Bimodal Task-Facilitation in a Virtual Traffic Scenario through Spatialized Sound Rendering

Matthias Bernhard[1]
Technical University of Vienna
and
Karl Grosse[2]
Technical University of Vienna
and
Michael Wimmer[3]
Technical University of Vienna

Audio rendering is generally used to increase the realism of Virtual Environments (VE). In addition, audio rendering may also improve the performance in specific tasks carried out in interactive applications such as games or simulators.

In this paper we investigate the effect of the quality of sound rendering on task performance in a task which is inherently vision dominated. The task is a virtual traffic gap crossing scenario with two elements: first, to discriminate crossable and uncrossable gaps in oncoming traffic, and second, to find the right timing to start crossing the street without an accident.

A study was carried out with 48 participants in an immersive Virtual Environment setup with a large screen and headphones. Participants were grouped into three different conditions. In the first condition, spatialized audio rendering with head-related transfer function (HRTF) filtering was used. The second group was tested with conventional stereo rendering, and the remaining group ran the experiment in a mute condition.

Our results give a clear evidence that spatialized audio improves task performance compared to the unimodal mute condition. Since all task-relevant information was in the participants' field-of-view, we conclude that an enhancement of task performance results from a bimodal advantage due to the integration of visual and auditory spatial cues.

Categories and Subject Descriptors: I.3.7 [**Computer Graphics**]: Three-Dimensional Graphics and Realism—*Virtual reality*; K.8.0 [**Personal Computing**]: General—*Games*; H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems—*Artificial, augmented, and virtual realities*

General Terms: Experimentation,Human Factors

Additional Key Words and Phrases: audio-visual perception, bimodal, spatialized audio rendering, task facilitation, pedestrian simulator, pedestrian safety, virtual environments, human-computer interaction

1. INTRODUCTION

Interactive Virtual Environments (VE) have become a part of many people's everyday experience. Thanks to recent progress in computer graphics, applications like video games, simulators, virtual worlds on the web, and virtual reality installations convey a high degree of realism and presence to their users. However, while improvements in the fidelity of VEs

increase immersion and presence, they do not necessarily increase the performance and efficiency with which the user carries out tasks [Bormann 2006]. Improving user cognition and interaction requires a perceptual evaluation to identify the most important cues which aid users in their interaction with an application. A rendering technology, e.g. stereo audio rendering, might contribute to a high degree to the fidelity of a VE, but at the same time fail to provide task-critical perceptual cues that support the user in application-specific tasks.

In this paper we investigate this problem in more detail by studying task performance in a specific challenging setting, namely a gap-crossing task in a traffic simulator. Traffic scenarios are perceptually challenging to road users, as users need to divide audio-visual attention to many moving objects, estimate complex spatio-temporal relations, be very attentive, and react quickly. The task carried out in our study involves complex cognitive processes to discriminate dynamic spatial relations, and requires high concentration, optimal action timing, and fast reaction. In this study, we focus especially on the role of audio rendering and its interaction with vision. Compared to most previous work on *auditory task facilitation*, we address a rather non-trivial case, since no task-relevant information is given by the audio modality alone, i.e., all information has to be taken from both modalities together. In the task studied, all sound is arising naturally (i.e., no feedback on a semantic level) and all task-relevant events occur in the visual field-of-view (FOV).

We hypothesize that thanks to the use of binaural filtering with Head Related Transfer Functions (HRTF) to spatialize sound signals, natural audio cues can be simulated in a sufficient quality so that a *bimodal task-facilitation* can be observed in a VE. Bimodal task-facilitation is assumed, for instance, due to a statistical advantage when redundant and uncertain information is perceived in two modalities and integrated.

The results obtained show a clear evidence that high-quality spatialization of auditory stimuli can improve task performance in a virtual audio-visual environment, notably action timing and estimation of spatial relations, whereas for conventional stereo rendering, no positive effect could be observed. This result is relevant for the design of interactive VE applications where user task performance depends on the way the VE is presented to the user. Example applications include high-quality simulators to train pilots and motor sportsmen, or simulators for pedestrian traffic safety training. Apart from VEs, the results might also find relevance for the design of real vehicles or aircraft, or helmets, where the audio feedback pilots or vehicle drivers perceive is heavily distorted, due to masking noise (e.g. engine), distraction through sonar occluders (e.g. shell of a helmet) or noise protectors (e.g. cockpit walls). In such cases, the audio scene could be reconstructed with less distortions in a way that improves task performance.

2. BACKGROUND

We first discuss the general crossmodal effects in multimodal perception that are so important for our experiment. Then we focus on previous work studying the use of spatialized audio to improve task performance, and finally address experiments and VE applications concerned with pedestrian safety.

2.1 Bimodal perception and task facilitation

When auditory perception provides information which is not available in the visual channel, like e.g. sounding events emerging behind us or auditory feedback with symbolic information (e.g. spoken instructions), auditory task facilitation has rather simple explanations. When task facilitation is observed due to added information in a second modality (e.g. audio) even though a single modality (e.g. vision) provides all task-relevant information, as is the case in our experiment, we assume a *bimodal task facilitation*, since redundant information is perceived in two modalities and effectively combined or integrated. As perception can not be studied independently for one modality, the current trend in Neuroscience and cognitive science is towards a multimodal understanding of perception, accounting for intermodal interactions of senses (e.g.[Stein and Meredith 1993]). To get a brief understanding how different senses are combined or integrated within or across modalities to resolve ambiguities or increase perceptional estimation precision, we recommend an enlightening review in [Ernst and Bülthoff 2004].

Through multimodal integration of corresponding stimuli, unimodal ambiguities are resolved faster and more precisely, thus enhancing recognition (e.g. [Suied et al. 2009]), localisation (e.g. [Alais and Burr 2004]) or orientation behaviors (e.g. [Jiang et al. 2002]). Task facilitation mainly results from improved *efficiency* and *accuracy* in the interpretation of audio-visual percepts.

Efficiency (i.e., reaction time) is improved due to the so-called redundant signal effect [Kinchla 1974]. A race model accounting for the statistical advantage when redundant information from two modalities is processed in parallel can explain faster reactions [Raab 1962]. But even shorter reaction times than were predicted by this model were observed and supposed to result from a coactivation [Miller 1982], which may be due to the existence of multimodal cells which are coactivated by different modalities [King and Palmer 1985].

Accuracy (i.e., discrimination performance) is improved through an optimal integration of uncertain information from two or more modalities. For instance, studies with visual-haptic size estimation tasks [Ernst and Banks 2002] and audio-visual localisation tasks [Alais and Burr 2004] both revealed that estimation precision approaches an optimum predicted by the Maximum-Likelihood Estimator (MLE).

In our experiment, participants need to combine accuracy (e.g. correctly discriminate crossable from uncrossable gaps) and efficiency (e.g. quickly decide when to cross a gap). Thus, we study a complex mixture of several perceptual factors instead of only one isolated effect as in many previous studies.

Apart from direct implications on task performance, bimodal perception also affects a user's impression of a VE. Due to the fused perception of congruent audio-visual stimuli, the perceived display quality of one modality can be crossmodally biased through the other modality. For instance, sound can perceptually increase the fidelity of the visual display [Davis et al. 1999; Storms and Zyda 2000], smoothen animations [Mastoropoulou et al. 2005], or lower visual quality discrimination when collision sounds enrich material properties [Bonneel et al. 2010].

2.2 Spatialized sound rendering to improve task performance

Using the audio channel to provide task-critical feedback can significantly improve user task performance. While visual feedback (e.g. menu icons) may distract attentional resources required to monitor task-relevant objects visually, auditory feedback can be perceived and processed to a certain degree in parallel (e.g. [Treismann and Davies 1973; Alais et al. 2006]). In particular, spatialized sound rendering has found great application in auditory displays, for example in cars [Sodnik et al. 2008] using spatially separated feedback sounds, for auditory menus for computer users with visual impairments [Barreto et al. 2007], and several Augmented Reality applications [Sodnik et al. 2006; Sundareswaran et al. 2003]. Spatialized sound was also used to resynthesize the external audio scene of aircrafts [Dell 2000], mostly in order to improve combat performance through audio cues to detect the direction of threats and targets.

In contrast to the above-mentioned examples, the application used in our experiment does not augment a user in her task by symbolic and artificial sounds (e.g. warning signals). Instead, we use sound to enhance task performance by following the laws of physics to provide VEs with a realistic context. Naturalistic cues let a user perceive additional properties of the environment and help, for instance, to localize and identify objects. [Bormann 2005] observed that distance attenuation is the most important cue to find sounding objects (e.g. playing radio) in a VE, while directional spatialization of sound, though increasing presence, was not proven to be useful for this task. On the other hand, for bimodal localization in a static scene, [Nguyen et al. 2009] reported that performance levels of real-world conditions are approached thanks to high-quality spatialization of sounds by HRTF filtering. Our work will contribute a more complex case of study (i.e., dynamic environment) to show that a realistic audio-visual VE can enhance task performance due to the bimodal task facilitation.

As high-quality audio rendering significantly increases the perceived fidelity and presence of a VE [Hendrix and Barfield 1995; Riecke et al. 2009], task facilitation may also be due to the rather simple explanation that participants act more seriously in a VE of high fidelity. However, [Bormann 2005] found that participants score the degree of presence independent of the audio cues that are relevant for task performance, and they found no correlation between presence and objective performance measures, like reaction time or task time [Bormann 2006].

2.3 Pedestrian safety

Studying pedestrians was approached from the perspective of different disciplines such as psychology [Michon and Denis 2001; Tom and Denis 2004], accidentology (e.g. [Carré and Arantxa 2005]), transport planning (e.g. [Fruin 1971]) or flow simulation (e.g. [Yang et al. 2006; Burstedde et al. 2001; Wan and Rouphail 2004]). Pedestrian road-crossing behavior has often been of particular interest. Established features for analysis include gap acceptance, crossing duration, time-to-contact estimation or waiting time. Many studies focussed on age-related differences [Connelly et al. 1998; Tung et al. 2008], often by means of comparing children versus adults [Te Velde et al. 2005; Plumert et al. 2007]. Studies in real-world setups are expensive and difficult. For instance, [Connelly et al. 1998] used a laser detector to measure car speed and distance. Hence, many attempts

were made to carry out studies with simulations in the lab where influential variables can be measured or controlled more easily. For instance, [Te Velde et al. 2005] simulated traffic environments with artificial indoor streets, or [Sidaway et al. 1990] and [Tung et al. 2008] used pre-recorded road environment videos in their experiments. However, VEs and Virtual Reality technology have emerged as the preferred tool to simulate interactive traffic environments (e.g. [Simpson et al. 2003; Seward et al. 2007]). Since safety-training can effectively improve pedestrian traffic safety [Barton et al. 2007], but reasonable real-world environments for traffic safety training are expensive and difficult to realize, computer generated VEs also found great reception for safety-training simulators [Oxley et al. 2008; Cavallo et al. 2009; Schwebel and McClure 2010]. However, research on Virtual Reality applications for virtual safety training is still in its infancy and many studies [Naveh et al. 2000; McComas et al. 2002; Bart et al. 2008; Schwebel et al. 2008] are mainly concerned with the validation of applicability and effectiveness of these tools.

While most work is focussed on the behavior of threatened groups (e.g. impaired people, seniors or children), on implications for safety design (e.g. pedestrian planning), on psychological factors, in particular attention, or on educational aspects such as safety training methods, we found only few related references addressing the unquestionably important role of auditory perception in traffic safety. For instance, [Barnecutt and Pfeffer 1995] investigated auditory distance perception in traffic. Apart from level differences they suggest that also non-auditory criteria, like past experience or visual imagery, are important factors for auditory interpretation. An interesting issue was considered in [Caelli and Porter 1980], where they observed different localization performance for different types of (emergency car) siren sounds. As auditory cues are of major importance for people with severe visual impairments, attempts have been made to provide maximum realism using HRTF filtering and reverberation cues in an auditory traffic simulator intended to be used as a safety training application for blind people [Takayuki et al. 2004].

With the virtual gap-crossing experiment in our work, we present a representative example case supporting our hypothesis that auditory perception is relevant for traffic safety and that it can be effectively utilized for better and faster decisions when moving through traffic, even though the pedestrian has no visual impairments and is able to obtain all safety relevant information from his environment visually.

3. SCOPE OF THIS WORK

It is obvious that auditory attention can increase overall safety in traffic when important or dangerous events are well audible, but can only be poorly recognized visually. Less obvious, however, is the impact of auditory cues in a situation where visual attention is already focussed on those objects (e.g. approaching cars) that are most relevant for the current action (e.g. crossing the street). It is clear that vision will dominate in such situations, since spatial information can be perceived more accurately in the visual channel than through auditory perception, but can auditory cues still improve task performance?

In this paper we describe a gap-crossing experiment designed to explore this less obvious latter case. In gap crossing, a user observes a stream of oncoming traffic and needs to decide when to cross the street. This includes two main perceptual tasks: determine a gap

which is crossable, and find the correct timing to actually cross the street. We focus on two basic questions:

—***Is bimodal task facilitation observable in a virtual traffic simulator?***

—***Do simulations with state-of-the-art audio-rendering techniques provide sufficient realism to reveal bimodal advantages?***

To make effects of bimodal perception measurable in task performance, we need to study a situation where the perceptual *and* cognitive limits of participants are approached. This is possible thanks to the use of a VE, as in real-world experiments it is not possible to count accidents. The variety of perceptual and cognitive activities involved in the gap-crossing task make it difficult to reveal the influence of auditory cues, since they are relatively subtle compared to the effects of individual subject-related factors. The key to obtain useful results was to configure the experiment so that the uncertainty of visual information is increased (but without hiding anything) and that participants do not develop individual strategies to accomplish the task. This was accomplished by increasing the difficulty of the task to a limit where participants need to utilize their full perceptual capabilities. We used gap sizes which can not be easily distinguished visually, thus introducing ambiguities that required participants to make spontaneous decisions not based on a cognitive process. This reduced individual strategies in participants and revealed that decisions when to cross a gap can rely also on auditory perception, though unconsciously.

Although implications on experimental design, i.e., getting confounding factors under control and finding a situation where bimodal perception is relevant, required us to study a case which is seemingly constructed, we think that it does represent particular safety critical situations. Dangerous moments often arise suddenly and a road user has little time to react while her perception is over-strained by an unusual and highly critical situation resulting in various kinds of perceptual uncertainties. To avoid losing safety-critical amounts of time, actions need to be carried out rather "intuitively" as it is impossible to deliberately evaluate the risk of each possible reaction.

Concerning technical issues, we investigate task performance with HRTF-filtering as representative state-of-the-art high-quality condition, as well as conventional stereo audio-rendering as minimum standard low-quality condition, and compare it to a unimodal control condition where participants acted without sound. Rendering conventional stereo sound provides an audio stimulus where, compared to our high-quality condition, several critical spatial cues (e.g. inter-aural time-differences) are missing. However, we assume that at least semantic congruency (i.e., car and engine-sound) [Laurienti et al. 2004] and temporal alignment are maintained in both audio-visual conditions.

4. METHOD AND APPARATUS

4.1 Setup

To provide a high degree of immersion, the experiments were carried out with a large projection screen (240cm by 185cm) displaying the scene in mono at a resolution of 1280 by 1024 pixels. The application was run on a laptop computer equipped with an Intel Core 2 Duo T9300 running at 2.5 GHz with 2 GB RAM, and an NVIDIA Geforce 8600M GT

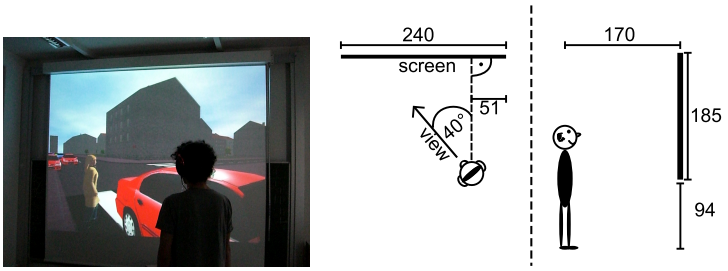


Fig. 1: *Experimental setup: Large screen VE (left) and schematic description of the viewpoint, screen size and listener position (right).*

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 rendering the visual scene, we used the open source engine Ogre ([Ogre3D]). 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 ([J. Blauert 1999]) using individual head-related transfer functions (HRTF) which were chosen for each participant from a database of example HRTFs (see Section 4.6). Besides HRTF-based filtering of sound signals, the sound library performs distance attenuation and simulates Doppler effects for sound sources in motion. For stereo rendering, 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. To avoid the necessity for head tracking, participants were instructed to stand in a fixed position, to look in a fixed direction and to avoid head motions, providing a fixed viewpoint. The viewpoint was off-axis and angled to the left. This configuration (see Figure 1) was chosen during the pilot studies in order to give the participants the best overview on the task-relevant screen regions, while maintaining a high degree of immersion. Head position, view direction and the coordinates of the screen corners were used to configure the engine so that viewpoint and listener position of the virtual environment align with the spatial layout of the setup and to account for off-axis viewing. This required us to reconfigure for each participant the height due to individual body heights. As input device we used the Nintendo Wii-Remote console controller, which is wireless and has a simple button layout, which is favorable when conducting the experiment with users without computer experience.

4.2 Scenario

The scenario rendered in the VE application is an urban environment containing buildings, roads, crosswalks, sidewalks and a sky-blue background at bright daylight. One vehicle type was used in the experiment, which was modelled to preserve real-world proportions and instanced with different colors to convey the impression of different cars, while maintaining a constant geometrical outline of each car. During the preparation of this experiment, we observed that using different car models leads to the effect that some participants tend to preferably cross the street after a car with certain geometrical properties. The car is a model of a 1992 Nissan Primera P10, with sound sources attached for the tires and the engine, which were all placed at the adequate position in the virtual carriage (Figure 2) and played back a looped sound with the corresponding noise. Sound files were taken

Fig. 2: *Visualization of the sound sources attached to the car to simulate the noise of the engine and tires.*

cue	Mute	Stereo	Spatial
Visual Depth & Motion Cues	+	+	+
Doppler Frequency Shift	-	+	+
Distance Attenuation	-	+	+
Interaural Level Difference	-	+	+
Interaural Time Difference	-	-	+
Monaural Spectral Cues	-	-	+

Table I: *Important cues for visual or auditory spatialization. "+" denotes whether the cue is rendered in the respective condition.*

from a free online database [Freesound]. For the engine we used an engine sound recorded from a Renault People Carrier. Since recording the sound of a single tire under anechoic conditions is very difficult and thus such recordings are not available, we used the sound of a vacuum cleaner, which was the most similar sound we found.

4.3 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 I we list the most important cues for distance and motion perception and whether they are absent or present in the respective conditions. Each participant performed the experiment in one of the three conditions, which we denote as follows:

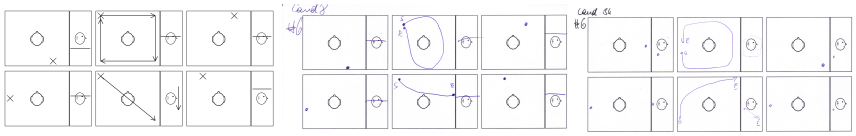
—*Spatial*: high-quality spatialized sound rendering with HRTF filtering

—*Stereo*: conventional stereo sound rendering with low-quality spatialization

—*Mute* : unimodal baseline condition

4.4 Participants

All participants were recruited from the university campus and were paid a participation fee. All tested participants reported normal or corrected to normal vision and normal hearing. Ages ranged from 19 to 32 (mean 24.3). In total we tested 48 participants (24 male) distributed uniformly over the three conditions. To avoid bias due to potential gender-related performance differences, male participants were also uniformly distributed over all



(a) **(b)** **(c)**

Fig. 3: *The test sheets used to score the applicability of a particular HRTF: reference sheet **(a)**, example sheets for a selected **(b)** and a rejected **(c)** HRTF candidate.*

conditions.

Apart from the need to avoid too many trials per participant, the use of a between-subject methodology was motivated by our experience that properly counterbalancing repeated measurements for within-subject analysis becomes problematic due to the fact that participants tend to adapt/train their perceptional strategy to a particular condition and continue to keep this strategy for the next condition being tested. For instance, when starting with the *Mute* condition, a participant develops skills particularly trained to base her recognition on pure visual properties. This can result in the tendency to fully ignore audio information provided in the next tested condition, since attention remains focused on visual properties as trained in the previous block.

4.5 Rationale

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*

4.6 HRTF selection

For the Spatial condition, an HRTF function is required for spatialized audio rendering. HRTFs should ideally be created for each individual participant. 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 ([Listen]) and determined the most appropriate for each participant. All HRTFs in this database were recorded with an azimuthal resolution of 5° and 11 elevations ($\pm 40^\circ$,$\pm 25^\circ$,$\pm 15^\circ$,$\pm 7.5^\circ$,0°,60°,90°).

To select an appropriate HRTF, [Moeck et al. 2007] used an application which lets candidates choose an HRTF by performing a “point and click” pre-test. However, after initial

experiments with this we were afraid of subjective mistakes caused by distraction and ventriloquism effects resulting from stimuli on a visual display. Therefore we designed a formalized selection procedure which requires no display. Using six candidate HRTFs, the method was efficient enough not to exceed the participants' patience. 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 her head into a transversal plane for the azimuthal angle and into a sagittal plane for the elevation (Figure 3). Each test scenario was presented until the participant reported confidence about her perception, assuring participants 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. After a pre-selection which removed all HRTF candidates that had a clear mistake (at least one clear front-to-back, left-right or top-bottom confusion), the HRTF with the best fit between actual and perceived angles was chosen by a subjective evaluation of the drawings by the person who conducted the experiment. In most cases, only one candidate was left after the pre-selection and in some cases two were left for deeper comparison. To avoid negative effects resulting from strong spatial misalignments of the perceived auditory scene and the visual scene, we replaced participants (about 40%) where no HRTF satisfied the pre-selection criteria.

It should be noted that an exact scientific evaluation of HRTF selection methods is beyond the scope of this paper, and the method described above is an intuitive and fast solution. If an "ideal" HRTF selection procedure is used, we expect the results of our experiment to become even more significant than those obtained in this work.

4.7 Procedure

Trial and task description. In each trial the gap crossing task is carried out as follows. The subject stands on the sidewalk facing towards the double-lane road in a first-person view. Cars come from the left side in a running stream at specified intervals and at a constant speed of 50 km/h. The test participant can voluntarily select a gap if she thinks it is safe to cross by pressing 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 (Figure 4). The goal is to cross the road several times without being hit by the oncoming cars. The virtual distance until the avatar reaches the Point of Safety is $2.14m$ ($= 1.9m$ car width$+0.24m$ distance to the lane). This corresponds to a walking time of $1.1s$ at a walking speed of $7km/h$ (Figure 5). A trial is counted as an accident if the bounding volumes of the car and the avatar have an intersection. It ends either with an accident or when the avatar reaches the point of safety. In case of an accident, the participant hears an acoustic alarm signal at the moment of the crash and gets a negative written feedback on the black screen displayed after each trial. Moreover, the black screen shows a status report about the current progress, and a success rate is displayed with a comment complimenting the participant's success to reward good performance and increase her motivation. By pressing a button, the participant controls when the next trial begins.

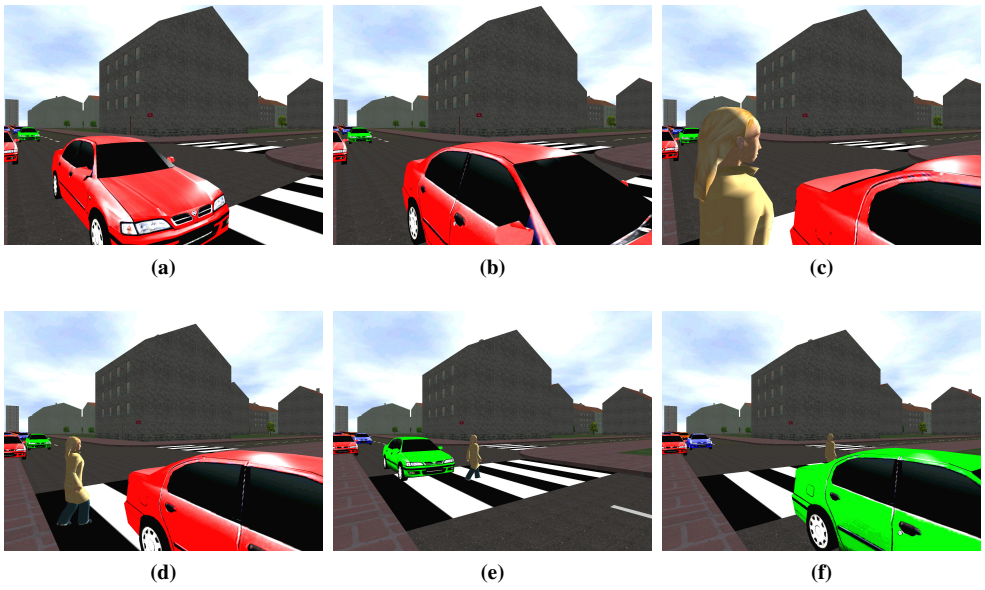


Fig. 4: ***Screenshots of the virtual gap crossing:*** ***(a)*** *first person view of car approaching,* ***(b)*** *starting to cross the street,* ***(c)*** *camera switches to third person view,* ***(d)*** *avatar is walking across the street,* ***(e)*** *avatar passes outer bound of car* ***(f)*** *avatar crossed the road successfully*

Note that in contrast to previous traffic gap choice experiments (e.g. [Clancy et al. 2006]), we had no motion-tracking system and the movement speed is not under the participant's control. 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 uncrossable after all. We would consider this a "negative outcome" of the trial since the participant decides to walk faster in order to avoid getting hit by a car due to a misjudgement.

Configuration. Pilot studies showed that effects of auditory cues can be better observed when we approach the limits of visual perception. If vision provides all information required to accomplish a task very clearly (e.g. strong contrast between crossable and uncrossable gap sizes), auditory information is of low value. We hence increased ambiguities in vision by using gap sizes which could be hardly distinguished visually. In particular, only two gap sizes were used: one assumed to be *crossable* and one assumed to be *uncrossable*. Both types of gaps appear randomly in the stream with an equal probability of 50%. Pilot testing revealed that a difference of 100 ms between both gap sizes at a car velocity of 50km/h was subtle enough to be visually distinguished only by guessing. Nevertheless, though participants reported that they had the impression not to be able to discriminate *crossable* and *uncrossable* gaps, there is a clear evidence that intuition allowed them to judge above chance level (Table II).

The size of the *uncrossable* gap was determined during a pilot study. By means of a staircase procedure, we found the threshold where participants have a chance below 10% to

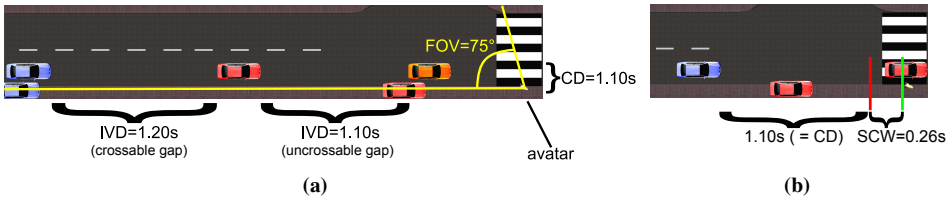


Fig. 5: ***(a):****Birds-eye-view with inter-vehicle distance (IVD) in time domain between the cars driving at a constant speed of 50 km/h. The IVD is depicted for uncrossable and crossable gaps. Two parking cars were placed in the field of view as distractors, but in a way that they do not significantly hide anything relevant for the task. The crossing duration (CD) to cross the dangerous part of the road refers to a walking speed of 7 km/h.* ***(b):*** *spatial illustration of the temporal Safe Crossing Window (SCW, green: begin time; red: end time) within which it is possible to start a successful street crossing (assuming a crossable gap was selected). Note that the start of the SCW is roughly in the middle of the lead car because the car moves this distance until the avatar reaches the part of the lane where cars actually move.*

	k $(avg \# of\ chosen\ crossable\ g.)$	$B_{60,0.5}(x \geq k)$
mute	37	0.02
stereo	39	0.007
spatial	42	0.0008

Table II: *Probabilities of observing a result of k or more correctly chosen gaps within 60 trials. The p-values were computed with the Cumulative Binomial Distribution, assuming that participants operate on chance level ($p = 0.5$). k is the average number of correctly identified crossable gaps in one condition. (The bars in Figure 6a show the same values as proportions.)*

cross the street without an accident. In particular, participants were presented with a scene where all gaps between cars were equal. The task was to cross the street without accident. Starting with a gap size of $1700ms$, the size was lowered by $10ms$ for the next trial in case of a success, otherwise it remained constant. The termination condition was that a participant failed 22 times to cross a gap of one size. Terminating after 22 consecutive failures yields a probability of $B_{22,0.1}(X <= 0) < 0.1$ (Binomial distribution) that we observe this pattern of failures under the assumption that the actual success rate is $P(success) >= 0.1$. In other words, we can reject the null-hypothesis $P(success) >= 0.1$ with a significance level of 10%. This low probability was necessary, since participants should experience a clear (negative) feedback when they choose an *uncrossable* gap to cross the street. The median ($1.10s$) of the thresholds measured in the pilot study was then used as inter-vehicle distance (IVD) for *uncrossable* gaps. Consequently an IVD of $1.20s$ was then used to define *crossable* gaps. Note that the smallest gap that could be theoretically crossed in our setup is $0.98s$. To illustrate the spatio-temporal relations of the task scenario, we labeled a birds-eye-view of the gap-crossing scene depicted in Figure 5.

Training block. To accustom participants with the virtual gap-crossing task, we found that the best training was to reuse the above-mentioned procedure (with 15 instead of 22 failure trials as stop condition and a decrement of $20ms$ instead of $10ms$), where participants crossed subsequently smaller gaps until the difficulty was too hard.

Main block. The main block comprised 60 trials. Participants were not informed that there are only two different gap sizes, where one is impossible to cross without getting hit by a

Tested factor	sel. *crossable* gaps (dv_1)	successr.: *crossable* gaps (dv_2)	successful trials (dv_3)
gender (ANOVA)	$F(2,43) = 0.32$, $p = 0.58$	$F(2,43) = 14.75$, $p < 0.001$'***'	$F(2,45) = 16.11$, $p < 0.001$'***'
wtime (Kendall)	$\tau = 0.47$, $p < 0.001$'***'	$\tau = -0.12$ $p = 0.25$	$\tau = 0.09$, $p = 0.35$
log(wtime) (ANOVA)	$F(2,44) = 30.0$, $p < 0.001$,'***' $coef = 0.13$	$F(2,43) = 4.64$, $p = 0.037$,'*' $coef = -0.09$	$F(2,44) = 0.11$, $p = 0.74$, $coef = 0.01$

Table III: ***Effects of confounding factors:*** *To estimate the effect of gender, an ANOVA between the linear model with and without this factor was computed. Kendall's τ was used to investigate potential correlations between waiting time and all dependent variables being analyzed. Moreover we also computed an ANOVA between the linear models with and without a log-linear regression function over waiting time. In the last row we also report the coefficients ($coef$) for the log-linear regression function which were obtained by a least-squares fit on the arcsine square root transformed data.*

car (*uncrossable*) and the other is possible to cross (*crossable*). Participants were instructed to intuitively choose a gap size they find safe to cross. For each trial we recorded whether the chosen gap was *crossable* and whether the participant was able to reach the other side of the street without an accident, i.e., whether the participant chose the correct moment to start crossing the gap.

5. RESULTS

5.1 Preliminaries

Dependent variables. Concerning the main block, our focus was on three variables: First, the proportion of trials where a *crossable* gap was selected (Figure 6a) (dependent variable dv_1), second, the fraction of successfully crossed *crossable* gaps (Figure 6b) (dv_2), and third, the overall success rate in the gap crossing task (dv_3), computed from the number of trials where a *crossable* gap was selected **and** successfully crossed, divided by the total number of valid trials (Figure 6c). A few of the *uncrossable* gaps were also crossed successfully (on average 0.5 gaps per participant), but we ignored those trials (Figure 7a) because those successes can be attributed to pure chance.

Stabilizing variance. Since all dependent variables are proportional data and thus derive from a binomial distribution, prior to any statistical analysis (but not in figures illustrating the results) we used the arcsine square root transform ($arcsin(\sqrt{x})$) to stabilize variance and increase the accuracy of the Gaussian approximation assumed in most statistical tools used in our analysis.

Condition and confounding factors (Table III). Apart from the independent variable condition (spatial, stereo, mute), we observed that two other variables, namely ***gender*** and ***average waiting time***, have significant effects on some of the investigated dependent variables:

- ***Gender:*** Gender is an independent variable which was controlled by distributing male participants equally over all conditions. An ANOVA revealed a highly significant ($p < 0.001$) effect of adding gender to the linear models for dv_2 and dv_3 (Table III).

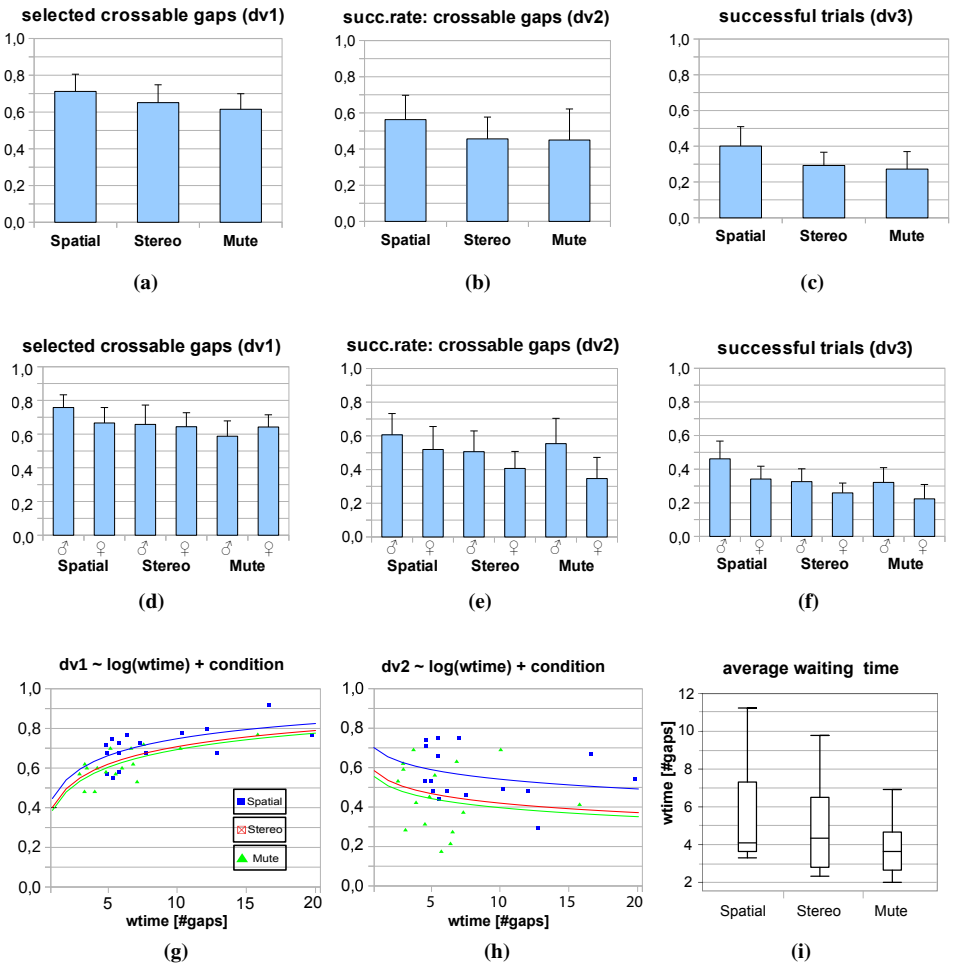


Fig. 6: ***Results :*** *Mean and standard-deviations of* ***(a)*** *proportion of trials where a crossable gap was selected ($dv1$);* ***(b)*** *fraction of successfully crossed crossable gaps ($dv2$);* ***(c)*** *proportion of successful trials ($dv3$). Figures* ***(d-f)*** *show the same results but split by gender. Figures* ***(g-h)*** *shall illustrate the effect of per participant average waiting times (wtime) on those dependent variables (i.e. dv1 and dv2) which are affected on a significant level (dv3 was not significantly influenced). To increase clarity of presentation the factor gender was ignored in the regression model used for illustration of dv2 and the log-linear regression, which was fitted to arcsin square root transformed data, was transformed back to the original scale. Figure* ***(i)*** *depicts distributions of per participant average waiting times (expressed as number of skipped gaps) as boxplots by condition.*

Overall, the trend is that female participants performed significantly worse than their male colleagues (see results split by gender in Figures 6d - 6f). Hence gender was added to the statistical models for $dv2$ and $dv3$, which increased statistical power due to a reduced variance and better fit with the normal distribution in the residuals.

- ***Average waiting time (wtime):*** The number of gaps participants waited until crossing the street could not be controlled in our experiment, since participants were free to choose when to cross. We computed the per-participant average waiting times

Alt.Hypoth.	sel. *crossable* gaps ($dv1$)	successr.: *crossable* gaps ($dv2$)	successful trials ($dv3$)
¬(Sp=St=M)	$F(2,44) = 2.44$ $p = 0.099$'.'	$F(2,45) = 5.74$ $p = 0.006$'**'	$F(2,46) = 11.17$ $p < 0.001$'***'
H_1(Sp > M)	$t = 2.13, p = 0.020$'.'	$t = 3.14, p = 0.002$'**'	$t = 3.32, p = 0.001$'**'
H_2(Sp > St)	$t = 1.62, p = 0.057$'.'	$t = 2.72, p = 0.005$'**'	$t = 3.31, p = 0.001$'**'
H_3(St > M)	$t = 0.64, p = 0.26$	$t = 0.41, p = 0.34$	$t = 0.51, p = 0.31$

Table IV: ***Hypothesis tests:*** *Results by dependent variable and alternative hypothesis. ANOVA was computed to evaluate general effects of condition and one-sided Student t-tests ($dof = 30$) to test our hypotheses (Section 4.5) by pairwise comparisons. Significance tests were corrected with the Bonferoni-Holm sequentially rejective procedure and denoted as '.' for a (global) $\alpha^{\cdot} = 0.1$,'**' for $\alpha^{**} = 0.01$ and '***' for $\alpha^{***} = 0.001$.*

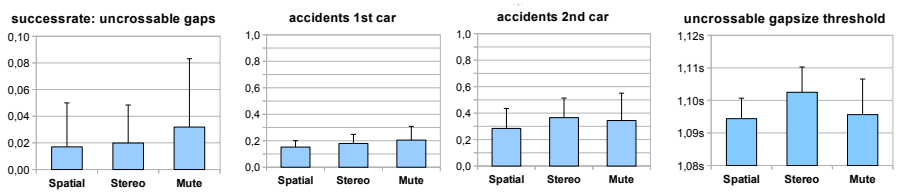
($wtime$) to be able to relate this factor with the dependent variables under investigation. For a better intuition, we used the number of gaps a participant saw before crossing the street as the units of $wtime$ [4]. This variable was not normally distributed in either of the conditions (Shapiro-Wilk: $W \leq 0.88$, $p \leq 0.04$), as is also obvious from the box-plots shown in Figure 6i. Though there was no significant effect of condition (Kruskal-Wallis rank sum test: $\chi^2 = 3.79, dof = 2, p = 0.15$), we suspect a trend that waiting time could be biased by condition. A main concern was that waiting time was by trend longest in the *Spatial* condition. So we tested whether any of the dependent variables was correlated with waiting time. Due to a significant violation of the normality assumption, a non-parametric correlation was computed. A Kendall's τ rank correlation (Table III) revealed a moderately positive ($\tau = 0.47$) but highly significant ($p < 0.001$) correlation between the average waiting time and the performance in the selection of *crossable* gaps (dv_1). There is a weak negative correlation between waiting time and dv_2 ($\tau = -0.12$) and a weak positive correlation with dv_3 ($\tau = 0.09$), which is not significant in either case. To compensate a potential bias due to a possible causal relation between waiting time and the independent variables, we used a log-linear regression, which yields the best fit (compared to a linear regression). An ANOVA (Table III) showed that effects of the log-linear regression over waiting time are highly significant for dv_1. Though the correlation between waiting time and performance in crossing *crossable* gaps (dv_2) is not significant ($p = 0.25$), it nevertheless turned out that a log-linear regression over waiting time improves the fit of the according linear model at a significant level ($p = 0.037$).

5.2 Testing our hypotheses (Table IV)

In order to investigate our main hypotheses, the effect of condition was tested on linear models where those factors (from $condition, gender$ and $wtime$) which were significant were included (according to the ANOVA p-values in Table III). None of the used models includes interactions between factors, since no significant interactions were observed when testing all possible variations.

The overall strategy was to test first for violations of normality and equality of variance assumptions in the residuals of the linear model used for regression, and second for the effect of condition by an ANOVA between the linear models with and without the factor

[4] With one unit corresponding in average to $1.48s = (1.1s + 1.2s)/2$ (average gap size) $+0.33s$ (length of a car)



(a) **(b)** **(c)** **(d)**

Fig. 7: ***(a) Ignored trials:*** *In a few trials, participants luckily managed to cross the street successfully though an uncrossable gap was chosen. We assume these successes result from luck rather than the condition. The rates were computed per participant by counting the number of successfully crossed uncrossable gaps and normalizing with the total number of uncrossable gaps being selected for crossing. No significant effect of condition (ANOVA: $F(2, 45) = 1.0$, $p = 0.38$) could be observed in these samples and we assume that these incidents are due to chance and ignored them.*
(b + c) Accidents with first or second car : *mean and standard-deviation of the proportion of trials with accidents with the respective car. Proportions were computed relative to the number of trials where a crossable gap was selected. Hence, the sum of both proportions equals the results shown in Figure 6b.*
(d) Results from the training block: *In the training block, participants had to cross subsequently smaller gaps until they failed to cross 15 times in a row. The bars show mean and standard deviation of the minimum gap sizes participants saw in the training.*

condition, and third, to perform pairwise comparisons with a one-sided Student's t-test according to the hypotheses specified in Section 4.5. To account for the fact that for pairwise comparisons three hypotheses (H_1, H_2, H_3) are tested simultaneously, significance levels were adjusted with the Bonferoni-Holm sequential rejective procedure [5] to compensate stochastic advantages.

- **Selected *crossable* gaps (dv_1):** For the proportion of trials where a *crossable* gap was selected, we used a linear model including condition and a log-linear function of average waiting time ($dv_1 \sim condition + log(wtime)$, Figure 6g). Neither normality (Shapiro-Wilk: $W = 0.97, p = 0.25$) nor equality of variance (Levene's Test statistic $= 0.61, p = 0.55$) assumptions were violated. The regression coefficient for the log-linear function of waiting time was 0.13, which shows that this factor and the performance in selecting *crossable* gaps are moderately proportional. The effect of condition had only a weak significance in the ANOVA test. After the Bonferoni-Holm correction on the Student's t results there was no significance for hypothesis H_3 ($p = 0.26$) and a weak significance ($\alpha^{\cdot} = 0.1$) for hypotheses H_1 ($p = 0.02 < \alpha^{\cdot}/3$) and H_2 ($p = 0.057 < \alpha^{\cdot}/2$). The respective offsets between the log-linear regression functions were decent: 0.04 (*Spatial* - *Stereo*) and 0.06 (*Spatial* - *Mute*) in the untransformed scale (i.e. differences in proportion values)[6].

  Without the log-linear regression over $wtime$, the effect of condition (ANOVA: $F(2, 45) = 9.09, p = 0.004$) is highly significant ($\alpha^{**} = 0.01$). And the pairwise comparisons show high significance for alternative hypothesis H_1 ($t = 3.14, p =$

[5] The Bonferoni-Holm sequential rejective procedure [Holm 1979] sorts all hypotheses by p-values in increasing order. All ($= n = 3$) hypotheses are tested sequentially on significance levels $\alpha/n - i$, with i as the rank in the sorted list (beginning with $i = 0$) until the first hypothesis is rejected.
[6] Since the arcsin square root function is a non-linear function, the back-transformed offsets may vary slightly with $wtime$. We chose to evaluate the difference for the back-transformation at the median $wtime = 4$.

$0.002 < \alpha^{**}/3$), but weak significance for hypothesis H_2 ($t = 1.81, p = 0.04 < \alpha/2$).

- **Success rate: *crossable* gaps (dv_2) :** The fraction of successfully crossed *crossable* gaps was fitted to a linear model containing condition, gender and a log-linear function of average waiting time ($dv_2 \sim condition + gender + log(wtime)$). Neither normality (Shapiro-Wilk: $W = 0.97, p = 0.35$) nor equality of variance (Levene's Test statistic $= 1.33, p = 0.28$) assumptions were violated. The regression coefficient of the log-linear function of waiting time was -0.09, indicating that waiting time and performance in crossing *crossable* gaps are marginally inversely proportional (see Figure 6h). The effect of condition is highly significant ($p = 0.006$). The Student's t-test with Bonferoni-Holm correction revealed no significance for hypothesis H_3 ($p = 0.34$) and high significance for both hypotheses H_1 ($p = 0.002 < \alpha^{**}/3$) and H_2 ($p = 0.005 < \alpha^{**}/2$). There are clear offsets between the log-linear regression functions: 0.12 for *Spatial–Stereo* and 0.14 for *Spatial–Mute* (in the untransformed scale evaluated at $wtime = 4$).

  The results are also significant without the factor $wtime$: Similarity of the residual distribution to a normal distribution is worse (Shapiro-Wilk: W=0.97,p=0.20), but still without significant deviation, whereas residual variances are more stable (Levene's Test statistic $= 0.35, p = 0.71$) than with the log-linear regression over waiting time. An ANOVA showed that the effect of condition is still significant ($F(2, 46) = 3.96, p = 0.026$) and also the pairwise comparisons yield significant results ($\alpha^* = 0.05$) for alternative hypotheses H_1 ($t = 2.53, p = 0.008 < \alpha^*/3$) and H_2 ($t = 2.33, p = 0.012 < \alpha^*/2$).

  Moreover we investigated how many of the accidents were with the first or second car of a gap (Figures 7b and 7c). We used for both variables a linear model including the factors condition, gender and a log-linear function of average waiting time ($x \sim condition + gender + log(wtime)$). There was no violation of the normality assumption for both variables (Shapiro-Wilk *1st car*: $W = 0.98, p = 0.59$; *2nd car*: $W = 0.97, p = 0.36$) and equality of variance can be assumed for the number of accidents with the second car (Levene's Test statistic$= 1.95, p = 0.15$), but for the number of accidents with the first car we should not assume equality of variance (Levene's Test statistic $= 2.97, p = 0.062$) and better use non-parametric measures. Though there is a significant effect of condition on dv_2 and hence in the sum of accidents with the first and second car, there is no significant effect on the number of accidents with the first car (Kruskal-Wallis: $\chi^2 = 2.08, dof = 2, p = 0.35$) and an effect with a weak significance for the number of accidents with the second car (ANOVA: $F(2, 45) = 3.09, p = 0.056$). Hence pairwise comparisons were only reasonable for the number of accidents with the second car. However, for none of our hypotheses the t-test gave a significant result (one-sided Student's t with 30 dof, H_1: $t = -0.86, p = 0.20$; H_2: $t = -1.53, p = 0.068$; H_3: $t = 0.43, p = 0.67$). A reason for the low significance could be that participants tend to a timing strategy with either a higher risk to hit the first car or a higher risk to hit the second car, and that hence the variance introduced by participant-related factors diminishes the effects of condition.

- **Successful trials (dv_3):** The proportion of trials where a participant succeeded in

crossing the street without an accident was fitted to a model including the factors condition and gender ($dv_3 \sim condition + gender$). The normality (Shapiro-Wilk: $W = 0.97, p = 0.38$) and equality of variance (Levene's Test statistic $= 0.42, p = 0.67$) assumptions were not violated. The effect of condition has a high significance ($p < 0.001$). The test results for alternative hypothesis H_1 ($p = 0.001 < \alpha^{**}/3$) and H_2 ($p = 0.001 < \alpha^{**}/2$) were highly significant, while there was no significance concerning hypothesis H_3 ($p = 0.31$).

Results do not change when adding $log(wtime)$ to the regression model for dv_3. The effect of condition is still highly significant (ANOVA: $F(2, 45) = 9.77, p < 0.001$), and also pairwise comparisons yield similar results (H_1:$t = 4.12, p < 0.001$, H_2:$t = 3.50, p < 0.001$, H_3:$t = 0.7, p = 0.49$).

Results from the training block. Figure 7d shows the mean gap-size thresholds where participants failed to cross in the training block. Probably due to the large discrete steps ($\Delta t = 20ms$) of the staircase procedure, the residuals were not normally distributed (Shapiro-Wilk: $W = 0.81, p < 0.001$) and we hence used a non-parametric test to test for the effect of condition, which is highly significant (Kruskal-Wallis rank sum: $\chi^2 = 11.12, dof = 2, p = 0.003$). There was no difference between the *Spatial* and *Mute* condition (two-sided Wilcoxon rank sum: $W = 127, p = 0.98$), whereas the (Bonferoni-corrected) contrast between *Stereo* and *Mute* conditions is significant ($W = 193.5, p = 0.008 < \alpha^*/2$) and highly significant between *Stereo* and *Spatial* ($W = 202.5, p = 0.002 < \alpha^{**}/3$), suggesting that participants from the *Stereo* group performed significantly worse than participants from other groups. However, results from the training block should not be taken too seriously. With its coarse step sizes, this procedure was not designed to be sufficiently accurate and sensitive to subtle effects of sound. The procedure was rather configured to familiarize participants with the difficult task as quickly as possible.

6. DISCUSSION

6.1 Clear effects of high-quality sound and no effects of stereo sound

Overall, we observed clearly positive results for the dependent variables reflecting performance in timing the crossing of *crossable* gaps (dv_2) and the gap-crossing task as a whole (dv_3), supporting the hypotheses H_1 and H_2 that spatialized audio rendering increases task performance. Compared to both control conditions, the high-quality sound condition yielded a relative increase of about 40% on average in the number of successful trials (dv_3). On the other hand, a statistically significant positive impact of conventional stereo rendering (H_3) could not be observed at all.

In general the results support our hypothesis that auditory spatial information rendered in high quality can be perceptually utilized to improve performance even in a higher-level task which is seemingly vision dominated. This applies in particular for a task involving accurate localization of moving objects and precise action timing. Since we could not observe a significant potential of standard stereo sound rendering to improve task performance, spatial cues provided by HRTF filtering seem to have distinct advantageous effects on task performance, whereas spatial cues in stereo sound, like distance attenuation and

interaural level differences, are less task supportive (in the case tested with our study). The results from the training block suggest that stereo sound may even decrease performance in some tasks where timing and estimation of spatio-temporal relations are crucial. Assuming that sound effects with wrong or inaccurate spatial cues can cause erroneous perception, the observation that low-quality sound may also decrease performance strengthens the hypothesis that spatial information of sound is perceptually utilized, if available.

6.2 Waiting time correlates with increased caution in selecting *crossable* gaps

Though the effect of condition on the *selection* of *crossable* gaps (dv_1) is highly significant without decorrelating waiting time ($p = 0.004$), the effect after decorrelation is not really significant ($p = 0.099$) and we need to reject our alternative hypotheses concerning this variable. A possible causal relation between waiting time and performance makes the decorrelation obligatory to exclude the possibility that other cognitive behaviours than bimodal integration are responsible for the effect of sound rendering. Though there was no significant effect of condition on the average per-participant waiting time, we suspect a trend that participants from the conditions with sound tend to wait longer – and longest in the high-quality sound condition. One reason could be that participants feel more immersed in an audio-visual environment, are more aware about the virtual danger of the situation and hence operate with enhanced carefulness. The highly significant correlation between waiting time and performance in choosing *crossable* gaps suggests that more deliberate behaviour increases the probability to select *crossable* gaps.

However, longer waiting times do not enhance performance in the timing of street crossings (dv_2) and the overall number of successful trials (dv_3). The relation between waiting time and dv_2 is rather marginally inverse proportional, i.e., participants with longer waiting times tend to perform worse in timing the street crossing actions. The different impact of waiting time on dv_1 and dv_2 can be due to the difference in the way decisions are carried out in the two corresponding sub-tasks: The selection of gaps to cross is a binary decision ("go or no-go?") and can involve a higher degree of deliberateness as a participant can withhold her decision until feeling more confident. On the other hand, the timing of street crossing actions requires a kind of quantitative response ("when?"), and thus a good sense for spatio-**temporal** relations. It is carried out on a rather reactive level and requires quick decisions for precise action timing. It seems that increased carefulness may rather hinder good reactions in this subtask.

The weak and non-significant correlation between waiting time and overall success rate suggests that advantages in the selection of *crossable* gaps through longer waiting times are partially compensated with a worse performance in crossing the street when a *crossable* gap was selected. An explanation for this behaviour could be that some participants focus on the selection task, while some others focus on the street crossing task. And higher waiting times correlate with a focus on the selection task. Hence, we think that the results for splitting the gap-crossing task into two subtask have to be interpreted with caution, since performance measures for both (dv_1 and dv_2) are not fully independent. We obtain more reliable and stable results when considering the performance in the overall task (dv_3), where the samples have lowest noise and yield the highest statistical power.

6.3 Bimodal integration reduces uncertainties in spatio-temporal perception

Though spatial precision in the auditory channel is (at least for normal people) much less precise than spatial perception in the visual channel, a significant effect measured by means of task performance could be revealed. Assuming an optimal audio-visual integration of uncertain spatial information according to the MLE model [Alais and Burr 2004], vision alone provides a lower precision than combined audio-visual percepts when participants estimate spatial properties of the traffic scene, such as the position of a moving car. While the spatial uncertainty of auditory stimuli depends on the spatial resolution of auditory perception, spatial uncertainty in visual perception is less intuitive to imagine. Experiments ([Alais and Burr 2004]) to validate the MLE model used blurring to simulate uncertainty of visual stimuli. But in a more general view, spatial uncertainty can arise also on other (higher) levels of ambiguity in visual perception.

We believe that in our experiment visual uncertainty is merely produced by motion and the inability of the user to keep direct gaze (i.e., foveal vision) simultaneously on every task-relevant region of the large screen, and audio-visual integration can particularly reduce the spatio-temporal uncertainty introduced by these factors. Our guess is that adding auditory information allows a user to update her internal mental model of the scenario more efficiently. For instance, while the car currently passing by is monitored mainly aurally, the visual focus can be oriented towards the oncoming car. The accuracy in spatialization of the aurally monitored closer car is maintained through effectively combining auditory cues with motion trajectory expectation (eased by constant velocity) and cues in peripheral vision. However, future studies, e.g. monitoring gaze with an eyetracker, are required to validate these speculations. Other relevant sources of uncertainty include the spatial extent and the perspective change of approaching cars, the subtle differences of *crossable* and *uncrossable* gaps and the projection of a 3D scene to a 2D screen.

7. CAVEAT

7.1 Methodological limitations

In general, our study has revealed the existence of bimodal advantages in virtual traffic scenarios, at least when sounds are spatialized with high-quality rendering. Unfortunately, investigating perceptual performance in a complex task scenario providing rich and naturalistic stimuli has to be traded against the detail at which perceptual factors can be analyzed. Neuro-scientific studies, for example, strictly isolate pure tasks (e.g. discrimination or detection) by strong simplifications under restrictive lab conditions and are hence able to much better control confounding factors and independent variables. With our experiment we can not answer questions like: does bimodal task facilitation result from a parallelization advantage due to separate processing of redundant information ([Raab 1962]) or from bimodal coactivation ([Miller 1982])? Or: does estimation accuracy actually approach the optimum of an MLE model ([Ernst and Banks 2002])? A main requirement of such experiments are unimodal control conditions in both modalities (1st question) or to sample data for various and perfectly controlled levels of uncertainty (2nd question). We rather addressed application-relevant issues, in particular the implications for VE technology, human computer interaction and traffic safety. Our conclusion is that bimodal (or multi-

modal) aspects of perception are an issue when realistic and task-supportive VEs should be designed. On the other hand, our simulator experiment indicates that bimodal perception can also be an issue of traffic safety, since pedestrians are capable of improving their perceptual performance due to audio-visual integration.

7.2 Limitations on implications for realistic traffic scenarios

Since we needed to push the perceptual system to its limits, the gap sizes in our experiment were exceptionally small compared to realistic traffic crossing. Together with the fact that participants had to select between only two marginally different gap-sizes (of $1.1s$ and $1.2s$ inter-vehicle distance), where the safer one is only crossable with a probability of about 50%, the degree of difficulty in our gap-crossing task should not be compared to standard traffic situations where pedestrians typically choose to cross gaps with about $3.5s$ to $5s$ in order to cross safely. Moreover, we used only one car speed and one vehicle type to avoid further variations in the variables being observed. But we expect that bimodal integration also increases accuracy in the perception of varying velocities. Another important cue in the perception of spatial properties of sound are changes of the auditory signal due to head movements and rotations. Due to the absence of a head-tracking system, participants were instructed to keep their heads in a constant predefined position. A more realistic scenario would also require to simulate pedestrian walking conditions in a way that perception can directly interact with motor control. As we used a simple button which is pressed once and the avatar then walks without any user control, we needed to place the avatar's starting point exceptionally close to the stream of cars (distance of 24 cm) and to use a relatively high walking speed (7 km/h) to maintain an immediate relation between perception, action and consequences. Otherwise, behaviour would be influenced by further cognitive processes, since a participant also has to make predictions when the avatar will arrive in the dangerous zone.

Overall, this experiment can be considered a preliminary investigation to clarify whether it is generally possible to observe bimodal effects in virtual environments and in particular in virtual traffic scenarios. To draw further conclusions about the role of bimodal integration for traffic safety issues or virtual safety training, our study can be used as a starting point, and future studies should find experimental designs where we can observe these effects in more realistic configurations. The most important questions that should be answered with follow-up studies concern the effects of sound under conditions where the speed, type and sound of vehicles varies, or in extended VE setups with stereoscopic displays, head-tracking to allow free head and viewpoint movements, and more realistic walking conditions (e.g. like in [Schwebel et al. 2008]).

However, our experience with this study and the preceding pilots suggests that bimodal advantages are difficult to observe in everyday actions in regular traffic situations (e.g. crossing gaps of $5s$ length). Since the rational behaviour of a traffic user does not accept any expectable risks and does not make decisions by fully exploiting perceptual capacities, most situations lead to a very deliberate behaviour where decisions are merely the result of cognitive reasoning, traffic regulations and other rules that a traffic user has learned in order to maximize safety. This results in the problem that complex cognitive processes are involved between the levels of perception where we assume the highest impact of sound and the behaviour we can observe by measuring task performance. Our strategy to solve

this measurement problem was to increase the significance of bimodal perception by pushing task difficulty to a level where tolerances for perceptual misjudgments are low and hence perceptual inaccuracies are directly observable in a decrease of task performance.

We believe that bimodal advantages are best observable in moments requiring spontaneous reactions. E.g. situations where threats arise surprisingly and cause affective reactions of a user trying to safe her health in the dangerous moment. Scenarios which resemble the perceptual challenges imposed by our experiment can be for example risky passing maneuvers which turn out to be dangerous since unexpected events occur or the driver of the passing car made some fatal misjudgments. Such scenarios impose a high challenge on perception and even small inaccuracies may result in an accident. Using VE platforms should allow designing experiments where perception and behaviour in rare but safety-critical situations, like proposed by the above-mentioned examples, can be studied in future work. We claim that bimodal perception is one important factor at least in such scenarios.

8. CONCLUSION

We studied the impact of high-quality sound rendering on task performance with a task which requires a high amount of cognitive and perceptual resources in the context of a virtual traffic scenario. While in many situations visual perception provides a higher degree of certainty than hearing for spatialization tasks (the most prominent example being the ventriloquism effect), our study revealed that in a naturalistic task and context there are meaningful cases where vision does not fully capture auditory perception, and both modalities significantly contribute to improvements of spatial perception through bimodal task facilitation. This occurs because visual spatialization in information-rich environments, such as traffic scenarios, can have significant uncertainties, which can be reduced by adding auditory spatial information. Actually, the potential accuracy of auditory spatial perception seems to be widely underestimated, when considering for example the work of [Schiff and Oldak 1990], who found that blind people (who are supposed to be especially trained in auditory perception) are almost as good in time-to-contact estimation tasks as sighted ones.

Already previous experiments carried out by Neuro-scientists were able to confirm that multimodal estimation of spatial properties performs near the statistical optimum predicted by the MLE model (e.g. [Ernst and Banks 2002; Alais and Burr 2004]). However, simplistic tasks and stimuli, such as Gaussian blurs or noise signals, were required to construct suitable cases to take the required measurements. First studies with more complex stimuli using 3D Computer Graphics to render complex objects in a realistic fashion were carried out very recently [Suied et al. 2009; Nguyen et al. 2009], but due to slightly different hypotheses being tested, the studied tasks were simplistic (i.e., object detection or localization) and their scenes did not contain moving sound sources. The gap-crossing task carried out in our experiment contributes one case of study with complex dynamic stimuli of high realism and an everyday task, which brings forward bimodal effects to an ecological example with direct implications on traffic safety and for the design of perceptually optimized VEs.

REFERENCES

ALAIS, D. AND BURR, D. 2004. The ventriloquist effect results from near-optimal bimodal integration. *Current Biology 14,* 3 (February), 257–262.

ALAIS, D., MORRONE, C., AND BURR, D. 2006. Separate attentional resources for vision and audition. *Proceedings of the Royal Society B: Biological Sciences 273,* 1592, 1339–1345.

BARNECUTT, P. AND PFEFFER, K. 1995. Auditory perception of relative distance of traffic sounds. *Current Psychology 17*, 93–101.

BARRETO, A. B., JACKO, J. A., AND HUGH, P. 2007. Impact of spatial auditory feedback on the efficiency of iconic human-computer interfaces under conditions of visual impairment. *Comput. Hum. Behav. 23*, 1211–1231.

BART, O., KATZ, N., WEISS, P., L., AND JOSMAN, N. 2008. Street crossing by typically developed children in real and virtual environments. *OTJR: Occupation, Participation and Health 28*.

BARTON, BENJAMIN, K., SCHWEBEL, DAVID, C., AND MORRONGIELLO, BARBARA, A. 2007. Increasing children's safe pedestrian behaviors through simple skills training. *Journal of Pediatric Psychology 32*, 475–480.

BONNEEL, N., SUIED, C., VIAUD-DELMON, I., AND DRETTAKIS, G. 2010. Bimodal perception of audio-visual material properties for virtual environments. *ACM Transactions on Applied Perception to appear*.

BORMANN, K. 2005. Presence and the utility of audio spatialization. *Presence: Teleoper. Virtual Environ. 14*, 278–297.

BORMANN, K. 2006. Subjective performance. *Virtual Real. 9,* 4, 226–233.

BURSTEDDE, C., KLAUCK, K., SCHADSCHNEIDER, A., AND ZITTARTZ, J. 2001. Simulation of pedestrian dynamics using a two-dimensional cellular automaton. *Physica A: Statistical Mechanics and its Applications 295,* 3-4 (June), 507–525.

CAELLI, T. AND PORTER, D. 1980. On difficulties in localizing ambulance sirens. *Human Factors: The Journal of the Human Factors and Ergonomics Society 22*, 719–724.

CARRÉ, J.-R. AND ARANTXA, J. 2005. *A new method for analysing the pedestrian activity during the daily urban mobility and for measuring the pedestrian risk exposure*. Institut national de recherche sur les transports et leur sécurité.

CAVALLO, V., LOBJOIS, R., DOMMES, A., AND VIENNE, F. 2009. 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*.

CLANCY, T. A., RUCKLIDGE, J. J., AND OWEN, D. H. 2006. Road-crossing safety: A comparison of adolescents with and without ADHD. *Journal of Clinical Child and Adolescent Psychology 35*, 203–215.

CONNELLY, M., CONAGLEN, H., PARSONSON, B., AND ISLER, R. 1998. Child pedestrians crossing gap thresholds. *Accident Analysis and Prevention 30*, 443–453.

DAVIS, E. T., SCOTT, K., PAIR, J., HODGES, L. F., AND OLIVERIO, J. 1999. 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*. Georgia Institute of Technology, 1197–1201.

DELL, W. 2000. The use of 3d audio to improve auditory cues in aircraft. Tech. rep., Department of Computing Science, University of Glasgow. Available with experimental software from http://www.dcs.gla.ac.uk/research/gaag/dell/.

ERNST, M. O. AND BANKS, M. S. 2002. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature 415*, 429–433.

ERNST, M. O. AND BÜLTHOFF, H. H. 2004. Merging the senses into a robust percept. *Trends Cogn Sci 8,* 4 (April), 162–169.

FREESOUND. The freesound project. http://www.freesound.org/.

FRUIN, JOHN, J. 1971. *Pedestrian Planning and Design*. Alabama: Elevator World Inc, New York, New York.

HENDRIX, C. AND BARFIELD, W. 1995. 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)*. IEEE Computer Society, Washington, DC, USA, 74.

HOFMAN, P. M., RISWICK, J. G. V., AND OPSTAL, A. J. V. 1998. Relearning sound localization with new ears. *Nature Neuroscience 1*, 417–421.

HOLM, S. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics 6*, 65–70.

J. BLAUERT. 1999. *Spatial hearing – The psychophysics of human sound localization*. The MIT Press.

JIANG, W., JIANG, H., AND STEIN, B. E. 2002. Two corticotectal areas facilitate multisensory orientation behavior. *Journal of Cognitive Neuroscience 14,* 8, 1240–1255.

KINCHLA, R. 1974. Detecting target elements in multielement displays: a confusability model. *Percept Psychophys 15*, 149–158.

KING, A. J. AND PALMER, A. R. 1985. Integration of visual and auditory information in bimodal neurones in the guinea-pig superior colliculus. *Exp. Brain Research 60*, 492–500.

LAURIENTI, P., KRAFT, R., MALDJIAN, J., BURDETTE, J., AND WALLACE, M. 2004. Semantic congruence is a critical factor in multisensory behavioral performance. *Experimental Brain Research 158,* 4, 405–414.

LISTEN. LISTEN HRTF Database. http://recherche.ircam.fr/equipes/salles/listen/.

MASTOROPOULOU, G., DEBATTISTA, K., CHALMERS, A., AND TROSCIANKO, T. 2005. Auditory bias of visual attention for perceptually-guided selective rendering of animations. In *GRAPHITE 2005, sponsored by ACM SIGGRAPH, Dunedin, New Zealand*. ACM Press.

MCCOMAS, J., MACKAY, M., AND PIVIK, J. 2002. Effectiveness of virtual reality for teaching pedestrian safety. *CyberPsychology & Behavior 5*, 185–190.

MICHON, P.-E. AND DENIS, M. 2001. When and why are visual landmarks used in giving directions? In *COSIT 2001: Proceedings of the International Conference on Spatial Information Theory*. Springer-Verlag, London, UK, 292–305.

MILLER, J. 1982. Divided attention: evidence for coactivation with redundant signals. *Cognitive psychology 14,* 2 (4), 247–279.

MOECK, T., BONNEEL, N., TSINGOS, N., DRETTAKIS, G., VIAUD-DELMON, I., AND ALLOZA, D. 2007. Progressive perceptual audio rendering of complex scenes. In *I3D '07: Proceedings of the 2007 symposium on Interactive 3D graphics and games*. ACM, New York, NY, USA, 189–196.

NAVEH, Y., KATZ, N., AND WEISS, P. 2000. The effect of interactive virtual environment training on independent safe street crossing of right CVA patients with unilateral spatial neglect. *Proc. 3rd Intl Conf. Disability, Virtual Reality & Assoc. Tech*, 243–248.

NGUYEN, K.-V., SUIED, C., VIAUD-DELMON, I., AND WARUSFEL, O. 2009. Spatial audition in a static virtual environment: the role of auditory-visual interaction. *Journal of Virtual Reality and Broadcasting 6,* 5 (Mar.). urn:nbn:de:0009-6-17640,, ISSN 1860-2037.

OGRE3D. Ogre 3D : Object oriented graphics rendering engine. http://www.ogre3d.org/.

OXLEY, J., CONGIU, M., WHELAN, M., D'ELIO, A., AND CHARLTON, J. 2008. 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.

PLUMERT, J. M., KEARNEY, JOSEPH, K., AND CREMER, JAMES, F. 2007. Children's road crossing: A window into perceptual-motor development. *Current Directions in Psychological Science 16*, 255–258.

RAAB, D. 1962. Statistical facilitation of simple reaction times. *Transactions of the New York Academy of Science 24*, 574–590.

RIECKE, B. E., VÄLJAMÄE, A., AND SCHULTE-PELKUM, J. 2009. Moving sounds enhance the visually-induced self-motion illusion (circular vection) in virtual reality. *ACM Trans. Appl. Percept. 6,* 2, 1–27.

SCHIFF, W. AND OLDAK, R. 1990. Accuracy of judging time to arrival: Effects of modality, trajectory, and gender. *Journal of Experimental Psychology: Human Perception and Performance 16,* 2, 303–316.

SCHWEBEL, D. C., GAINES, J., AND SEVERSON, J. 2008. Validation of virtual reality as a tool to understand and prevent child pedestrian injury. *Accident Analysis and Prevention 40,* 4, 1394 – 1400.

SCHWEBEL, D. C. AND MCCLURE, L. A. 2010. Using virtual reality to train children in safe street-crossing skills. *Injury Prevention 16*, e1 – e5.

SEWARD, A. E., ASHMEAD, D. H., AND BODENHEIMER, B. 2007. Using virtual environments to assess time-to-contact judgments from pedestrian viewpoints. *ACM Trans. Appl. Percept. 4*, 18.

SIDAWAY, B., FAIRWEATHER, M., SEKIYA, H., AND MCNITT-GRAY, J. 1990. Time-to-collision estimation in a simulated driving task. *Human Factors 38 38,* 1, 101–113.

SIMPSON, G., JOHNSTON, L., AND RICHARDSON, M. 2003. An investigation of road crossing in a virtual environment. *Accident Analysis and Prevention 35*, 787–796.

SODNIK, J., DICKE, C., TOMAIČ, S., AND BILLINGHURST, M. 2008. A user study of auditory versus visual interfaces for use while driving. *Int. J. Hum.-Comput. Stud. 66,* 5, 318–332.

SODNIK, J., TOMAZIC, S., GRASSET, R., DUENSER, A., AND BILLINGHURST, M. 2006. Spatial sound localization in an augmented reality environment. In *OZCHI '06: Proceedings of the 18th Australia conference on Computer-Human Interaction*. ACM, New York, NY, USA, 111–118.

STEIN, B. E. AND MEREDITH, M. A. 1993. *The Merging of Senses*. MIT Press, Cambridge, MA.

STORMS, R. L. AND ZYDA, M. J. 2000. Interactions in perceived quality of auditory-visual displays. *Presence: Teleoper. Virtual Environ. 9,* 6, 557–580.

SUIED, C., BONNEEL, N., AND VIAUD-DELMON, I. 2009. Integration of auditory and visual information in the recognition of realistic objects. *Experimental Brain Research 194,* 1 (March), 91–102. http://www.springerlink.com/content/jj2w6g7366271237/fulltext.pdf.

SUNDARESWARAN, V., WANG, K., CHEN, S., BEHRINGER, R., MCGEE, J., TAM, C., AND ZAHORIK, P. 2003. 3d audio augmented reality: Implementation and experiments. In *ISMAR '03: Proceedings of the 2nd IEEE/ACM International Symposium on Mixed and Augmented Reality*. IEEE Computer Society, Washington, DC, USA, 296.

TAKAYUKI, S., KIYOHIDE, I., AND KAZUHIKO, M. 2004. The development of virtual 3d acoustic environment for training 'perception of crossability'. In *ICCHP* (2004-07-05), J. Klaus, K. Miesenberger, W. L. Zagler, and D. Burger, Eds. Lecture Notes in Computer Science, vol. 3118. Springer, 476–483.

TE VELDE, A., VAN DER KAMP, J., BARELA, J., AND SAVELSBERGH, G. 2005. Visual timing and adaptive behavior in a road-crossing simulation study. *Accident Analysis and Prevention 37*, 399–406.

TOM, A. AND DENIS, M. 2004. Language and spatial cognition: comparing the roles of landmarks and street names in route instructions. *Applied Cognitive Psychology 18*, 1213–1230.

TREISMANN, A. AND DAVIES, A. 1973. Divided attention to ear and eye. *Attention and Performance 4*, 101–117.

TUNG, Y.-C., LIU, Y.-C., AND OU, Y.-K. 2008. The pedestrian road-crossing behaviors between older and younger age groups. In *Proceedings of the 9th Asia Pacific Industrial Engineering & Management Systems Conference, APIEMS 2008*. 6–12.

WAN, B. AND ROUPHAIL, N., M. 2004. Using arena for simulation of pedestrian crossing in roundabout areas. *Transportation Research Record: Journal of the Transportation Research Board* 1878, 58–65.

YANG, J., DENG, W., WANG, J., LI, Q., AND WANG, Z. 2006. Modeling pedestrians' road crossing behavior in traffic system micro-simulation in china. *Transportation Research Part A: Policy and Practice 40,* 3 (March), 280–290.