

threshold. The reason for this is mere luck, the amount is nearly equal in all conditions and there is no statistically significant difference towards one of them. Furthermore the number is far below 10% as intended by the design of the training phase (Figure 6.7), so we ignored those trials in our statistics.

The normal distribution of the records was checked with Shapiro-Wilk's test for normality, the equality of variances of the samples was checked with Levene's test. Hypotheses in *Task A, Task B* and *Task A○B* were tested with the Bonferroni-Holm method (Table 6.4). A one-way ANOVA was computed to assess the effect of condition in general. For all three tasks, we observed clearly positive results, supporting the hypotheses H_1 and H_2 that spatialized audio rendering increases task performance, whereas no positive impact of conventional stereo sound rendering (H_3) can be observed in any case.

Task A (♂+ ♀) $p_1 = 0.002 < p_2 = 0.04 < p_3 = 0.13$	
H_1 (Spatial > Mute)	$p_1 < \alpha_1^{**}$
H_2 (Spatial > Stereo)	$p_2 < \alpha_2^{\bullet}$
H_3 (Stereo > Mute)	$p_3 > \alpha_3^{\bullet}$

Task B (♂+ ♀) $p_2 = 0.01 < p_1 = 0.024 < p_3 = 0.45$	
H_1	$p_1 < \alpha_2^{*}$
H_2	$p_2 < \alpha_1^{*}$
H_3	$p_3 > \alpha_3^{\bullet}$

local α values		
$\alpha_1^{**} = 0.00\bar{3}$	$\alpha_2^{**} = 0.005$	$\alpha_3^{**} = 0.001$
$\alpha_1^{*} = 0.01\bar{6}$	$\alpha_2^{*} = 0.025$	$\alpha_3^{*} = 0.05$
$\alpha_1^{\bullet} = 0.0\bar{3}$	$\alpha_2^{\bullet} = 0.05$	$\alpha_3^{\bullet} = 0.1$

Task A∘B (♂+ ♀) $p_2 = 0.0007 < p_1 = 0.001 < p_3 = 0.26$	
H_1	$p_1 < \alpha_2^{**}$
H_2	$p_2 < \alpha_1^{**}$
H_3	$p_3 > \alpha_3^{\bullet}$

Table 6.4: Significance tests (see Appendix B.2) of our hypotheses; three different shades highlight test results with weak (light shade) , normal (mid shade) or high significance (dark shade); note that we used three different significance levels for our hypotheses denoted by "$\bullet$" ($p < 0.1$),"$*$" ($p < 0.05$) or "$**$" ($p < 0.01$)

6.8 Discussion

The results support our hypotheses H_1 and H_2 that for all our three defined tasks (*Task A, Task B, TaskA∘B*) auditory information can be perceptually utilized to improve performance in a task which is seemingly vision dominated and requires visual estimation of complex spatio-temporal relations and optimized action timing, whereas no statistically significant impact of conventional stereo (H_3) can be observed. In particular the auditory cues from high-quality HRTF spatialization are an important factor and according to our results conventional stereo audio is statistically not better compared to no audio.

Our vision normally tends do dominate over our hearing, an example for this can be found in the ventriloquist effect. Recently the term refers to the phrase vision "captures" sound, meaning a good or clear visual stimulus dominates over sound in our perception. For example in television or cinema movies, sound seems to emanate from the actors' lips rather than from the loudspeakers. However, if the visual stimulus is blurred

and therefore not clearly locatable, its influence decreases and the sound captures vision, e.g. a sound event in a badly lit room. Hence bimodal localization is usually better than either the visual or auditory representation alone [AB04]. According to this, we assume that audio-visual spatialization follows the rules of a near-optimal integration of information from both channels. Lowering one modality, so called blurring, introduces uncertainty about the particular stimulus. For audio this corresponds to the spatial resolution of auditory perception. The visual blur consists of the movement of the vehicles, the subtle differences between *safe* and *unsafe* gaps, the perspective changes in size of the vehicles, and the relatively large screen which prevents the user from keeping a long gaze on every task-relevant region of the scene simultaneously. This visual uncertainty cannot be compensated by the auditory cues of stereo sound rendering, leading to a wrong evaluation of the situation by the user.

Due to the relatively acute-angled trail of consecutively driving vehicles, masking effects make it almost impossible to hear more distant cars, resulting in a relatively short time window in which one can hear the close vehicle driving by and the incoming next vehicle simultaneously. This together with the performance increase in *Task A* brings us to the hypothesis, that the user is able to update the internal mental model of the scene representation more efficiently by monitoring the current car passing by aurally and putting the visual focus on the next incoming car. The inter-vehicle distance between the aurally observed car in the peripheral vision and the next incoming, car combined with motion trajectory expectation (eased by constant vehicle speed) provides enough information to distinguish between *safe* and *unsafe* and furthermore allows efficient coordination and timing for a safe crossing of the road. Nevertheless, to confirm this hypothesis, further studies are necessary on the gap crossing experiment with an eye tracking device monitoring participants' gaze for regions of interest evaluation.

Task A ♂ $p_1 = 0.0005 < p_2 = 0.03 < p_3 = 0.102$	
H_1	$p_1 < \alpha_1^{**}$
H_2	$p_2 < \alpha_2^{\bullet}$
H_3	$p_3 > \alpha_3^{\bullet}$

Task A ♀ $p_1 = 0.28 < p_2 = 0.31 < p_3 = 0.47$	
H_1	$p_1 > \alpha_1^{\bullet}$
H_2	$p_2 > \alpha_2^{\bullet}$
H_3	$p_3 > \alpha_3^{\bullet}$

Task B ♂ $p_2 = 0.066 < p_1 = 0.23 < p_3 = 0.74$	
H_1	$p_1 > \alpha_2^{\bullet}$
H_2	$p_2 > \alpha_1^{\bullet}$
H_3	$p_3 > \alpha_3^{\bullet}$

Task B ♀ $p_2 = 0.01 < p_1 = 0.04 < p_3 = 0.15$	
H_1	$p_1 < \alpha_1^{*}$
H_2	$p_2 < \alpha_2^{\bullet}$
H_3	$p_3 > \alpha_3^{\bullet}$

Task A∘B ♂ $p_2 = 0.005 < p_1 = 0.006 < p_3 = 0.45$	
H_1	$p_1 < \alpha_2^{*}$
H_2	$p_2 < \alpha_1^{*}$
H_3	$p_3 > \alpha_3^{\bullet}$

Task A∘B ♀ $p_1 = 0.006 < p_2 = 0.015 < p_3 = 0.17$	
H_1	$p_1 < \alpha_1^{*}$
H_2	$p_2 < \alpha_2^{*}$
H_3	$p_3 > \alpha_3^{\bullet}$

Table 6.5: Significance tests (see Appendix B.2) of our hypotheses (grouped by gender); three different shades highlight test results with weak (light shade) , normal (mid shade) or high significance (dark shade); three different significance levels for our hypotheses denoted by "•" ($p < 0.1$),"$*$" ($p < 0.05$) or "$**$" ($p < 0.01$)

6.9 Other Factors

A deeper investigation of the data revealed that there are also other factors which influence task performance. First we regard the results within the two groups male and female. For women, the hypotheses H_1 and H_2 hold for *Task B* and *Task A∘B*, whereas for men the hypotheses H_1 and H_2 apply for *Task A* and *Task A∘B* (Table 6.5). One can interpret the results in such a way that men are better at identifying a safe gap with spatialized sound, however they make their decision on whether to cross or not too late, which leads to a non-significant result in *Task B* for any condition. As we intended to have a 50% fraction of female participants for every condition, we balanced gender related performance differences.

For the comparison of gender-specific achievement under exclusion of the conditions, it shows up that there is no statistically significant difference for *Task A* between male and female participants, but for *Task B* and *Task A∘B* in favor of the men (Figure 6.10a).

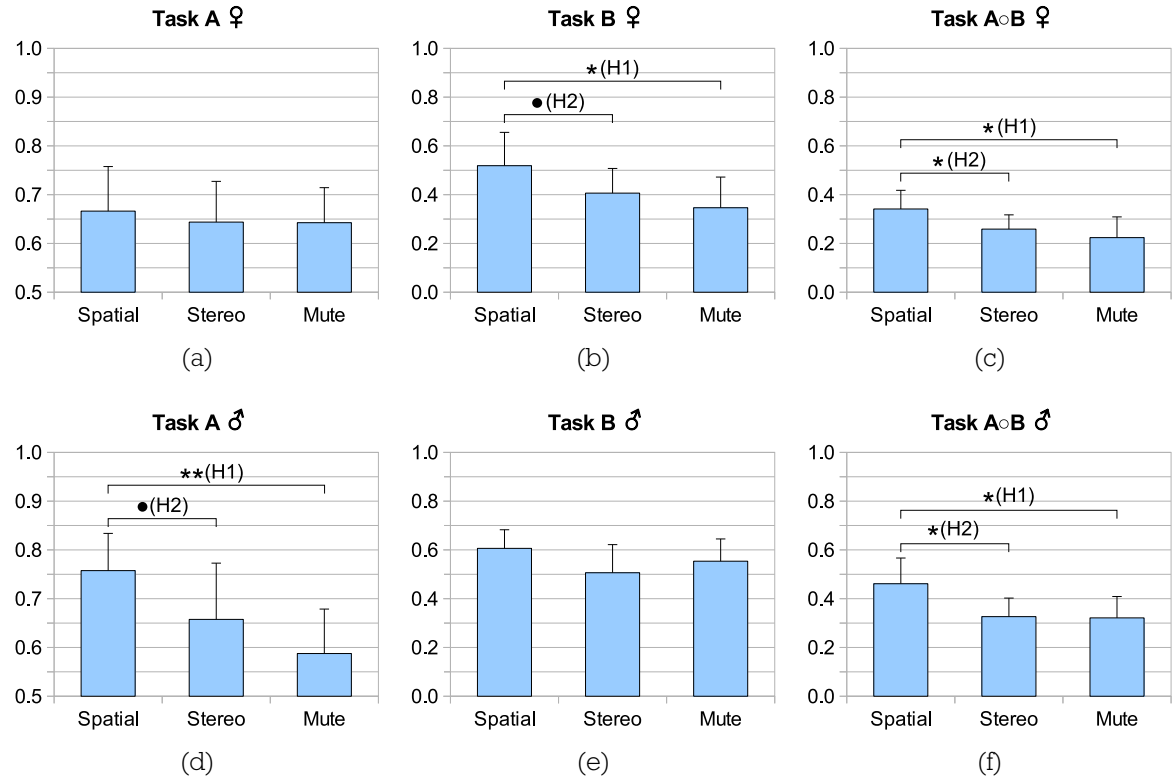


Figure 6.9: (a,d) ratio of safe gaps chosen (*Task A*), (b,e) success rate when safe gaps were crossed (*Task B*), (c,f) success rate among all trials (*Task A○B*); significance of a certain hypotheses is denoted by "•" ($p < 0.1$),"∗" ($p < 0.05$) or "∗∗" ($p < 0.01$)

This could be connected with the game-experience indicated in the questionnaire (Figure 6.10b). Therefore men have a clearly higher experience with computer games than women. This correlation between gender and experience brings us to a more exact analysis of the correlation between experience and performance. The linear regression with trend lines, regression coefficients and probabilities are to be seen in Figure 6.10. No connection exists between experience and performance for men, however there exists a weak coherency for women. The assumption that men with high game experience obtain a better result is not correct. Either their experience is not decisive for their achievement, or they tend to overestimate their experience, whereas women state their computer game skills in a rather modest way. Due to the random assignment, we managed to balance experience over all conditions. At the moment we are not sure about the reasons for gender-related differences.

Nevertheless, experience and gender could be latent factors one has to take care of either by a screening of participants or by a prior test to categorize their potential.

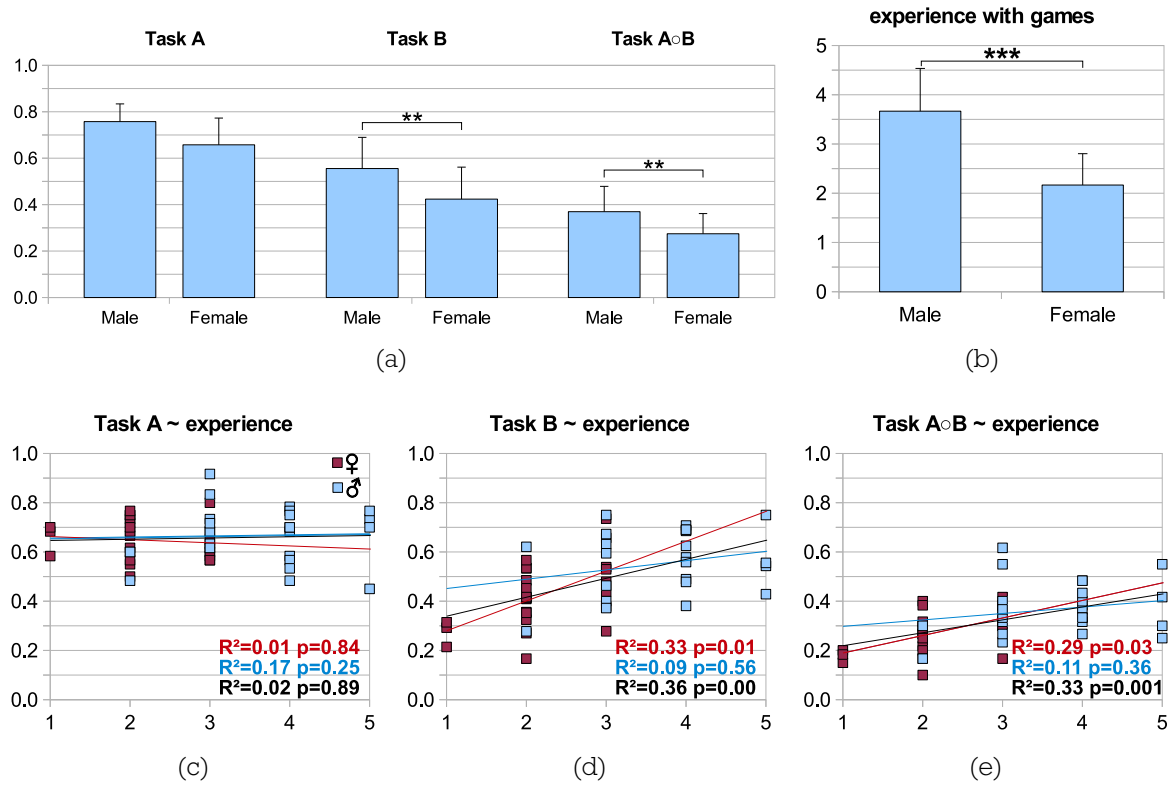


Figure 6.10: (a) gender separated task performance, (b) candidates' experience in computer games significance is denoted by "**" ($p < 0.01$) or "***" ($p < 0.001$), (c,d,e) male (blue), female (red) and combined (black) correlation between task performance and computer game experience

Chapter 7

Conclusion and Outlook

7.1 Conclusion

In this thesis, we provided an overview of the human senses responsible for visual and auditory perception highlighting the bimodal merging of our senses. After explaining visual rendering and stereo sound synthesis, the process of high-quality spatialized audio rendering starting with proper HTRF acquisition to the output via headphones or multi speaker setup is described and the existing methods are compared. An application simulating an urban environment for the use in a virtual gap crossing scenario was implemented. It was later used in a study with 48 participants to highlight the performance increase through binaural sound rendering.

Making the decision to cross the road is a highly complex day-to-day task which requires efficient perceptual and cognitive processes. Therefore we picked up this task and continuously enhanced the concept to elaborate the sensitivity of our study to the effects of auditory cues. For this purpose we had to provoke a situation where visual cues alone do not carry enough information, so that participants have to act on intuition about the situation rather than on confidence. Pilot studies enabled us to eliminate thought and reasoning in the decision making process to reduce variance due to strong effects of evolving individual crossing strategies. That way, we were able to study a scenario where the impact of auditory cues becomes evident. Results showed a significance far above chance that spatialized audio rendering with HRTF filtering provides sufficient perceptual information to increase performance, while no significant effect for conventional stereo rendering could be observed. As we were interested in identifying spatialized auditory information as a positive factor for accuracy and reliability,

which has been accomplished, the gap crossing experiment was designed to mimic a risky but still occurring traffic situation.

Our perception is a multimodal process were various stimuli collected by different sense organs lead to sometimes redundant but augmented sensations. This percept facilitates cognition as the process of forming a mental image of the situation in the brain and helps to come to a good, and reliable decision following the MLE model. Situations which carry a lot of information, such as traffic scenarios, can have a significant visual uncertainty, which can be reduced by adding auditory spatial information. Depending on the application-specific task, audio quality is an important issue in the design of VE applications which cannot be compensated by high-quality visual stimuli. This impact on accuracy by auditory information, although known for quite some time, is often underestimated. [SO90] showed that the TTC judgments of blind people are almost as accurate as those of sighted ones.

In our study, we maintained a realistic alignment between the virtual scene's audio scale and the visual display. The parameters were fed into the application and recalibrated for each participant, but maybe a perfect adjustment between them is not the crucial point for our results regarding task-performance. This would concur with [EB04], who describe perception as a combination and integration of various sensory stimuli, collected by our sense organs. These often redundant signals have different modalities and come in different units. To determine a sound's position, the brain must learn and calibrate these cues, using accurate spatial feedback from other sensorimotor systems [HVOVR99]. The recalibration of the human auditory system based on multimodal sensory feedback (e.g. exploiting semantic congruency) can compensate for spatial deviations caused by bad auditory VE configuration (e.g. slightly improper HRTF). The other way around bimodal task facilitation could also be expected for small display setups as long as the user can adapt to the auditory scene of a different scale.

Spatialized auditory information together with visual-immersive VEs can emphasize our bimodal perception. Furthermore VEs help to teach safe situation-handling in a risky and often unsafe environment. Increasing traffic volume and migration into cities demands for a technology assisted urban development, and specialized pedestrian training simulators to protect especially the youngest road users from danger.

7.2 Outlook

As this thesis tries to synergize computer graphics with applied perception, we targeted an application-oriented human computer interaction issue. Of course there are things left for future research of the gap crossing scenario in the field of audio-visual perception in conjunction with VEs:

Eye tracking An eye tracking device could give clear evidence about the various regions of interest on the screen to investigate our assumption about the crossmodal aural monitoring of the current car passing by while the visual focus resides on the next car.

Improve visual stimulus Introduce stereoscopic display techniques like e.g. polarized 3D glasses as a new condition for the experiment and measure the performance in different combinations of HQ/LQ/AUDIO/VIDEO.

Semantic incongruity Explore the effect of semantical coincidence between visual and auditory representation of the vehicle by replacing the engine audio with an analytical tone or a completely different sound and measure performance in the conditions.

In-/Decrease immersion Typically, a greater field of view results in a greater sense of immersion and better situational awareness. The usage of a multi-screen setup facilitates a broader FOV near the optimal 180° human field of view.
Contrary to that, the visual scene could be ported to a very small screen while the sound scene remains unchanged. Performance measures could show a balancing effect of LQ visual representation by HQ auditory cues.

HRTF generation/selection Most of the established approaches suffer certain disadvantages. As the HRTF is the key to realistic spatialized audio rendering via headphones, new algorithms could provide a solution to the time and money consuming process of individual HRTF measurement. To minimize the need for specially designed hardware, a perception-based approach operating on a user-application-feedback system combining psychological, physiological and computational neuroscience concepts could be employed.

Appendix A

Acronyms

AI artificial intelligence

ANOVA analysis of variance

API Application Programming Interface

AR Augmented Reality

CD crossing duration

Cg C for graphics

FPS frames per second

GLSL OpenGL Shading Language

GPU graphics processing unit

HCI Human-Computer Interaction

HLSL High Level Shading Language

HMD head-mounted display

HQ high-quality

HRIR Head-Related Impulse Response Function

HRTF Head-Related Transfer Function

ILD interaural level difference

ITD interaural time difference

IVD inter-vehicle distance

LQ low-quality

MRI magnetic resonance imaging

OGRE Object-Oriented Graphics Rendering Engine

OIS Object-Oriented Input System

PC Personal Computer

PCA principal components analysis

SCW safe crossing window

TTC time to contact

VE Virtual Environment

VR Virtual Reality

Appendix B

Supplements

B.1 Termini

Time to Contact (TTC)

$TTC = distanceToObserver/drivingSpeed$

A general case for the car starting at a predefined distance and traveling at a certain speed is difficult to construct. The distance of the car's starting position away from the observer's position and different driving speed demands the definition of distance in seconds rather than in meters as it would not make sense to start a car traveling with 40 km/h at the same position as a car traveling with 80 km/h.
This time is referred to as the time to contact (TTC). It describes how long it takes the car to reach the observer's position.

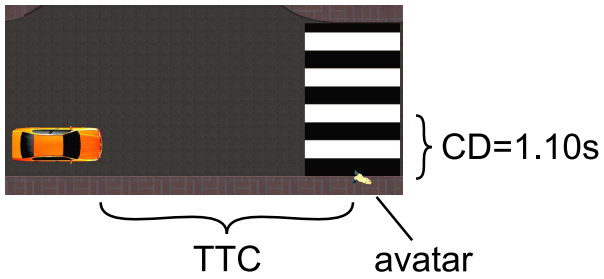


Figure B.1: Top view with TTC and CD; crossing duration is based on 7 km/h walking speed

Crossing Duration (CD)

$$CD = distanceToSafety/walkingSpeed$$

The time needed to cross the street is called crossing duration (CD). It describes how long it takes the observer to reach the safe position where the far outer bound of the car cannot hit him anymore.

Safety Ratio

$$safetyratio = CD/TTC$$

The safety ratio is calculated by dividing the time it takes the pedestrian to cross the street through the time to contact of the incoming car. The result indicates:

$$accident = safetyratio < 1.0$$

$$safecrossing = safetyratio \geq 1.0$$

This means a safe gap becomes unsafe after some time elapsed, in other words the TTC becomes smaller and so does the safety ratio. Figure B.1 explains the connection of CD and TTC, the car is driving towards the observer's position and will arrive in a certain TTC. The observer can cross the street with a certain CD.

Safe Crossing Window (SCW)

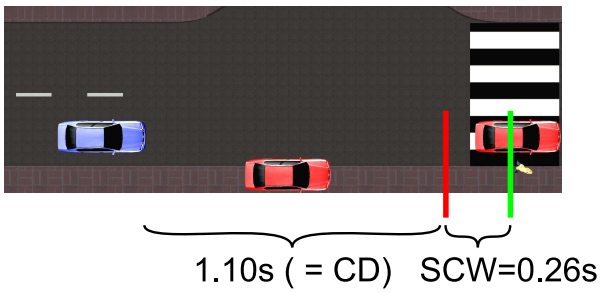


Figure B.2: Top view with CD and SCW; crossing duration is based on 7 km/h walking speed; safe crossing window starts at green line and ends at red line

The safe crossing window (SCW) denotes the span from the moment a car is passable to the point in time where a safe gap becomes unsafe meaning:

$$safetyratio < 1.0$$

The walking procedure has to be triggered within the SCW, otherwise the pedestrian will run into the close vehicle next to him or will be overrun by the next incoming car.

B.2 Bonferroni-Holm

As we stated three hypotheses (see Chapter 6.4), using a t-test to test each hypothesis against each other leads to two problems:

1. Inconsistencies: One wants to compare the expected values μ_1, μ_2, μ_3. Pairwise testing of $\mu_1 = \mu_2, \mu_1 = \mu_3, \mu_2 = \mu_3$ does not reject the null-hypotheses but only the hypothesis $\mu_1 = \mu_2 = \mu_3$
2. α-error inflation: The problem of multiple comparison discriminates between local α (single hypothesis) and global α (whole set of hypotheses). In case of independent tests, local α can be adjusted according to $\alpha_{local} = 1 - (1 - \alpha_{global})^{(1/k)}$

We address this problem by using the conservative Bonferroni-Holm method:

1. define α_{global}
2. perform pairwise tests to calculate p-values
3. sort p-values ascending
4. calculate each α_{local} according to: $i = 1, ...k$
   $\alpha_1 = \frac{\alpha_{global}}{k}$
   $\alpha_2 = \frac{\alpha_{global}}{k-1}$
   $\alpha_i = \frac{\alpha_{global}}{k-i+1}$
5. compare p-values to calculated, sorted local α (starting with α_1) and repeat this step until $p_i > \alpha_i$
6. reject all null-hypotheses where $p_i < \alpha_i$
   if $p_i > \alpha_i$ then stop and accept all null-hypotheses that have not been rejected at previous steps

Appendix C

Bibliography

[AB04] David Alais and David Burr. The ventriloquist effect results from near-optimal bimodal integration. *Current Biology*, 14(3):257–262, February 2004.

[Acc] Statistik austria. Straßenverkehrsunfälle.

[ADS03] *Sixty Years of Aeronautical Research in Australia: Research Overview*. Australian Defence Science and Technology Organisation, March 2003. http://www.dsto.defence.gov.au/.

[Ale04] Petter Alexanderson. Peripheral awareness and smooth notification: the use of natural sounds in process control work. In *NordiCHI '04: Proceedings of the third Nordic conference on Human-computer interaction*, pages 281–284, New York, NY, USA, 2004. ACM.

[AMB06] David Alais, Concetta Morrone, and David Burr. Separate attentional resources for vision and audition. *Proceedings of the Royal Society B: Biological Sciences*, 273(1592):1339–1345, 2006.

[AMHH08] Tomas Akenine-Möller, Eric Haines, and Natty Hoffman. *Real-Time Rendering 3rd Edition*. A. K. Peters, Ltd., Natick, MA, USA, 2008.

[Beg00] Durand R. Begault. *3-D sound for virtual reality and multimedia*. Academic Press Professional, Inc., San Diego, CA, USA, 2000.

[BJH07] Armando B. Barreto, Julie A. Jacko, and Peter J. Hugh. Impact of spatial auditory feedback on the efficiency of iconic human-computer interfaces under conditions of visual impairment. *Computers in Human Behavior*, 23(3):1211–1231, 2007.

[BKSZ01] Carsten Burstedde, Kai Klauck, Aandreas Schadschneider, and Johannes Zittartz. Simulation of pedestrian dynamics using a two-dimensional cellular automaton. *Physica A: Statistical Mechanics and its Applications*, 295(3-4):507–525, June 2001.

[BKWJ06] Orit Bart, Noomi Katz, Patrice L. Weiss, and Naomi Josman. Street crossing by typically developed children in real and virtual environments. *2006 International Workshop onVirtual Rehabilitation*, pages 42–46, 2006.

[Bla99] Jens Blauert. *Spatial hearing – The psychophysics of human sound localization*. The MIT Press, 1999.

[BSM07] Benjamin K. Barton, David C. Schwebel, and Barbara A. Morrongiello. Increasing children's safe pedestrian behaviors through simple skills training. *Journal of Pediatric Psychology*, 32(4):475–480, 2007.

[CA05] Jean-René Carré and Julien Arantxa. *A new method for analysing the pedestrian activity during the daily urban mobility and for measuring the pedestrian risk exposure*. nstitut national de recherche sur les transports et leur sécurité, 2005.

[CCPI98] Marie L. Connelly, Helen M. Conaglen, Barry S. Parsonson, and Robert B. Isler. Child pedestrian's crossing gap thresholds. *Accident Analysis and Prevention*, page 443–453, 1998.

[CLDV09] Viola Cavallo, Régis Lobjois, Aurélie Dommes, and Fabrice Vienne. Elderly pedestrians' visual timing strategies in a simulated street-crossing situation. In *PROCEEDINGS of the Fifth International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design*, 2009.

[CRO06] Tamara A. Clancy, Julia J. Rucklidge, and Dean H. Owen. Road-crossing safety: A comparison of adolescents with and without adhd. *Journal of Clinical Child and Adolescent Psychology*, pages 203–215, 2006.

[Del00] William Dell. The use of 3d audio to improve auditory cues in aircraft. Technical report, Department of Computing Science, University of Glasgow, 2000.

[DHJA01] Niranjan Debnath, Zul Azizi Hailani, Sakinah Jamaludin, and Syed A. K. Aljunis. An electronically guided walking stick for the blind. In *Engi-*

neering in Medicine and Biology Society, 2001. Proceedings of the 23rd Annual International Conference of the IEEE, pages 1377–1379, 2001.

[DN66] Ole-Johan Dahl and Kristen Nygaard. Simula: an algol-based simulation language. *Commun. ACM*, 9(9):671–678, 1966.

[DPT+08] Matteo Dellepiane, Nico Pietroni, Nicolas Tsingos, Manuel Asselot, and Roberto Scopigno. Reconstructing head models from photographs for individualized 3d-audio processing. *Computer Graphics Forum (Special Issue - Pacific Graphics 2008 Proc.)*, 27(7):1719–1727, 2008.

[DSP+99] Elizabeth T. Davis, Kevin Scott, Jarrell Pair, Larry F. Hodges, and James Oliverio. Can audio enhance visual perception and performance in a virtual environment? In *Proceedings of the 43rd Annual Meeting of the Human Factors and Ergonomics Society*, pages 1197–1201, 1999.

[EB] Britannica.com, encyclopædia britannica. http://www.britannica.com/.

[EB04] Marc O. Ernst and Heinrich H. Bülthoff. Merging the senses into a robust percept. *Trends in Cognitive Science*, 8(4):162–169, April 2004.

[Fru71] John J. Fruin. *Pedestrian Planning and Design*. Alabama: Elevator World Inc, New York, New York, 1971.

[Gie92] Hans W. Gierlich. The application of binaural technology. *Applied Acoustics*, 36:219–243, 1992.

[Gus00] Rainer Guski. *Wahrnehmung*. Kohlhammer, Stuttgart, 2000.

[HB95] Claudia Hendrix and Woodrow Barfield. Presence in virtual environments as a function of visual and auditory cues. In *VRAIS '95: Proceedings of the Virtual Reality Annual International Symposium (VRAIS'95)*, page 74, Washington, DC, USA, 1995. IEEE Computer Society.

[HLJ+06] Hu Hongmei, Zhou Lin, Zhang Jie, Ma Hao, and Wu Zhenyang. Computer simulation of hrtfs for personalization of 3d audio. *International Conference on Computational Intelligence and Security*, pages 1829–1832, 2006.

[HVOVR99] Paul M. Hofman, A. John Van Opstal, and Jos G. A. Van Riswick. Relearning sound localization with new ears. *Journal of the Acoustical Society of America*, 105(2):1035, 1999.

[INR] Institut national de recherche en informatique et automatique. http://www.inria.fr/.

[KP85] Andrew J. King and Alan R. Palmer. Integration of visual and auditory information in bimodal neurones in the guinea-pig superior colliculus. *Exp. Brain Research*, 60:492–500, 1985.

[Len08] Tobias Lentz. *Binaural technology for virtual reality*. PhD thesis, RWTH Aachen, 2008.

[Lis] Listen hrtf database. http://recherche.ircam.fr/equipes/salles/listen/.

[LVK02] Pontus Larson, Daniel Vastfall, and Mendel Kleiner. Better presence and performance in virtual environments by improved binaural sound rendering. In *Proceedings of the 22nd International Conference: Virtual, Synthetic, and Entertainment Audio*, pages 31–38, Espoo,Finland, June 2002. AES.

[LYM84] David N. Lee, David S. Young, and Carmel M. McLaughlin. A roadside simulation of road crossing for children. *Ergonomics*, 12:1271–1281, 1984.

[MBT+07] Thomas Moeck, Nicolas Bonneel, Nicolas Tsingos, George Drettakis, Isabelle Viaud-Delmon, and David Alloza. Progressive perceptual audio rendering of complex scenes. In *I3D '07: Proceedings of the 2007 symposium on Interactive 3D graphics and games*, pages 189–196, New York, NY, USA, 2007. ACM.

[MD01] Pierre-Emmanuel Michon and Michel Denis. When and why are visual landmarks used in giving directions? In *COSIT 2001: Proceedings of the International Conference on Spatial Information Theory*, pages 292–305, London, UK, 2001. Springer-Verlag.

[MDCT05] Georgia Mastoropoulou, Kurt Debattista, Alan Chalmers, and Tom Troscianko. The influence of sound effects on the perceived smoothness of rendered animations. In *APGV '05: Proceedings of the 2nd symposium on Applied perception in graphics and visualization*, pages 9–15, New York, NY, USA, 2005. ACM.

[MMP02] Joan McComas, Morag MacKay, and Jayne Pivik. Effectiveness of virtual reality for teaching pedestrian safety. *CyberPsychology & Behavior*, 5(3):185–190, 2002.

[MTNK08] Parham Mokhtari, Hironori Takemoto, Ryouichi Nishimura, and Hiroaki Kato. Computer simulation of hrtfs for personalization of 3d audio. *Second International Symposium on Universal Communication*, pages 435–440, 2008.

[MWRZ05] Georg F. Meyer, Sophie M. Wuerger, Florian Röhrbein, and Christoph Zetzsche. Low-level integration of auditory and visual motion signals requires spatial co-localisation. *Exp Brain Res*, 166(3-4):538–547, 2005.

[NKW00] Yuval Naveh, Noomi Katz, and Patrice Weiss. The effect of interactive virtual environment training on independent safe street crossing of right cva patients with unilateral spatial neglect. *Proceedings of the Third ICDVRAT*, pages 243–248, 2000.

[OCW+08] Jennifer Oxley, Melinda Congiu, Michelle Whelan, Angelo D'Elio, and Judith Charlton. Teaching young children to cross roads safely. *Annual proceedings / Association for the Advancement of Automotive Medicine. Association for the Advancement of Automotive Medicine*, 52:215–23, 2008.

[Ogr] Ogre 3d: Object oriented graphics rendering engine. http://www.ogre3d.org/.

[OIF+05] Jennifer Oxley, Elfriede Ihsen, Brian Fildes, Judith Charlton, and Ross Day. Crossing roads safely : An experimental study of age differences in gap selection by pedestrians. *Accident analysis and prevention*, pages 962–971, 2005.

[Ols67] Harry F. Olson. *Music, Physics and Engineering*. Dover Publications, Mineola, NY, 1967.

[O'S91] Robert P. O'Shea. Thumb's rule tested: visual angle of thumb's width is about 2 deg. *Perception*, 20(3):415–418, 1991.

[PKC07] Jodie M. Plumert, Joseph K. Kearney, and James F. Cremer. Children's road crossing: A window into perceptual-motor development. *Current Directions in Psychological Science*, 16:255–258, 2007.

[RVSP09] Bernhard E. Riecke, Aleksander Väljamäe, and Jörg Schulte-Pelkum. Moving sounds enhance the visually-induced self-motion illusion (circular vection) in virtual reality. *ACM Trans. Appl. Percept.*, 6(2):1–27, 2009.

[SAB07] Elizabeth A. Seward, Daniel H. Ashmead, and Bobby Bodenheimer. Using virtual environments to assess time-to-contact judgments from pedestrian viewpoints. *ACM Trans. Appl. Percept.*, 4(3):18, 2007.

[Sch57] Herbert Schober. *Das Sehen*, volume 1. Fachbuchverlag Leipzig, Leipzig, 1957.

[SD79] William Schiff and Mary L. Detwiler. Information used in judging impending collision. *Perception*, 8:647–658, 1979.

[SDTB08] Jaka Sodnik, Christina Dicke, Sašo Tomaič, and Mark Billinghurst. A user study of auditory versus visual interfaces for use while driving. *Int. J. Hum.-Comput. Stud.*, 66(5):318–332, 2008.

[SJR03] Gordon Simpson, Lucy Johnston, and Michael Richardson. An investigation of road crossing in a virtual environtment. *Accident Analysis & Prevention*, 35(5):318–332, 2003.

[SM93] Barry E. Stein and M. Alex Meredith. *The Merging of Senses*. MIT Press, Cambridge, MA, 1993.

[SO90] William Schiff and Rivka Oldak. Accuracy of judging time to arrival: Effects of modality, trajectory, and gender. *Journal of Experimental Psychology: Human Perception and Performance*, 16:303–316, 1990.

[ST74] Edgar A. Shaw and R. Teranishi. Transformation of sound-pressure level from the free field to the eardrum in horizontal plane. *Journal of the Acoustical Society of America*, 56:1848–1861, 1974.

[ST96] Richard A. Schieber and Nancy J. Thompson. Developmental risk factors for childhood pedestrian injuries. *Injury Prevention*, 2(3):228–236, 1996.

[STG+06] Jaka Sodnik, Saso Tomazic, Raphael Grasset, Andreas Duenser, and Mark Billinghurst. Spatial sound localization in an augmented reality environment. In *OZCHI '06: Proceedings of the 18th Australia conference on Computer-Human Interaction*, pages 111–118, New York, NY, USA, 2006. ACM.

[SWC+03] Venkataraman Sundareswaran, Kenneth Wang, Steven Chen, Reinhold Behringer, Joshua McGee, Clement Tam, and Pavel Zahorik. 3d audio augmented reality: Implementation and experiments. In *ISMAR '03: Proceedings of the 2nd IEEE/ACM International Symposium on Mixed and*

Augmented Reality, page 296, Washington, DC, USA, 2003. IEEE Computer Society.

[SZ00] Russell L. Storms and Michael J. Zyda. Interactions in perceived quality of auditory-visual displays. *Presence: Teleoper. Virtual Environ.*, 9(6):557–580, 2000.

[TD04] Ariane Tom and Michel Denis. Language and spatial cognition: comparing the roles of landmarks and street names in route instructions. In *Applied Cognitive Psychology*, volume 18, pages 1213–1230, 2004.

[TKK04] Shiose Takayuki, Ito Kiyohide, and Mamada Kazuhiko. The development of virtual 3d acoustic environment for training 'perception of crossability'. In Joachim Klaus, Klaus Miesenberger, Wolfgang L. Zagler, and Dominique Burger, editors, *ICCHP*, volume 3118 of *Lecture Notes in Computer Science*, pages 476–483. Springer, 2004.

[TLO08] Ying-Chan Tung, Yung-Ching Liu, and Yang-Kun Ou. The pedestrian road-crossing behaviors between older and younger age groups. Proceedings of the 9th Asia Pacific Industrial Engineering & Management Systems Conference, APIEMS 2008, 2008.

[TVVdKBS05] Arenda F. Te Velde, John Van der Kamp, José A. Barela, and Geert J. P. Savelsbergh. Visual timing and adaptive behavior in a road-crossing simulation study. *Accident Analysis and Prevention*, 37:399–406, 2005.

[Wii] Wiiyourself library. http://wiiyourself.gl.tter.org/.

[Wik] Wikipedia the free encyclopedia. http://www.wikipedia.org/.

[WR04] Baohong Wan and Nagui M. Rouphail. Using arena for simulation of pedestrian crossing in roundabout areas. *Transportation Research Record: Journal of the Transportation Research Board*, pages 58–65, 2004.

[XL03] Tian Xiao and Qing Huo Liu. Finite difference computation of head-related transfer function for human hearing. *The Journal of the Acoustical Society of America*, 113(5):2434–2441, 2003.

[Xmd] Cross-modal perceptual interaction and rendering. http://www-sop.inria.fr/reves/CrossmodPublic/index.php.

[YDW+06] Jianguo Yang, Wen Deng, Jinmei Wang, Qingfeng Li, and Zhaoan Wang. Modeling pedestrians' road crossing behavior in traffic system micro-

simulation in china. *Transportation Research Part A: Policy and Practice*, 40(3):280–290, March 2006.

[ZF03] Ying Zhang and Terrence Fernando. 3d sound feedback act as task aid in a virtual assembly environment. In *TPCG '03: Proceedings of the Theory and Practice of Computer Graphics 2003*, page 209, Washington, DC, USA, 2003. IEEE Computer Society.

[ZFXT06] Ying Zhang, Terrence Fernando, Hannan Xiao, and Adrian R. L. Travis. Evaluation of auditory and visual feedback on task performance in a virtual assembly environment. *Presence: Teleoper. Virtual Environ.*, 15(6):613–626, 2006.

[ZHDD03] Dmitry N. Zotkin, Jane Hwang, Ramani Duraiswaimi, and Larry S. Davis. Hrtf personalization using anthropometric measurements. *Applications of Signal Processing to Audio and Acoustics*, pages 157–160, 2003.