THESIS FOR THE DEGREE OF LICENTIATE OF ENGINEERING

Technical Report No. R037/2005

Self-motion and Presence in the Perceptual Optimization of a Multisensory Virtual Reality Environment

ALEKSANDER VÄLJAMÄE

CHALMERS

Department of Signals and System
Division of Communication Systems
CHALMERS UNIVERSITY OF TECHNOLOGY
Göteborg, Sweden 2005

Self-motion and Presence in the Perceptual Optimization of a Multisensory Virtual Reality Environment

ALEKSANDER VÄLJAMÄE

This thesis has been prepared using Microsoft Word™

Technical Report No. R037/2005
ISSN 1403-266X

Department of Signals and System
Communication Systems Group
Chalmers University of Technology
SE-412 96 Gothenburg, SWEDEN

Front cover:
POEMS project (IST-2001-39223) logo by Pontus Larsson ©

Back cover:
Motion simulation setup at Chalmers Cognitive Room Acoustics Group (CCRAG), photograph by William Martens ©

Printed in Sweden
Chalmers Reproservice
Göteborg, Sweden, November 2005

Посвещается моей бабушке,
Маргарите Сергеевне Зайцевой

To my grandmother,
Margarita Sergeevna Zaitseva

Self-motion and Presence in the Perceptual Optimization of a Multisensory Virtual Reality Environment

ALEKSANDER VÄLJAMÄE

Division of Communication Systems
Division of Applied Acoustics
Chalmers University of Technology

Abstract

Determining the perceptually optimal resolution of multisensory rendering might help to foster the development of cost-effective, highly immersive multi-modal displays for mediated environments (e.g. virtual and augmented reality). The required sensory depth of stimulation can be quantified using human centered methodologies where end user experiences serve as a basis for uni- and cross-modal optimization of the sensory inputs. In the psychophysical studies presented in this thesis, self-reported presence and illusory self-motion (vection) indicated salience of auditory and multisensory cues in design of perceptually optimized motion simulators.

Contribution of auditory cues to illusory self-motion has been largely neglected until very recently and papers A and B present studies on purely auditory induced vection (AIV). Paper A shows that rotating auditory scenes synthesized using individualized Head-Related Transfer Functions (HRTFs) are more instrumental for presence compared to generic binaural synthesis. Study on translational AIV in paper B shows that inconsistent auditory scene might significantly decrease self-motion responses. Paper C and D demonstrate that bi-sensory stimulations increase presence and self-motion ratings as expected. In paper C additional vibrotactile stimulation increased translational AIV and presence ratings, especially for the stimuli containing the auditory-tactile engine metaphor. Paper D extended paper A results for rotational AIV showing that spatial resolution of rotating auditory scenes can be greatly reduced when combined with visual input.

This thesis shows that sound plays important role in the illusory self-motion perception and it should be carefully used in multi-modal motion simulators. The presented findings suggest that a minimum set of acoustic cues can be sufficient for eliciting a self-motion sensation, especially if other modalities are involved. However, perceptual consistency of the created auditory and multimodal scenes should be assured in the design of the next generation of motion simulators.

Keywords: virtual reality, spatial audio, auditory scene synthesis, cognitive acoustics, multisensory optimization, presence, illusory self-motion, vection.

List of Appended Papers

This thesis is based on the following four publications:

A. Auditory presence, individualized head-related transfer functions and illusory ego-motion in virtual environments

B. Traveling without moving: Auditory scene cues for translational self-motion

C. Vibrotactile enhancement of auditory induced self-motion and presence

D. Sound can compensate for a restricted field-of-view in self-motion simulators

These papers are referenced in the text using their associated letters.

Other Papers

The author has also contributed to following publications related to sound perception, virtual acoustics and presence topics:

Väjamäe, A., Kohlrausch, A., van de Par, S., Västfjäll, D., Larsson, P., & Kleiner, M. (2006). Audio-visual interaction and synergy effects: implications for cross-modal optimization of virtual and mixed reality applications. *Submitted to the Handbook of Presence*

Väljamäe, A., Västjäll, D., Larsson, P., & Kleiner, M. (2006). Perceived sound in mediated environments. *Submitted to the Handbook of Presence*

Larsson, P., Väljamäe, A., Västjäll, D., & Kleiner, M. (2006). Auditory induced presence in mediated environments and related technology. *Submitted to the Handbook of Presence*

Acknowledgements

I started to write the introduction for this thesis from this acknowledgements section. I am sitting on the bus to Pärnu, where my grandmother lives (as you probably saw, this thesis is dedicated to her)[1]. There are many people who helped and encouraged me to enter this wonderful world of Das Glasperlenspiel and I will try to mention some of the key figures.

At first, I want to thank my friend and very good journalist Tamara Kalantar for convincing me to go and study abroad. My trip to Sweden would not happen if Alfred Ots' (1918-1992) scholarship foundation would not exist. I remember reading about this foundation on the advertisements board at Tallinn University of Technology, Tallinn, in 1998 when I was still on the third year of my Bachelor's degree. Of course, I forgot about it and rediscovered it 2 years later while checking the opportunities to visit Iceland in the foreign affairs office at TUT. I would like to thank Kristel Virula, a secretary of the office, for encouraging me to apply for the Alfred Ots' stipend – the stipend, which supported me for 4 years (both my MSc and major part of this Licentiate) and gave me an ultimate opportunity to study what I like. I am indebt to Prof. Em. Mart Mägi, the head of the foundation, for believing in me, for his advices and hospitality (Maxi ja Mart, tänan teid külalislahkuse ja sooja ning sõbraliku suhtumise eest).

Next I want to applaud my dear friends from International Master's programs 2001 – Gerasimos Savramis, Fabien Batejat, Pablo Hernandez and Rodrigo Parra. Each of you indirectly contributed to this work (not only by been an endless source of good mood and encouragement by visiting, calling or sending e-mails). Because of Pablo, I discovered 3D sound topic and attended Prof. Mendel Kleiner's inspiring lecture on sound holography at Applied Acoustics department (which later become a topic of my MSc. thesis). ¡Pablo, gracias por la zamarra! Fabien taught me not to fear working late or even during the night – before I sincerely believed that I can not effectively think after 6 pm. (who knows, maybe it is still just an illusion). I shamelessly copied Rodrigo's idea to become a graduate student instead of desperately looking for a job. And finally, Gerasimos' ancient Greek idea of a time pace helped me to keep a necessary balance between work and personal time (e.g. Kozyrev & Nasonov, 1980).

Surprisingly, having a stipend and a will to pursue a PhD degree is only a half of the story – you need to convince someone to be your supervisor. I was lucky that my master's supervisor, Dr. Thomas Eriksson, believed in my ability to become a researcher and made me a member of the Communication Systems group (thanks everyone in CS, it was an unforgettable experience). I am really grateful for his supervision style which gently taught me that an interesting research topic is not the most crucial part of your studies – it is people working with you, a team, which makes it interesting. I continued to work on the new compression methods for the multichannel audio reproduction but it soon became an exploration trip into several neighbor research disciplines. Studies on the spatial sound quality evaluation methods and the binaural hearing models brought me back to Applied Acoustics department. I should say that most probably I would not be brave enough to continue my studies without masterful supervision of Dr. Daniel Västfjäll. Carefully listening to all my ideas, picking up the best of them, giving further

[1] As the acknowledgments section allows for some entertainment I wanted to comment that in the times of writing my first thesis in Estonian and asking my father Tiit to help me with the language, he always wondered why I liked so much to write in the parentheses. Later, I discovered that I can also insert an additional, often irrelevant information in the footnotes. Tiit, I really tried to minimize both vices in this thesis.

subtle hints – all these nourished my self-confidence as a researcher. Quite often I was developing the research idea and then realizing that almost unconsciously was using the hints from the fruitful discussions with Daniel. It should be noted that both Daniel and Thomas have a great sense of humor, which is a crucial component in graduate student supervision.

Together with Daniel, Dr. Pontus Larsson always treated me as an equal in our joint work on the POEMS project and this was very encouraging for me. During my first conference experience, he taught me the most essential components of the conference life where one should know how to keep a balance between a huge amount of compressed (sometimes even uncompressible) information, new contacts and new places which you may discover at the expense of the conference time.

Ana, everything I will write here would be mundane and, anyway, this is not the write place for writing that. Thank you for coming back to Sweden and forcing me to make this thesis faster ;-). Both you and Dr. Gargantjev helped me to avoid extreme workaholism and made my stay in Sweden a never ending adventure. I also want to thank my friends for all these nice moments during this last 3 years in *Estonia* and in Sweden: - A.T., *K.T.*, N.P., S.R., *T.K.*, *A.T.*, B.S., *G.G.*, G.M., D.P., D.V., F.O., *J.T.*, *K.V.*, M., *O.B.*, P.L., *R.F.*, R.G., *V.P.*

Finally, I want to thank my parents, Lena and Tiit, for their patience and support over all this time (Спасибо за моральную поддержку в трудные минуты). Jag är imponerad av att du lärde dig svenska mycket bättre än mig. I also want to thank my ”swedish mother” Ann-Christine Andersson-Arntén for knowing that I could always ask for her help if needed.

Last but not the list, I want to thank Ana, Andy, Daniel, Mendel, Pontus, Rodrigo and Thomas for reading and correcting my English and making this thesis more coherent and reader friendly.

P.S. *I should definitely agree with other people's opinion about the joy of writing the acknowledgement part – this sensation can be comparable to flying and I am glad that I will hopefully experience that again in a few years.*

This work was partly supported by the European Community under the FET Presence Research Initiative project POEMS (Perceptually Oriented Ego-Motion Simulation), IST-2001-39223 and the Swedish Research Council (VR project 40499601).

Abbreviations and terms

3DS	3-dimentional, spatial Sound
AIV	Auditory Induced Vection
AV	Audio-visual
CCRAG	Chalmers Cognitive Room Acoustics Group
Ecological acoustics	Related to everyday listening experience
EDA	Electro Dermal Activity
FB confusion	Front Back confusion
FOV	Field Of View
HRTFs	Head Related Transfer Functions
IHL	In Head Localization
ILD	Interaural Level Differences
ITD	Interaural Time Differences
KEMAR	Knowles Electronic Manikin for Acoustic Research
LF	Low Frequency
LFMV	LFV and MV presented simultaneously
LFV	LF-induced Vibrations
MV	Mechanically-induced Vibrations
POEMS	Perceptually Oriented Ego-Motion Simulation
PSE	Point of Subjective Equality
SIT	Still Images Train
SVUP	Swedish Viewer-User Presence questionnaire
TOJ	Temporal Order Judgment
VBAP	Vector Based Amplitude Panning
VE	Virtual Environment
Vection	illusory self-motion
VR	Virtual Reality
WFS	Wave Field Synthesis

Contents

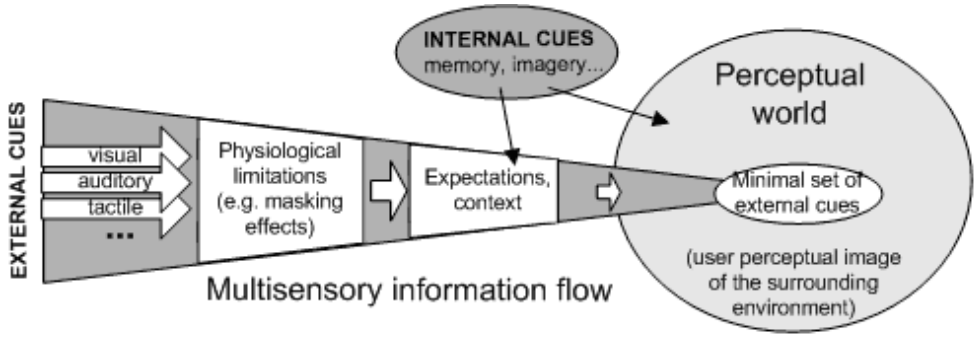
Preface

Being a member of the POEMS[2] project (Perceptually Oriented Ego-Motion Simulation) and working on advanced research topics in the area of an illusory self-motion made me curious about commercially available motion simulators. At the time of writing this thesis I have tried two publicly available attractions. The 3D animation film “Alien Adventure 3D” (Stassen, 1999) projected on an IMAX screen and containing numerous roller-coaster rides, and a sports car ride simulator mounted on a moving platform at FAO Schwarz, the biggest toy shop in New York. What struck me in both of these simulations was a very ad-hoc sound design which sometimes destroyed the sensation of self-motion rather than contributed to it. Sound rendering is a rather inexpensive but often underestimated resource in self-motion simulations if compared to the costs of a large IMAX screens or complex motion platforms with several degrees of freedom. So how beneficial is it to add sound information in self-motion simulators, and what auditory cues would be the most instrumental for creating an illusory self-motion sensation? An even more important question is how various sensory inputs should be optimally combined in order to produce additive rather than conflicting multisensory effects. The research presented in this thesis addresses these questions and, hopefully, provide some answers.

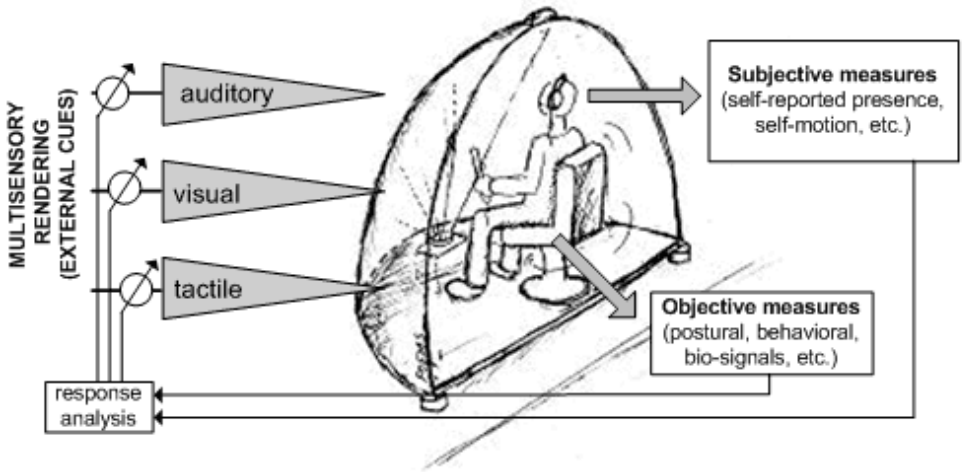
[2] The POEMS project (IST-2001-39223) has been funded through the Future and Emerging Technologies (FET) Presence Research initiative.

1 Introduction

In the last decades, research on multisensory interaction enjoyed growing attention from various disciplines. However, it is still an ongoing process of communicating the recent findings from cross-modal perception into research areas such as Virtual Reality (VR) and multimedia, where rendering of different sensory inputs has been traditionally optimized separately. On the contrary, in the new, human-centered design of multimodal media one should rather concentrate on perceptual dimensions of space and time, optimizing them across all contributing sensory inputs. In such a design, the end users' percepts of self-motion and presence should serve as the evaluation criteria of immersive multisensory displays.



(A) Virtual environment perception as product of external and internal cues



(B) Human-centered experimental loop for multisensory optimization of virtual environment (self-motion simulator case)

Figure 1: *A) A minimal set of multi-modal cues can be sufficient for creating perceptual image of surrounding environment. B): multi-modal rendering can optimization can be based on subjective (this thesis) or objective measures (future motion simulator sketch by Pontus Larsson).*

Figure 1 represents the idea of perceptual optimization in rendering of multi-modal environments. It is well known that our perception of the external world depends on various perceptual and cognitive factors (Figure 1A). For example, your sound localization can be biased by your vision (the ventriloquism effect, see section 3) or presented environment context (cf. elevation judgments bias of naturalistic sounds reported by Hughes et al. (2004)). Therefore it is likely that a reduced amount of sensory stimulation could be sufficient to produce a surrounding environment which might be perceptually identical to the real world. In order to find this minimum set of multi-modal cues one should examine end users' percepts in rendered Virtual Environments (VEs) using subjective or objective measures, e.g. verbal responses, psychophysical measures etc. (Figure 1B). For example, user presence sensation or "presence response" to virtual environments is believed to serve as such an evaluation criterion, and a growing body of research tackles the perceptual roots of presence (for a recent review see (Sanchez-Vives & Slater, 2005)). In this thesis, subjective sensations of self-motion and presence served as a basis for psychophysical studies (4 studies, 86 participants) on cross-modal optimization of self-motion simulators.

This thesis consists of an introductory part followed by 4 papers (papers A-D) serving as the main body of the presented research. In **Section 2**, the notions of illusory self-motion and presence are first specified from the traditional unisensory perspective. **Section 3** provides a brief overview of recent findings in audio-visual interaction where the concepts of perceptual dimensions of time and space are introduced. Possible implications for new perceptually optimized design of multi-modal environments are then discussed and exemplified (section 3.4). At this point, studied research questions can be defined thus creating **Section 4**. Next, the methodology used throughout the presented work (papers A-D) is described in **Section 5**. The summaries of the results from the individual papers in **Section 6** are followed by the concluding **Section 7** where interconnections between the papers are highlighted. Possible future research topics are discussed in **Section 8**.

2 Experiencing self-motion and presence

We can hardly imagine our everyday life experience without the ability to move around and actively explore the surrounding environment. This is also true for virtual environments where the ongoing research on self-motion simulators occupies an important niche in VR – related studies. Similar to other VR applications, designers and engineers aim at the high multi-modal rendering fidelity in motion simulators and new technical innovations gradually reduce the gap between simulation and reality. However, it is important to have the ability to access the end-user sensations during motion simulation as we know that often perceptual rather than pictorial realism matters in VR (e.g. Slater, 2002). In this human centered approach, we are particularly interested in self-motion and presence sensations which are elicited by the simulator. Accessing these user experiences should help to validate the effectiveness of existing technologies and further optimize motion simulators design.

2.1 Spatial sound rendering in virtual environments

Since the invention of the phonograph in 1877, sound recording and reproduction techniques have been continuously evolving. The aim of spatial sound rendering is to create an impression of a sound environment surrounding a listener in 3D space, thus simulating auditory reality. This goal has been assigned many different terms including *auralization*, *spatialized sound* or *3-D sound (3DS)*, and *virtual acoustics*, which have been used interchangeably in the literature to refer to the creation of virtual listening experiences. For example, the term auralization was defined by Kleiner et al. (1993) as "*. . . the process of rendering audible, by physical or mathematical modeling, the sound field of a source in a space, in such way as to simulate the binaural listening experience at a given position in the modeled space*".

One common criterion for technological systems delivering spatial audio is the level of perceptual accuracy of the rendered auditory environment, which may be very different depending on application needs (e.g. VR simulator or videoconferencing). Apart from the qualitative measure of perceptual accuracy, spatial audio systems can be divided into *head-related* and *soundfield-related* methods. Interested readers can find detailed information on these two approaches in the books by Begault (1994) and Rumsey (2001) respectively (see also a recent review by Shilling & Shinn-Cunningham, 2002 and new book edited by Blauert (2005)). The following subsections will briefly describe current state-of-the-art technology for reproduction of spatial audio used in VE environments.

2.1.1 Soundfields reproduction

Soundfield-related or multichannel audio reproduction systems can give a natural spatial impression over a certain listening area – the so called *sweet spot* area. The size of this sweet spot area mainly depends on the number of audio channels used. At the present time, 5-channel, often referred to as *surround sound* systems, have become a part of many audio-visual technologies and standards in digital broadcasting and in the cinema domain. Next generations of multichannel audio rendering systems are likely to have a larger number of channels providing better spatial sound quality like 10.2-channel system (Zimmermann et al., 2004), Vector Based Amplitude Panning - VBAP (Pulkki, 1997), Ambisonic (Gerzon, 1985), or Wave Field Synthesis - WFS (Berkhout, 1988, Horbach et al., 2002). WFS can create a correct spatial impression over an entire listening area by using large loudspeaker arrays (typically >100 channels). However, direct recording/transmission of spatial audio using WFS principles is difficult and requires novel multichannel audio compression methods (see Väljamäe, 2003 for a review). Currently the WFS concept has been coupled with object-based rendering principles, where the desired soundfield is synthesized at the receiver side from separate signal inputs representing the sound objects and data representing room acoustics (Horbach & Boone, 1999).

2.1.2 Binaural synthesis

Head-related audio reproduction systems, also referenced as binaural or 3DS systems, are based on special pre-filtering of sound signals imitating mainly the outer ears (the pinnae) effects. Pre-measured catalogues of Head-Related Transfer Functions (HRTFs) are used for binaural sound synthesis, where a non-spatialized (“dry”) sound is convolved with transfer functions corresponding to the desired spatial position of the source. Currently, most 3DS rendering systems use generic HRTFs catalogues due to the lengthy procedure of recording listeners’ own HRTFs, however, individualized catalogues are proven to enhance presence (Paper A).

When generic HRTFs are used, the most common problem is in-head localization (IHL), meaning that sound sources are not externalized but rather perceived as being inside the listener’s head (Blauert, 1997). Another known artefact is the high rate of reversals in perception of spatial positions of the virtual sources as binaural localization cues can be ambiguous (cone of confusion), e.g. front-back confusion (Begault, 1994). Errors in elevation judgments can be also observed for stimuli processed with non-individualized HRTFs (Wenzel et al., 1993). These problems are believed to be reduced when head-tracking and individualized HRTFs are used (Blauert, 1997). At present time it is popular to use anthropometric data

(pinnae and head measurements) for choosing "personalized" HRTFs from a database containing HRTFs catalogues from several individuals (Zotkin, 2004). However, as the auditory system exhibits profound plasticity in the spatial hearing mechanisms, a person can usually adapt to localize sounds with some generic HRTFs catalogue. One could see these processes as re-learning to hear with modified pinnae as was shown in Hofman et al. (1998). This natural ability to adapt to new HRTFs catalogues might be used when specifically modified, "supernormal" as termed by Durlach et al. (1993), transfer functions are introduced in order to enhance localization performance, e.g. to reduce front-back confusions (Gupta et. al., 2002).

Generally, binaural systems are used for sound reproduction over headphones, which make these techniques very attractive for wearable augmented reality applications (see the excellent review by Härmä et al., 2004). However, binaural sound can be also reproduced by a pair of loudspeakers if additional processing is applied (*cross-talk cancellation* technique), which sometimes is used in teleconferencing (Evans et al., 1997).

2.2 Illusory self-motion

What is an illusory self-motion sensation? Imagine you sitting on the train and waiting for its departure. From the window, you see a train on the neighboring track. Finally, the scene starts to move – your journey has begun. Suddenly you realize that it is the neighboring train which departed instead. This miss-perception (illusion) can be even stronger if your train's engine is already switched on and the seat is slightly vibrating. Self-motion illusions can occur in similar situations on a bus, a car or a ferry – typically in the cases where a visual scene you believed to be stationary suddenly starts to move, thus fooling your sense of space.

The first studies on self-motion and illusory self-motion (often referred to as vection) in the 19th century involved specific machinery where participants or various stimuli could be physically moved (typically rotated). As professor Viktor Urbantschitsch noted in his work: "self-motion sensation is so slight that one should pay special attention to notice it and a specific methodology should be used in the experiments" (Urbantschitsch, 1897, p. 236). Nowadays, illusory self-motion effects can be reliably elicited in laboratory conditions. Usually, a rotating visual stimulus with simple geometrical patterns is presented to the participant sitting in the center of a rotating optokinetic drum or observing a projection screen having a large field of view (FOV). In such experiments observers first perceive motion of the visual stimulus (object motion) during the first few seconds. After some time, the perceived object motion unwillingly merges into a strong self-motion sensation where the moving visual stimulus is now perceived as a result of this self-motion, such as a scene

observed from the window of a moving vehicle. Effects of self-motion illusion for rotation, translation and translation in the vertical plane have been studied for more than a century and extensive reviews can be found in the literature (Andersen, 1986; Warren & Wertheim, 1990).

While a large body of research has been focused on vection elicited by visual stimuli, research on purely auditory-induced vection (AIV) has received little attention in the past, and only in recent years this area of research experienced substantial growth (see the introduction in paper B for references). Although the renewed interest in auditory-induced vection is supported by multi-sensory research topics, this phenomenon can be also interesting for applications for visually impaired people.

Auditory induced self-motion can be elicited using moving sound fields, either real (e.g. loudspeaker array presentation) or virtual ones (typically headphone reproduction). Binaural technology where non-spatialized ("dry") sounds are convolved with pre-measured HRTFs of the corresponding spatial positions (see section 2.1.2) provides the most flexible and cost-effective way of creating sound environments for self-motion related research.

2.3 Presence

Unlike the research on illusory self-motion, presence research is a relatively new but rapidly growing area. The feeling of presence is often described as a sensation of "being there", inside rendered VE, whereas several other definitions exists (see seminal review by Lombard & Ditton, 1997). Being widely explored on the subjective level using questionnaires, presence is believed to have neurophysiological correlates and the ongoing research has to provide further insights to this question (Sanchez-Vives & Slater, 2005). As mentioned earlier, accessing user's presence sensation or presence response within rendered multi-modal virtual environments is usually believed to be crucial for human centred evaluation (Lombard & Ditton, 1997) and further optimization of media content, including VR applications (e.g. self-motion simulators).

Virtual reality development and related presence studies have been long time visually-dominant but at the present time more multisensory approaches start to emerge. Studies on sound influence on users' presence response, or auditory presence research, has been significantly smaller than corresponding visual studies (for extensive recent review on auditory presence research see Larsson et al. (2006)). Auditory presence research questions involve both spatial sound rendering issues and other sonic environment aspects, for example sounds contributing to one's self-representation in VR (Väljamäe et al., 2005).

3 Multimodal experience[3]

In the previous section the experience of self-motion and presence was described from the "unimodal" perspective – e.g. visually induced or auditory induced. However, our perception processes are multimodal in nature, even in the cases where a certain sensory input acts as a main information carrier. For example, visual input appears to be an important component for speech intelligibility and can lead both to its enhancement or degradation (in the case of delay between auditory and visual streams or improper recreation of audio-visual speech of virtual characters). The roots of multimodal perception mechanisms can be traced to early stages of human development and are shaped by the surrounding environment where multisensory events and objects tend to dominate. This section aims to give a brief overview of recent findings from the rapidly growing area of multisensory perception research in relation to VR design.

Creating an immersive Virtual Environment (VE) resulting in high levels of presence is a challenging task, and cross-modal stimulation is an important tool for achieving this goal (Durlach & Mavor, 1995). The visual modality had been often dominant in VR technologies, where high visual fidelity of rendered objects was desired. When designing a multimodal VE, a common approach is to add other sensory input on top of the existing visual rendering technology. It is tempting to extending the logic of unisensory rendering to multisensory displays, thinking that a higher resolution (sensory depth) for all modalities will assure the best results. Although such an approach is absolutely necessary for some VR applications where physically correct rendering is crucial (e.g. basic research on perception), there are several constraints to it. Apart from the computational load, a VE providing detailed information for each modality may overload the user (the cognitive capacity hypothesis; Västfjäll et al., 2005). Moreover, recent findings suggest that the realism of VE's may not be as important as a set of necessary "minimal cues" for creating a sense of presence (Slater, 2002). A search for optimal cross-modal combinations in VE rendering will contribute to our growing knowledge about this "minimal cues" set.

Advances in compression algorithms for visual and auditory information serve as a good example how our knowledge on underlying perception mechanisms can be used for media processing. Data storage, transmission and rendering can be optimized without the user being aware of changes in the initial material. Cross-modal perceptual optimization is thus a logical step in the development of multi-media and VR technologies. Recent results from multisensory perception research suggest that the surrounding environment dimensions, such as space and time, rather than

[3] This section contains parts from the chapter draft by Väljamäe, A., Kohlrausch, A., van de Par, S., Västfjäll, D., Larsson, P., & Kleiner, M. (Väljamäe et al. 2006a)

individual sensory inputs are used as critical modules in our perception of the world (Guttman et al., 2005). Moreover, the user in a multi-modal VE has a strong, ecologically motivated, tendency to attribute inputs from different modalities either to single (unitary) event or to separate objects and events (Welch, 1999). Therefore, for a future VE designer it might be beneficial to think in the categories of created environment, objects and events, desired affective responses and search for optimal combinations of perceptually salient modalities thus assuring temporal and spatial congruency of this new unitary perceptual world.

The idea of audio-visual optimization is not new and has been widely used in cinema history, where cross-modal unity effects compensated for imperfections in a specific modality. For example, in the George Lucas film *The Empire Strikes Back* a sound effect of an automatic spaceship door opening was often used to compensate for a missing visual event – what viewers actually saw was successive static shots of first the open and then the closed door (Chion, 1994). Correspondingly, a spatially discrepant sound can be perceived as originating from the same position as the associated visual object - this effect is known as the ventriloquist effect or ventriloquism, referring to the sensation of watching a skilled ventriloquist agitating the lips of a dummy in synchrony with the speech he/she produces, preferably with as little visible articulation as possible (Witkin, 1953; recent review in Vroomen & de Gelder, 2004a). Although there is a spatial discrepancy between sound and the image, we perceive the voice as if it was originating from the actor's mouth. The ventriloquist effect makes feasible rendering of the wide auditory space in front of the viewers, where typically three frontal loudspeakers work for visually continuous space on the cinema screen.

3.1 Spatial unity and disparity

Visual information usually dominates proprioceptive or auditory inputs in the perception of spatial parameters (Welch & Warren, 1980) as demonstrates the ventriloquism effect described above. This apparent dominance of one sensory modality over another might be a result of differences in the suitability of the different modalities for perceptually coding of a certain stimulus feature. According to the modality appropriateness hypothesis, the sensory modality which is more suitable, usually established by the accuracy of coding of this stimulus feature, is assumed to be the dominant sensory modality (Welch 1999; Welch & Warren, 1980). This holds for the case of the ventriloquism, where spatial acuity is about 1 min of arc or 1/60 degrees for vision (Howard, 1982) and 1 degree for hearing (Mills, 1958).

What happens when the visual input is distorted and stimuli cannot be easily localized - can the auditory system gain weight in this situation? A

recent study on the ventriloquist effect by Alais and Burr (2004) used severely blurred visual stimuli (object extension to 60 degrees) and found that in this situation sound captured vision. By varying the level of visual distortion they showed that localization could be modeled as a bimodal integration, were weights of the auditory and visual modalities are inversely proportional to the localization variance. In the experiment, a limited number of auditory spatial cues were used and the authors suggested that with a better spatial audio rendering quality, less blurring of the visual object (down to 10 degrees) would be sufficient for allowing sound to capture vision. A more complex statistical model for modality integration was applied by Battaglia et al. (2003) to explain results from a study on visual bias for the localization in depth, but it should be noted that auditory and visual distance perception mechanisms differ significantly from the ones applied for horizontal localization (Blauert, 1997).

Auditory perception of distances is known to be rather imprecise. As summarized in a recent review article by Zahorik et al. (2005), humans typically overestimate the distance of sources with a distance below about 1 meter. For more distant sources, the perceived distance typically is an underestimation of the physical distance (Bronkhorst & Houtgast, 1999; Nielsen 1991). Due to this inaccuracy in auditory perception of distance, it might be expected that, similar to the ventriloquism effect for direction disparities, visual information also influences perceived auditory distance. Such an influence of visual information on spatial perception might be of relevance when judging room acoustic properties, and it might also influence the way how 3D audio and video are reproduced, such as in stereoscopic television, 3-D TV (Västfjäll, Larsson & Kleiner, 2005).

The scarce literature about the visual influence on auditory distance perception reveals different results regarding across-modal effects. The study by (Nathanail et al., 1997) used 3-D visual projection and natural room acoustics providing sufficient auditory cues for distance perception. The authors found a visually induced bias of the perceived depth, but this required a certain training of the participants. Another study, based on naturalistic audio-visual stimuli, failed to observe visually induced shift in distance judgments, but found an increased accuracy of distance judgments (Zahorik, 2001). Interestingly, a visual stimulus can successfully bias judgments on auditory distance if it is placed near the observer at the reachable distance (Brown et al., 1998). Inconsistencies in the results on audio-visual interaction might arise due to the different strategies for encoding information about distant or near, within-the-reach space (Grazziano et al., 1999). A growing body of research from neurophysiological, cognitive and functional imaging studies provides evidence of multisensory representation of near-peripersonal space (see Ladavas and Farne (2004) and references therein).

It is important to note that intersensory bias effects, the effects where a stimulus in one modality affects perception of a stimulus in another modality, can be influenced by pre-cognitive factors - a number of amodal stimulus properties shared by two modalities (Welsh & Warren, 1980). Common spatial location, temporal patterning or rate, size, shape, orientation, intensity, motion and texture can contribute to the percept of a single object or event, which has been termed as the “unity assumption” (Welsh, 1999) or “pairing” (Bertelson & de Gelder, 2004). Using stimuli with weak, mild, or strong support for the unity assumption in an experiment on the ventriloquist effect (Welsh & Warren, 1980), found a gradual increase in visual bias as the number of amodal properties increased. Pairing of audio-visual stimuli can also occur for “ecologically valid” combinations, such as in the case of approaching or receding object simulation. Several studies confirm that presentation of specifically paired visual and auditory stimuli (e.g. increasing object size coupled with increasing sound intensity) leads to a percept of a unitary audio-visual object, which reflects in stronger after-effects compared to effects from unisensory stimulation (Zetzsche et al. 2002, Kitagawa & Ichibara, 2002). It is important to mention the asymmetry between the approaching and receding objects perception, where the former are proven to be more biologically salient and leading to stronger perceptual effects (judgments on loudness, distance traveled) for both auditory only (Hall & Moore, 2003) and audio-visual stimuli (Mayer et al., 2004).

3.2 Temporal unity and disparity

In the vein of the modality appropriateness hypothesis, the auditory system is better suited for processing of temporal information and therefore shows a dominating role for audio-visual interaction effects in the time domain. For example, visual “flicker” fusion occurs at rates above about 50 to 100 Hz (Landis, 1954; van der Zee & van der Meulen, 1982). At the same time, the auditory system is able to detect amplitude modulations for rates of up to 600 Hz, if the carrier is a high-frequency sinusoid (e.g., Kohlrausch et al., 2000). This implies that the ability to represent and perceive fast variations in stimuli is better developed in the auditory system than in the visual system, and that in an AV context the auditory system may be dominant in determining rate perception. Such dominance is sometimes referred to as temporal ventriloquism, a term encompassing several different effects described in the text below.

A temporal interaction effect known since several decades is the mutual influence in perceived rate between visual “flicker” and auditory “flutter”. This interaction is revealed if subjects are asked to match the rates of regularly-modulated stimuli within and between the auditory and visual modalities (e.g., Gebhard & Mowbray, 1959). The authors of this study

found that the precision of rate matching was much lower for across-modality than for within-modality matches. When the rates of variation in the visual and the auditory stimulus were different, a bias of the auditory rate on the perceived visual rate was observed, while the opposite, visual rate influencing auditory rate was not found. Apart from the auditory bias of visual stimuli presented at a certain rate (see Recanzone (2003) for recent review), the same can be observed for a single audio-visual stimulus, where sound can sharpen the temporal boundaries of a visual event (Vroomen & de Gelder, 2004b).

Another case of the temporal ventriloquism can be observed when performing a visual temporal order judgment (TOJ) task (Morein-Zamir et al., 2003). In their study, performance in a TOJ task significantly improved when two visual flashes were temporally "framed" by two sounds presented before and after the visual stimuli. More importantly, this effect could be observed only in a certain time interval between the auditory beep and the corresponding visual flash (approximately 200 ms) after which TOJ performance started to decrease. This time interval can be termed an *intermodal temporal contiguity window* during which stimuli in different modalities are perceived as simultaneous even if they are not precisely aligned in time (Lewkowicz, 1999).

The idea of a temporal contiguity window is very important for multi-media and VR applications, where asynchrony between modalities can destroy the congruence of the perceptual world. For example, spatial audio-visual interaction in the ventriloquism effect decreases if sufficient delay (200 ms and higher) between the auditory and visual temporal patterns is introduced (Radeau & Bertelson, 1977; Warren et al., 1982, Wallace et al., 2004). AV synchrony has mainly been studied with respect to two perceptual aspects, i.e. the detectability of asynchrony and the effect of AV asynchrony on perceived quality (for a recent review see Kohlrausch and van de Par (2005) and references therein). The tolerance to AV asynchrony varies depending on the stimuli type used, with highest asynchrony tolerance for speech, then for single audio-visual stimuli pairs and lowest for repetitive stimuli (Miner & Caudell, 1998). It has to be noted that more systematic research has to be done before drawing final conclusions on the underlining mechanisms of this phenomenon.

Interestingly, the audio-visual temporal contiguity window is asymmetric. This asymmetry can be expressed by means of the point of subjective equality (PSE), which is computed as a mean between asynchrony detection thresholds for auditory lead and lag relative to visual stimuli. PSE occurs in most studies for a positive delay of the auditory stimulus of about 30 to 50 ms relative to the visual stimulus (Kohlrausch & van de Par, 2005). The discussion of the reason for this observation focuses in nearly all studies on the physical delay in the audio signal caused by the rather low speed of sound of 340 m/s. If the perceptual system was adjusted

to perceive AV objects from a certain distance, the reference point for AV synchrony should not be the head of the observer, but the position of the object. Since the perception experiments reveal a synchrony offset of 30 to 50 ms, the distance of this reference point should be about 10 to 15 m. It should be noted that recent studies on adults show that the temporal contiguity window can be changed if participants are exposed to a constant audio-visual delay prior to the main experiment, thus suggesting that experience might play important role in the process of recalibration of our perceptual sensitivity (Vroomen et al., 2004, Fujisaki et al., 2004).

Guttman et. al. (2005) studied a phenomenon called cross-modal encoding of visual temporal structure, where the authors suggest a process of unimodal temporal rhythm registration and memorizing based on auditory code. In their initial experiment, visual stimuli consisting of striped patches with temporally changing contrast were presented in conjunction with sound clicks, matched or mismatched with the visual rhythm. In addition, a condition with no sound was included. The participants' task was to judge whether stimuli in two successive trials had the same visual temporal pattern or not. It was found that mismatched auditory information significantly reduced participants' performance in the same/different judgment compared to trials with no sound. Moreover, a matched auditory stimulus significantly increased participants' performance in the task compared to both no-sound and mismatched sound trials. Further experiments with the same experimental design (Guttman et al., 2005), confirmed that 1) incongruent auditory information, even being task-irrelevant, significantly impaired visual rhythm memory, and 2) observed cross-modal interference disrupts the encoding but not the retrieval of visual rhythm.

Results from the experiments on the cross-modal encoding of rhythm can have an important implication on our understanding how the illusion of unitary perceptual world is created in the brain. Guttman et al. (2005) suggest that general temporal processing and representation is auditory in its nature. The reverse process can be seen for auditory space encoding where the unified audio-visual space encoding is driven by the visual modality (Knudsen & Brainard 1995, Knudsen 2002, Zwiers et al., 2003). This assigned sensory-dependent encoding of more general space and time information might be explained from the perspective of the modality appropriateness hypothesis (Welsh, 1999), where auditory or visual encoding mechanisms were evolutionary chosen because of their performance accuracy in spatial or temporal tasks. It is interesting that the same line of thought about audio-visual information appears in the area of film theory, where Chion (1994) insightfully suggests that spatial components in a film are encoded as "visual impression" and temporal components as "auditory impression"

3.3 High-level processing effects on cross-modal interaction

Research advances in the area of cross-modal perception development provides a growing evidence that higher order processes (e.g. attention, motivation) may modulate multisensory processing and supports an integrative view of human perception (see Wallace (2004) and references therein). Cross-modal perception relies on multisensory neurons, whose responses to multimodal stimuli largely differ from being stimulated by single modality stimuli (Meredith and Stein, 1993). These neurons responses can be either enhanced (superadditivity) or depressed depending on whether the presented multimodal stimulus pair is spatially or/and temporally synchronous or not.

A striking demonstration of how the higher order processes can influence the lower perceptual processes was given by the connection recently found between multisensory neurons in the cerebral cortex (related to "association" functions) and the superior colliculus (midbrain area involved in lower-level functions like motor control of orientation behaviors). Artificially induced deactivation of the cerebral cortex area was found to deactivate multisensory neurons in the midbrain area, which in turn was reflected in inadequate behavioral responses of test species (Wallace, 2004).

Recent neurophysiological findings reviewed in (Wallace, 2004) are in line with the growing research evidence suggesting an important role of experience in shaping the information processing in our brain (e.g. Lewkowicz, 1999, Vroomen et al., 2004). These processes start already in the prenatal period where modified sensory stimulation during this time results in changes in perception and behavior after birth (Lickliter & Bahrick, 2000 and references therein). It is important to note that this experience can also be acquired without awareness where frequent appearance of certain features in the environment is equated by our brain with ecological salience (Watanabe et al., 2001). The insightful conclusion by Watanabe and his colleagues warns us that these natural adaptive mechanisms of our brain can be unprepared for new environments created by the "manipulative modern-day media".

3.4 Cross-modal optimization of VR and media applications

Real-time rendering of highly immersive multi-modal environments is a computationally demanding task, and reducing the computational load and the amount of presented information can be beneficial for many applications. For example, in the case of distributed VE applications, reducing irrelevant multisensory information can be highly advantageous in the light of transmission bandwidth limitations or latency restrictions. Although various perceptual coding algorithms are widely used in audio and

video formats, the possibilities of cross-modal perceptual optimization have received little attention until recently (Kankanhalli et al., 2004) and have been mostly explored in an intuitive way. For example, the ventriloquism effect supports the illusion of wide spatial sound image in front of the cinema viewers when only three loudspeakers are typically used[4]. The growing knowledge on cross-modal interaction and synergy effects reviewed in this chapter can allow for a systematic multisensory optimization, where the resolution of a rendered VE can be adjustable either within or across sensory inputs. Coming back to the cinema example, a sufficiently high spatial sound resolution[5] should be allocated for the off-screen sound (outside the viewer's field of view), where no support from visual localization is available. On the contrary, visible objects can be accompanied with a lower resolution spatial sound rendering. In general, the new algorithms for multi-modal rendering optimization should aim for preserving the unity of the perceived environment, objects and events, therefore preserving only the optimal combinations of perceptually salient modalities. In a similar vein, sound dominance in the temporal domain can be efficiently used when representing dynamics of multimodal VE, which is often more important than visual realism of moving objects as pointed out by Luciani (2004).

The attention of end user is likely to play a very important role in the perceptual optimization of the displayed information. Selective rendering of the visual scenery is a growing area of research, where various techniques are used for tracking or predicting the user's attentional focus at a particular moment in time (see Cater, 2004, O'Sullivan et al., 2004 and references therein). Typically, the eye-tracking techniques (Baudisch et al, 2003), continuously monitoring user's gaze are used for allocating the highest rendering resolution in display. However, systems based on monitoring of the brain activity related to spatial attention (Müller et al., 1998) can emerge in the areas, where brain-computer interfaces are already used for user interaction in a VE (e.g. Friedman et al., 2004; for review see Pfurtscheller et al., 2005).

The perceptual optimization of visual displays based on visual-attention models is also a rapidly growing area of research where the perceptually salient scene regions are predicted and selectively rendered. However, the user's attention can be also guided by the non-uniform, *foveated* display resolution (Itti & Koch, 2001). Following this line of research, one could expect that the cross-modal attention models can be used in the future optimization of multi-modal environments.

---

[4] It is difficult to preserve characteristics of spatial image over a wide area with using common sound panning techniques – reason why third loudspeaker was introduced instead of stereophonic setup in cinema (Rumsey, 2001).

[5] Spatial sound localization accuracy varies depending on the sound source position being highest (1 deg) in front of the listener in the horizontal plane (Blauert, 1997).

The perceptual optimization of multi-modal information can benefit from computer graphics optimization research where attention processes in the human visual system, such as *inattentional blindness* and *change blindness*, have recently been studied (Cater, 2004). In the case of inattentional blindness, the visual scene objects, which are not attended by the viewer, are not actually seen by them (Mark & Rock, 1998). In specific task-related scenarios the effect of inattentional blindness can be used for selective visual scene rendering (Cater, 2004). Change blindness is defined as the viewer's inability to detect large changes between two images presented consequently if some sudden interruption (e.g. brief flicker, film cut or eye blink) occurred between them (Rensink, 1997). The effect of change blindness cannot be easily applied to the selective rendering of the continuous visual stream; however, it might be beneficial for situations where consequent still images are used in a slide show fashion. A classic cinematographic example where successive still images create a coherent perceptual world is the photo-novel *La Jetée* by Chris Marker (1962). An accompanying voice of a narrator keeps together the visuals and guides the viewer through the story. Another example can be provided by Armenian director Sergei Paradjanov's films, where still images or long still shots serve transforms motion picture into motion painting (see Holloway, 1994 and references therein).

Representing a scene by series of still images instead of a continuous video stream becomes a more and more common technique in music video clips and advertisements, where fast-editing techniques are typically used to catch the user's attention (Fahlenbrach, 2002). However, presenting an audio-visual stream as a sequence of still images can be seen also as a way of reducing the cognitive load for an end-user of multisensory displays. In a way, such "slide show"-like presentation of a visual stream resembles the common technique in visual and multi-modal information retrieval where only key video frames are used (Snoeg & Worring, 2002).

An interesting question is to what extent visual information can be compensated by other modalities in new media applications. For example, a video stream can be divided into a sequence of still images (*still images train - SIT*) combined with a continuous binaural soundtrack. In this case spatial sound might contribute to the recreation of continuous perceptual world compensating for the reduced visual input. In the cinema, for example, visual editing effects are often masked by the continuity of an accompanying sound background therefore "greezing the cut", as an editor would call it (Eidsvik, 2005). Currently, our group investigates the feasibility of such reduction of audio-visual material in presented VE (Väljamäe et al., in preparation). It should be noted that the reduction of the visual information to SIT format can be of great interest for applications having strict bandwidth or storage capacity limitations. It should be noted that the SIT idea is a logical extension of new emerging audio-visual

applications where combination of single still photographs with binaural sound is used (Hoisko, 2003; Bitton & Agamanolis, 2004).

An interesting initial attempt for audio-visual optimization in the temporal domain has been conducted by Mastoropoulou and Chalmers (2004), where the effects of music tempo on the frame rates of the accompanying visual sequence were investigated. Although no significant effects were found, slow music showed a marginal decrease in the participant's discrimination performance between video sequences with frame-rates of 12 and 16 frames per second. One possible explanation for the insignificant results of this experiment is the stimulus selection where arbitrary auditory and visual rhythms have been used instead of consistent audio-visual pairs. One could argue that the temporal cross-modal optimization might be feasible in the situations when unitary audio-visual objects or events are present (Kubovy & Valkenburg 2001, see also Väljamäe et al. (2006b) and references therein). The perception of such unitary audio-visual pairs might involve mechanisms of time-varying attention, similar to the model of *attention rhythm* described by Jones (2004). Furthermore, in the case of the still image train technique described above, a variable, attention-optimized frame rate rather than a constant frame rate might lead to an expected compensation of reduced visual information by the auditory modality.

Apart from reducing the perceptual irrelevance of cross-modal information one could think of another type of optimization – namely affective optimization. The superadditivity of multisensory neurons described above might lead to a stronger sensory stimulation and new perceptual experiences compared to the situations were unsynchronized auditory and visual information is presented. Interestingly, in the experiments by Guttman et al. (2005) on interaction between auditory and visual rhythms (see section 3.2), several participants perceived a complex audio-visual rhythmic structure while being aware of separate information presented in auditory and visual modalities. Though anecdotal, further evidence for that can be felt while watching the animation films by Walt Disney (e.g. Fantasia), where music was directly used as a reference for the animators' work (Frank & Johnston, 1981), or the abstract films by Len Lye (e.g. Colour Flight, Rhythm, Free Radicals and others, Len Lye filmography, 2005) where he was visualizing the music rhythm by painting or scratching directly on celluloid. Tight audio-visual synchronization leads to stronger emotional responses and this fact has been successfully exploited in music video-clips (Fahlenbrach, 2002). To conclude, one can predict that finding the optimal cross-modal pairings across modalities should lead both to more effective and affective ways of presenting the audio-visual and other multisensory information.

3.5 Multisensory optimization: summary

Systematic optimization the VR and multimedia applications is difficult without better understanding of underlying perceptual mechanisms of cross-modal interaction and synergy effects. The results from multisensory research reviewed in this chapter suggest that for a future VE designer it might be beneficial to think in the amodal categories of the unitary space, time, desired affective responses etc. Knowledge about the modality-specific properties of our perception should help him/her in the search for optimal combinations of sensory inputs salient for a specific task (e.g. auditory alerts in the air traffic control (Cabrera et. al, 2005). In the case of VEs, temporal and spatial congruency of the simulated unitary perceptual world is a crucial component for user's overall presence response to this new environment.

4 Studied research questions

This thesis addresses several research questions (RQs) which might be essential for the successful, perceptually optimized design of the next generation self-motion simulators.

RQ1 – *Contribution of auditory cues to self-motion simulation*. It is usually assumed that increasing the number of sensory inputs should result in stronger VE experience, in our case, higher self-motion sensation. However, it is interesting to evaluate how efficient is purely auditory stimulation in eliciting self-motion sensation, and how much one can gain by adding proper auditory scene rendering in multi-modal motion simulations.

RQ2 – *Minimum auditory cues for self-motion simulating*. Current virtual environments often provide satisfactory spatial sound rendering, either via speakers or headphones. As described above, binaural synthesis is better fitted for accurate rendering of sound scenes in VR applications. In self-motion simulation applications, binaural synthesis of moving soundfields is a computationally demanding task, where various factors have to be taken into account, e.g. spatial resolution of the HRTFs catalogue, room acoustics modelling, sound source characteristics, etc. Therefore, finding the auditory cues which are most instrumental in inducing self-motion is an important step towards perceptually optimized VR motion simulators.

RQ3 – *Perceptual scene consistency*. It is necessary to assure that a created virtual or augmented environment is perceived as perceptually coherent by the end user. For example, one should ask how important the temporal and spatial consistency is for eliciting presence and self-motion responses (see section 3.4 for general discussion). Perceived consistency of the created uni- or multimodal scenes can be also dependant on contextual information and users' expectations.

5 Method considerations

All experiments reported in this thesis share a similar methodology, and papers B, C and D have been conducted in the setup built by engineers from Max Planck Institute of Biological Cybernetics, Tübingen, Germany (see Figure 2). This setup replicates the apparatus in Germany where it is mounted on top of a 6 degree-of-freedom motion platform, which allows studies on vestibular cues for self-motion. Studies at Chalmers did not involve physical motion, and only illusory self-motion effects have been investigated in auditory, visual and tactile domains.

5.1 Apparatus

The experimental allows choosing between a flat or curved projection screen (2 m curvature radius), which participants viewed from 1.7-1.8 m distance. Black curtains cover the side and top of the cabin surrounding the projection screen in order to increase immersion and block visibility of the outside room.

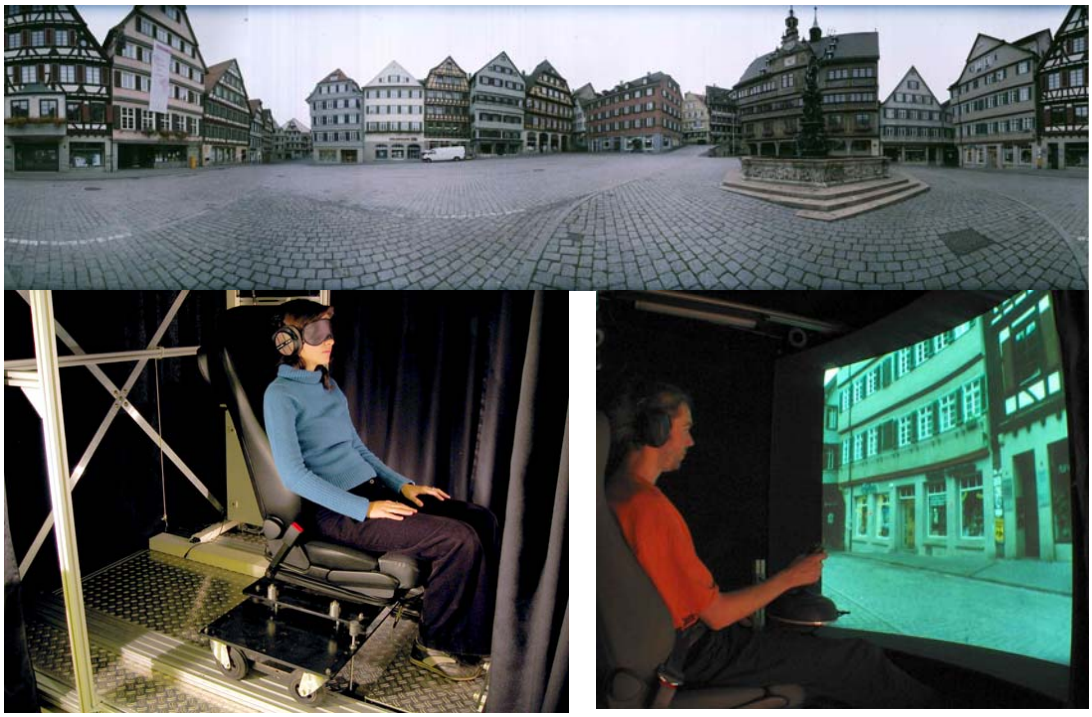


Figure 2. *Top: 360° roundshot photograph of the Tübingen market place. Bottom left: Auditory experiment with a participant sitting on a chair mounted on a wheeled platform coupled with a footrest. Bottom right: Participants were seated at a distance of about 1.8m from a curved projection screen (Riecke, 2005a).*

Visual stimuli are rendered in real time using the VElib software (VElib). Auditory stimuli are played back with Beyerdynamic DT-990Pro

circumaural headphones, and real-time sound rendering is possible via the Lake Huron auralization system. Vibrotactile stimulation is delivered via mechanical shakers mounted under the chair and the footrest.

5.2 Stimuli

In the majority of the experiments (papers A, C and D) ecological, close to everyday experience, stimuli have been presented. We decided to concentrate on the ideas of ecological acoustics which studies sound perception from the perspective of every-day listening experiences (Gaver, 1993). Previous research by Larsson et al. (2004) showed that the experience of self-motion was significantly higher for sound fields with sound sources clearly recognizable as "acoustic landmarks". The visual stimulus used in paper D consisted of a photorealistic view of the Tübingen market place that was generated by wrapping a 360˚ roundshot (4096 × 1024 pixel) around a virtual cylinder (see Figure 2, top). Auditory-tactile naturalistic pair was used in paper C, where additional vibrotactile stimulation was synchronized with a sound representing a virtual engine.

All experiments were conducted using virtual auditory space and the stimuli were synthesized in Matlab™ using catalogues of generic HRTFs or individualized HRTFs in paper A. Binaural synthesis and headphone based sound reproduction provided best flexibility in creating and rendering virtual auditory environments. Moreover, the author was interested in how AIV can be affected by imperfections in spatial sound rendering due to the use of generic HRTFs (see subsection 2.1.1). The generic HRTFs catalogue was measured from a KEMAR mannequin using the procedure described in paper A. Only one horizontal plane (-4 degree elevation) with a 5 degree resolution was used for stimuli synthesis. No acoustic environment rendering was applied, as the virtual scene under investigation – middle of Tübingen market square – did not provide perceptually distinct acoustic cues (Larsson et al., 2004).

5.3 Specific features of the experimental setup

It should be noted that special measures were taken to amplify the auditory self-motion sensation – several studies confirm that the visual appearance (e.g. seen loudspeakers) of the experimental setup can significantly bias participants' responses (Larsson et al., 2005; Gustaviano et al., 2005). We also believe that in the context of self-motion stimulation, seeing a completely immovable setup could negatively bias the AIV responses. Therefore, in these translational AIV experiments (papers B and C) participants were seated on a chair mounted on a wheeled platform connected to a wheeled footrest[6] (see Figure 2, bottom left). Participants

[6] In paper A, a turntable was used for the same purpose.

were not explicitly told that the chair could physically move, but the facts that the wheels were visible and that the chair moved slightly when they sat down on it should have insured them that the simulator setup potentially could move.

5.4 Verbal measures

To assess self-motion and subjective presence sensation, three direct measures were used in the presented papers: presence, vection intensity and convincingness of vection direction.

Presence was defined in the questionnaire as "a sensation of being actually present in the virtual world". This item has been chosen from the Swedish Viewer-User Presence (SVUP) questionnaire developed and verified in (Larsson et al., 2001). Vection intensity corresponded to the level of the subjective sensation when experiencing self-motion. In the experiments in papers A, B and C, participants were blindfolded and were asked about the direction of perceived self-motion ("heading") and how convinced they were about this direction[7]. It should be noted that convincingness and intensity ratings are usually highly correlated but in some specific cases they can differ (e.g. see paper C). Ratings of all three measures were given on a 0-100 scale where 100 corresponded to a sensation equal to that of a real event (physical motion of a chair or "being there").

Questionnaire-based measures are often criticized for being noisy and not reflecting user internal state (Slater, 2004). Within the experiments in the POEMS project conducted at Chalmers, self-motion sensation has also been accessed by EDA (electrodermal activity) measures, and a strong correlation with the verbal response was seen. Therefore the author decided to rely on verbal and behavioral measures for self-motion throughout the experiments presented in this thesis.

Compared to self-motion reports, participants found it more difficult to give verbal ratings on presence sensation. Using questionnaire specifically developed for media applications (e.g. ITC-SOPI which accesses various dimensions of presence (Dillon et al., 2000) might provide better sensitivity). However, applying multi-dimensional questionnaires after each trial would require significantly longer time for experiments (average experiment duration was 40-45 minutes).

Additionally, we expected to find a correlation between presence and self-motion reports as both responses reflect the degree of motion simulator effectiveness. Interestingly, a strong correlation between vection and presence responses has been found in the experiments containing visual stimuli (e.g. Riecke et al., 2005a, paper D). Therefore, in order to find

[7] Due to artefacts in binaural synthesis, e.g. front-back confusions, reported direction of self-motion served as a quality indicator for rendered auditory scenes.

possible individual differences in participants self-reports, the experiments in papers B, C and D were accompanied by questionnaires on imagery (Sheehan, 1967), need for cognition (Cacioppo et al., 1984) and tests on cognitive style (Witkin & Goodenough). Analysis of this data is still in progress. Apart from the direct measures of self-motion, a binary measure reflecting the number of self-motion experiences was used. While vection onset time often is used in experiments with visual stimuli (e.g. Riecke et al. 2005a), onset time was not measured in the presented studies since previous experiments (Larsson et al., 2004) on auditory-induced vection indicated that this measure resulted in large inter-individual variance.

6 Research results summaries

The work in this thesis represents a fraction of the POEMS project research conducted at Chalmers. The research pathway of this thesis is graphically shown on the tree-like structure on Figure 3. The experiments presented in this work investigated either rotational (papers A and D) or translational self-motion (papers B and C). At the first stages of the project, purely auditory-induced self-motion effects have been investigated (papers A and B). Later on, other contributing modalities – vibrotactile (paper C) and visual (paper D) stimulations have been added to presented auditory scenes.

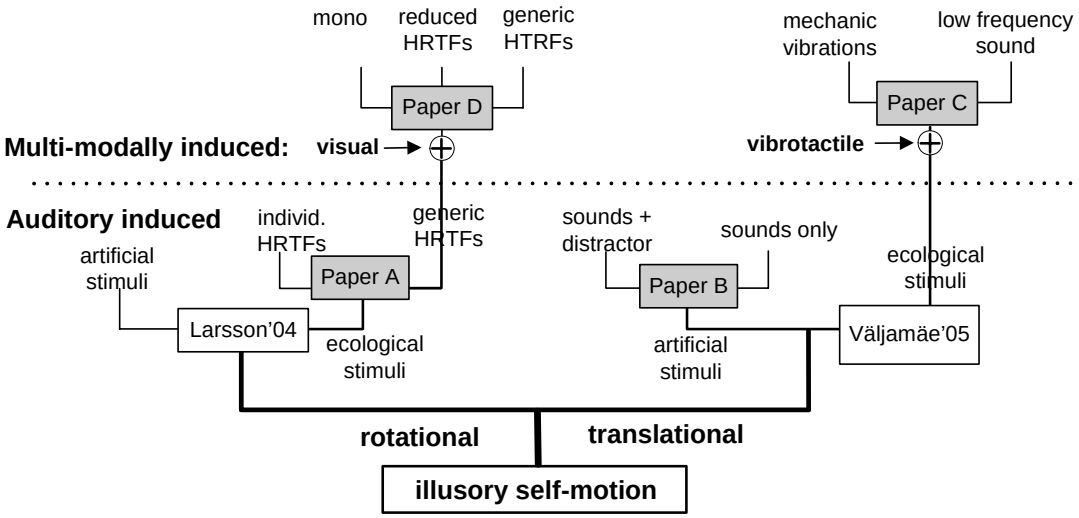


Figure 3. *Schematic representation of the work presented in this thesis (papers A-D). Branches from each experiment "box" represent the main conditions varied in the experiment. Papers C and D represent the experiments with multi-modally induced self-motion complementing the auditory only experiments from papers A and B.*

I joined the project after its first year and paper A has been based on previous work by Larsson et al. (2004). In paper A, the same ecological stimuli as in (Larsson et al., 2004) have been used for binaural synthesis with either generic or individualized HRTFs. Later on, the same acoustic landmarks have been used together with ecological visual stimuli (the market square roundshot) in paper D.

The work on translational illusory self-motion started with a separate experiment with artificial sounds, as described in paper B. A similar study but with ecological sound types is not presented in this thesis (see Väljamäe et al., 2005). Paper C represents the translational self-motion experiments where ecological stimuli have been combined with vibrotactile stimulation.

6.1 Paper A: Auditory presence, individualized head-related transfer functions and illusory ego-motion in virtual environments.

The main aim of paper A was to investigate how individualized binaural synthesis, where participants' own HRTFs had been used, would affect self-motion and presence ratings. In the previous experiment by Larsson et al. (2004) the same auditory scenes were synthesized using generic HRTFs, and post-experimental verbal probing revealed that participants often experienced artefacts typical for binaural reproduction (e.g. distorted sound sources trajectories or in-head localization, see subsection 2.1.2). In the current experiment, 12 participants listened to stimuli synthesized using either generic or their own pre-measured HRTFs.

There are several conclusions from this initial investigation. First, it was found that individualized HRTFs increased presence ratings. Second, the results were consistent with the previous results reported by Larsson et al. (2004), where the number of rotating sound sources influenced both presence ratings and self-motion experiences. Third, inter-group differences were found within the subjects, which were most likely caused by errors occurring during the fast measurement of individualized HRTFs catalogues. Surprisingly, participants from the group with poor localization performance showed no discrimination in AIV responses between conditions with lower or higher number of sound sources in rotating auditory scenes. Finally, it is important to note that stimuli processed with generic (KEMAR) HRTFs also induced self-motion, regardless of the lowered spatial consistency of the auditory scene.

6.2 Paper B: Traveling without moving: Auditory scene cues for translational self-motion

The main aim of paper B was to study auditory-induced translational self-motion. A literature review on auditory cues for motion perception in paper B suggested that sound intensity and interaural delay (ITD) are typically the most salient cues for auditory motion perception. In the study, artificial stimuli were used and several parameters were varied: sound velocities, velocity types (accelerating vs. constant) and initial positions of the sound sources. Additionally, a focused attention task was used in half of the trials, where participants had to report changes in a stationary tonal sound added to the auditory scene containing moving sound sources.

A main finding in this study was that the additional tonal sound (focused attention task) significantly decreased self-motion and presence ratings. One explanation to this result could be a difference in methodology: focused attention task versus "relaxed" listening conditions. However, the author did not found similar significant decrease in AIV and presence

ratings as observed in paper B for other experiments containing focused attention task (work in progress). Another explanation for the found negative effect is the reduced auditory scene consistency – participants often commented that a tonal sound was perceived as highly unnatural and disturbing compared to the moving sounds composed from band-passed pink noise. Importantly, participants usually created meaningful objects and scenes when exposed to the abstract stimuli (e.g. being in the train, metro or driving in a tunnel). Frequency-modulated tonal sound usually was perceived as distracting and "artificial" compared to the moving noises. Interestingly, further studies on translational AIV with ecological sounds showed an opposite effect - a stationary sound resembling engine noise had a positive effect on self-motion and presence ratings (Väljamäe et al., 2005).

In general, the results from the studies in paper B showed that translational AIV is more easily induced than rotational AIV. This is not surprising as translational movement is a more frequent experience from an ecological validity standpoint. Additionally, the almost equal percentage for front-back and back-front reversals was found, when typically localization reversals are rather asymmetric (front-to-back dominates, Begault, 1994). The reason why such asymmetry was not found in the current results can be explained by the fact that people are simply more used to move in forward direction. This finding suggests that an ecological context might influence auditory scene perception.

This experiment and the previous findings in paper A suggest that the auditory scene consistency and ecological validity plays a crucial role in AIV. In the current experiment, we deliberately used only the most salient acoustic cues (sound intensity and ITD) for moving sound fields simulation, and the quality of the rendering was determined by a generic HRTFs catalogue. The results suggest that this reduced level of details in spatial sound rendering can be sufficient for creating a self-motion sensation. This, in turn, would allow allocating sound processing resources for other tasks including, for example, low latency rendering of real-time interaction.

6.3 Paper C: Vibrotactile enhancement of auditory induced self-motion and presence

Paper C investigated the effects of additional vibrotactile stimulation on translational AIV. The methodology resembled the one in paper B but, instead of artificial sounds, ecological stimuli were used. As in paper B, only a minimal set of acoustic cues was applied for rendering of the moving sound fields, which lead to some perceptual artefacts of the auditory scene (e.g. distortions in the perceived trajectories of sources' motion). The auditory scenes contained moving sound fields (acoustic landmarks) and/or non-spatialized sound representing a sonic motion metaphor (engine sound).

The main finding confirmed that the vibrotactile stimulation significantly enhanced self-motion and presence responses for AIV, similar to vibrotactile enhancement of visually-induced self-motion reported in (Schulte-Pelkum et al., 2004). In the current AIV experiment, the observed self-motion facilitation via vibrotactile stimulation might be seen as a way to compensate for a rendering quality of spatial sound.

Interestingly, vibrotactile stimulation effects varied depending on the auditory scene content. The cross-modal enhancement of AIV and presence ratings was strongest for the auditory-vibrotactile engine metaphor, suggesting a cognitive nature of the observed effect. The starting of the engine sound was synchronized with the vibrations initial burst, leading to a recognizable perceptual event (cf. strong negative effect for unrecognizable, distracting sound in paper B).

6.4 Paper D: Sound can compensate for a restricted field-of-view in self-motion simulators

Paper D represents the most recent study among the experiments presented in this thesis. This study replicated and extended the findings from the audio-visual experiment by (Riecke et al., 2005a). The study showed a small but significant facilitation of self-motion and presence responses for conditions where spatial sound was added to the visual stimulus. However, the observed auditory enhancement could be weak because of the ceiling effect due to the strong visual input (field of view of 54˚×45˚ was used). In the current experiment two visual stimuli conditions with smaller FOV were combined with three spatial sound quality conditions (mono, reduced generic HRTFs and generic HRTFs).

The results showed that spatial sound can significantly increase presence and self-motion responses. Compared to (Riecke et al., 2005a), where the spatial sound condition resulted in a 16% increase in vection intensity ratings, a 27% increase was registered for conditions with medium FOV (20˚×15˚) in our study. However, for a smaller FOV (10˚×7.5˚) the auditory facilitation effect became insignificant. One possible explanation for this finding is the large spatial incongruence between auditory and visual stimuli, as the small FOV condition resulted in a very scarce representation of the market square and heard sound objects could not be seen. The currently conducted audio-visual experiments are investigating in more detail the observed effects for smaller FOV (10˚×7.5˚).

In addition, the results also showed that using generic HRTFs with reduced spatial resolution (only 6 directions were used) was sufficient to observe auditory enhancement of vection. Interestingly, this effect replicates the visually induced ceiling effect found by (Riecke et al., 2005a) for the auditory domain and suggests that a high resolution spatial sound rendering might not necessarily facilitate audio-visual vection and presence (cf. AIV

experiment results in paper A). The sound condition with reduced spatial quality is comparable to the common surround sound standard with 5-loudspeaker configuration, and this finding can encourage vection oriented audio-visual materials production and broadcasting, for example, for home theatres use.

6.5 Comparison between the presented results

It is difficult to compare the results presented in papers A[8] to D because of the variation in the number of participants and changes in methodology. For example, in the experiments with visual stimuli in paper D, self-motion intensity ratings were lower than in the experiments where participants were blindfolded and imagery could affect the reported results (e.g. Marx et al., 2003). Currently, data from participants imagery scores (Sheehan, 1967) is being collected and analyzed. Nevertheless, a good illustration of the major effects described above can be given by a binary vection measure which shows how many participants reported self-motion for a given stimuli type (Figure 4).

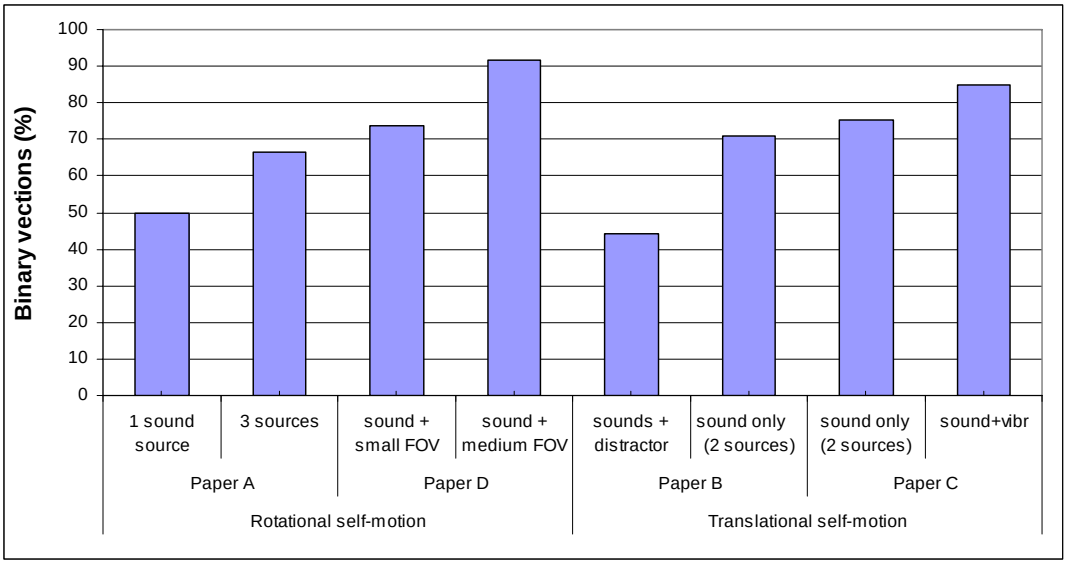


Figure 4: *Binary vections (100% correspond to stimuli where all participants in given experiment experienced vection). The number of participants was 12 (Paper A), 24 (Paper D), 24 (Paper B) and 23 (Paper C).*

First, one can see that translational AIV is easier to induce than rotational AIV: 50% (1 sound source) and 67% (3 sound sources) in paper A compared to 71% and 75% (2 sound sources) in papers B and C. It can be also seen that stimuli with moving noises in paper B are comparable to the

[8] Results for paper A – the effects of individualized HRTFs – did not show any difference in the binary vection measure, and for illustration purposes the conditions with single vs. multiple sound sources contained in the rotating sound fields are presented.

results obtained with ecological stimuli in paper C. This once again shows that participants in experiment B were creating a context from the presented scenes (e.g. “like being in the train”). Number of experienced vections in B experiment significantly dropped to 44% when the unrecognizable tonal (distracter) sound was added. Bimodal enhancement effects are also clearly represented: vibrotactile stimulation resulted in 10% increase (from 75 to 85%) and visual stimuli with medium FOV[9] in 25% (from 67 to 92%).

[9] The small FOV condition (10˚×7.5˚) did not show any significant effects for self-motion or presence.

7 General conclusions

Throughout the results presented in this thesis, the reader can see that sound plays an important role in self-motion perception, thus, supporting the importance of multisensory displays in VR and other perceptually mediated environments. Interestingly, the research results from auditory-vibrotactile and auditory-visual stimulation showed that spatial sound rendering, even with reduced quality (non-individualized HRTFs in paper A, limited spatial resolution in paper D), might be sufficient for eliciting self-motion sensation. In addition, the results from the presented experiments - decreased AIV for distracter sound in paper B; increased AIV for an auditory-vibrotactile engine representation in paper C; observed floor effect for a small FOV in paper D, (see also Riecke et al., 2005b) - suggest that consistency of the presented multi-modal scenes is one of the key factors for a cost-effective design of self-motion simulators.

Returning to the examples of commercially available products aiming to produce self-motion experience (IMAX and motion simulator at FAO Schwarz) mentioned in the introduction, one should ask how the results presented in this thesis could enhance the existing technologies with a minimum effort. Presented research shows that sound appears to be an important component in self-motion simulators having both positive and negative effects on the end user's perception.

One could distinguish three characteristics of rendered sound which can be instrumental for the enhancement of a self-motion simulator content. First, there is a number of auditory cues related to motion perception (see paper B for a review), such as sound intensity and related "point of closest passage" of the objects passing by, binaural cues (ITD, ILD) and Doppler effect (if higher velocities need to be simulated). Second, from the ecological acoustics perspective, sound source types like "acoustic landmarks", sounds creating representing motion (e.g. engine sound) and contextual information also play an important role in self-motion sensation enhancement (Larsson et al. 2004, Väljamäe et al. 2005, paper B). Finally, the spatial properties of the rendered sound are important not only from the binaural motion cues perspective, but also as carrying the information about the VEs' spatial characteristics, the so called "spaciousness" of the perceived environment (Rumsey, 2002). It should be noted that both binaural cues and "spaciousness" could be rendered using conventional 5-channel reproduction systems or using binaural synthesis with reduced spatial resolution (see paper D for a detailed discussion).

The misuse of sound in motion simulators can be related, for example, with a randomly selected background music with fast tempo incongruent with the rhythm conveyed in other modalities (visual, vibrotactile, etc.) and overall motion pattern. As discussed before (see section 3.4), one should rather aim for composing a multi-modal rhythmic structure where clearly

recognizable events (e.g. objects passing by or hitting a vehicle, road bumps, etc.) are represented synchronously in several modalities. This cross-modal rhythm perception will be strongly driven by auditory information (see sections 3.2 and 3.4 for discussion and references). Still, it is a fascinating research question to which extent the temporal structure of VE can be distributed across modalities – for example, a visual event followed by other auditory event followed by another visual event etc.

8 Future work

This thesis presented a fraction of the research conducted within the POEMS project and the findings described here establish a good basis for the follow up studies. An interesting topic to address is individual differences in the perception of multimodal environments in general and self-motion simulators in particular. Indirect, more objective measures of presence and self-motion (see Figure 1 and introduction) would also help to fine-tune the methodology used throughout this thesis.

First indicative results on the importance of the presented multi-modal scenes consistency on self-motion and presence responses should be further elaborated. It appeared that scene context might influence the perceived virtual world and compensate for rendering imperfections in one of the modalities (cf. auditory-vibrotactile interaction effect in paper C). In addition, paper B demonstrated that the front-back confusion rate can be biased by the auditory scene context. Further studies should provide better understanding how cognitive effects can contribute to the creation of a unified perceptual image from a reduced set of multisensory cues.

References

Alais, D., & Burr, D. (2004) The ventriloquist effect results from near optimal crossmodal integration. *Current Biology, 14* (3), 257–262.

Battaglia, P.W., Jacobs, R.A., & Aslin, R.N. (2003). Bayesian integration of visual and auditory signals for spatial localization. *Journal of the Optical Society of America, 20* (7), 1391-1397.

Baudisch, P., DeCarlo, D., Duchowski, A. T., & Geisler, W. S. (2003). Focusing on the essential: considering attention in display design. *Communications of the ACM 46* (3), 60–66.

Begault, D.R. (1994). *3D Sound for Virtual Reality and Multimedia*. London: Academic Press Professional.

Berkhout, A.J. (1988). A holographic approach to acoustic control. *Journal of the Audio Engineering Society, 36*(12), 977-995.

Bertelson, P., & de Gelder, B. (2004). The psychology of multimodal perception. In C. Spence & J. Driver (Eds.), *Crossmodal space and crossmodal attention* (pp. 141-178). New York: Oxford University Press.

Bitton, J., & Agamanolis, S. (2004). RAW: Conveying minimally-mediated impressions of everyday life with an audio-photographic tool. *Proceedings of CHI 2004,* ACM Press, 495-502.

Blauert, J. (1997). *Spatial Hearing*. Cambridge, MIT Press, rev. edition

Bronkhorst, A., & Houtgast, T. (1999). Auditory distance perceptions in rooms. *Nature, 397,* 517-520.

Brown, J. M., Anderson, K. L., Fowler, C. A., & Carello, C. (1998). Visual influences on auditory distance perception. *Journal of the Acoustical Society of America, 104,* 1998.

Cacioppo, J.T., Petty, R.E., & C.F. Kao (1984). The efficient assessment of need for cognition, Journal of Personality Assessment, vol. 48, pp. 306-307

Cater, K. (2004). Detail to attention: Exploiting limits of the human visual system for selective rendering. (Doctoral dissertation, University of Bristol, 2004).

Chion, M. (1990), *Audio-Vision: Sound on screen*. New York: Columbia University Press.

Dillon C, Keogh E, Freeman J, Davidoff J (2000) Aroused and immersed: the psychophysiology of presence. In: Proceedings of 3rd International Workshop on Presence, Delft University of Technology, Delft, The Netherlands, March 2000, pp 27–28

Durlach, N. I., & Mavor, A. S. (Eds.). (1995). *Virtual reality: Scientific and technological challenges* (pp. 161-187). Washington, D. C.: National Academy Press.

Eidsvik, C. (2005). Background tracks in recent cinema. In J.D. Anderson & B.F. Anderson (Eds.), *Moving image theory: Ecological condisderations* (pp. 70-78). Carbondale, IL: Southern Illinois University Press.

Evans, M.J., Tew, A.I., & Angus, J.A.S. (1997). Spatial audio teleconferencing: Which way is better? *Proceedings of the fourth International Conference on Auditory Display (ICAD '97)*, Palo Alto, California, 29-37.

Fahlenbrach, K. (2002). Feeling sounds. Emotional aspects of music videos. *Proceedings of IGEL 2002 conference*, Pécs, Hungary, 2002

Friedman, D., Leeb, R., Garau, M., Guger, C., Keinrath, C., Pfurtscheller, G., et.al (2004). Navigating virtual reality by thought: First steps. *Proceedings of 7th International Conference on Presence, Presence 2004,* 160-167. Valencia, Spain.

Fujisaki, W., Shimojo, S., Kashino, M., & Nishida, S. (2004). Recalibration of audiovisual simultaneity. *Nature Neuroscience, 7,* 773 – 778.

Gebhard, J. W., & Mowbray, G. H. (1959). On discriminating the rate of visual flicker and auditory flutter. *The American Journal of Psychology, 72,* 521–528.

Gerzon, M.A. (1985). Ambisonics in multichannel broadcasting and video. *Journal of the Audio Engineering Society, 33*(11), 859-871.

Graziano, M.S.A., Reiss, L.A., & Gross, C.G (1999). A neuronal representation of the location of nearby sounds. *Nature, 397,* 428-430

Gupta, N., Barreto, A., & Ordonez, C. (2002). Spectral modification of head-related transfer functions for improved virtual sound spatialization. *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing,* Orlando, FL, 2, 1953–1956.

Guttman, S. E., Gilroy, L. A., & Blake, R. (2005). Hearing what the eyes see: Auditory encoding of visual temporal sequences. *Psychological Science, 16* (3), 228-235.

Hall, D.A., & Moore, D.R. (2003). Auditory neuroscience: the salience of looming sounds, *Current Biology, 13* (3), R91-R93.

Härmä, A., Jakka, J., Tikander, M., Karjalainen, M., Lokki, T., Hiipakka, J., et.al. (2004). Augmented reality audio for mobile and wearable appliances. *Journal of the Audio Engineering Society, 52*(6), 618-639.

Hoisko, J. (2003). Early experiences of visual memory prosthesis for supporting episodic memory. *International Journal of Human- Computer Interaction, 15* (2), 209-320.

Holloway, R (Director). (1994). Paradjanov: A requiem [Motion picture]. O-Film Productions

Horbach, U., & Boone, M.M. (1999). Future transmission and rendering formats for multichannel sound. *Proceedings of the AES 16th Conference on Spatial Sound Reproduction*, Rovaniemi, Finland.

Howard, I. P. (1982). *Human visual orientation.* New York: Wiley.

Hughes, D. E., Thropp, J., Holmquist, J., & Moshell, J. M. (2004). Spatial perception and expectation: factors in acoustical awareness for MOUT training. *Proceedings of Army Science Conference (ASC) 2004*, Orlando, FL.

Itti, L., & Koch, C. (2001). Computational modeling of visual attention. *Nature Reviews Neuroscience, 2* (3), 194-203.

Jones, M.R. (2004). Attention and timing. In J. G. Neuhoff (Ed.), *Ecological psychoacoustics* (pp. 49-88). San Diego: Elsevier Academic Press.

Kankanhalli, M.S., Wang, J., & Jain, R. Experiential sampling in multimedia systems, in *IEEE Multimedia Journal* (submitted 2004).

Kitagawa, N., & Ichibara, S. (2002). Hearing visual motion in depth. *Nature, 416,* 172-174.

Kleiner, M., Dalenbäck, B-I., & Svensson, P. (1993). Auralization: An overview. *Journal of the Audio Engineering Society, 41*(11), 861-875.

Knudsen, E.I. (2002). Instructed learning in the auditory localization pathway of the barn owl. *Nature, 417,* 322–328.

Knudsen, E.I., & Brainard, M.S. (1995). Creating a unified representation of visual and auditory space in the brain. *Annual Review of Neuroscience, 18,* 19-43.

Kohlrausch, A., Fassel, R., & Dau, T. (2000). The influence of carrier level and frequency on modulation and beat-detection thresholds for sinusoidal carriers. *Journal of the Acoustical Society of America, 108,* 723–734.

Kohlrausch, A., van de Par, S. (2005) Audio-visual interactions in the context of multimedia applications In: J. Blauert, *Communication Acoustics* (pp 109-138). Springer

Kozyrev, N.A. & Nasonov V.V. "О некоторых свойствах времени, обнаруженных астрономическими наблюдениями наблюдения//Проявления космических факторов на Земле и звездах" [On some properties of time, discovered by astronomical observations//Influence of cosmic factors on Earth and starts], Проблемы исследования Вселенной, vol 9, pp. 76-84, 1980.

Kubovy, M., & Van Valkenburg, D. (2001). Auditory and visual objects. *Cognition, 80*, 97-126.

Làdavas, E., Farnè, A. (2004). Neuropsychological evidence for multimodal representations of space near specific body parts. In: C. Spence & J. Driver (Eds.), *Crossmodal space and crossmodal attention* (pp 69-98). Oxford: Oxford University Press.

Landis, C. (1954). Determinants of the critical flicker-fusion threshold. *Physiological Reviews, 34,* 259–286.

Len Lye filmography. Retrieved 22 May, 2005 from the Len Lye Foundation Web site: http://www.govettbrewster.com/NR/lenlye/foun/default.htm

Lewkowicz, D. J. (1996). Perception of auditory-visual temporal synchrony in human infants. *Journal of Experimental Psychology: Human Perception and Performance, 22* (5), 1094-1106.

Lewkowicz, D. J. (1999). The development of temporal and spatial intermodal perception. In G. Aschersleben, *Cognitive contributions to the perception of spatial and temporal events* (pp. 395-420). Amsterdam: Elsevier.

Lickliter, R., & Bahrick, L. E. (2000). The development of infant intersensory perception: Advantages of a comparative, convergent-operations approach. *Psychological Bulletin, 126,* 260–280.

Luciani, A. (2004). Dynamics as a common criterion to enhance the sense of presence in virtual environments *Proceedings of 7th International Conference on Presence, Presence 2004,* 96-103. Valencia, Spain.

Mack, A. and Rock, I. (1998). *Inattentional Blindness*. Massachusetts Institute of Technology Press.

Maier, J.X., Neuhoff, J.G., Logothetis, N.K., & Ghazanfar, A.A. (2004). Multisensory integration of looming signals by rhesus monkeys. *Neuron, 43,* 177-181.

Marker, C. (Producer & Director). (1962). *La Jetée* [Motion picture]. France: Argos Film.

Marx, E., Stephan, T., Nolte, A., Deutschländer, A., Seelos, K.C., Dieterich, M. and Brandt, T., 2003. Eye closure in darkness animates sensory systems. *NeuroImage* **19**, pp. 924–934.

Mastoropoulou, G., & Chalmers, A. (2004). The effect of music on the perception of display rate and duration of animated sequences: an experimental study. *Proceedings of Theory and Practice of Computer Graphics 2004* (TPCG'04) (pp. 128 - 134). IEEE Computer Society.

Mills, A. W. (1958). On the minimum audible angle. *Journal of the Acoustical Society of America, 30,* 237-246.

Miner, N., & Caudell, T. (1998). Computational requirements and synchronization issues for virtual acoustic displays. *Presence: Teleoperators and Virtual Environments, 7* (4), 396-409.

Morein-Zamir, S., Soto-Faraco, S., & Kingstone, A. (2003). Auditory capture of vision: Examining temporal ventriloquism. *Cognitive Brain Research*, *17*, 154–163.

Müller, M. M., Teder-Salejarvi, W. & Hillyard, S. A. (1998). The time course of cortical facilitation during cued shifts of spatial attention. *Nature Neuroscience, 1,* 631-634.

Nathanail, C., Lavandier, C., Polack, J. D., & Warusfel, O. (1997). Influence of sensory interaction between vision and audition of the perceived characterization of room acoustics. *Proceedings of the International Computer Music Conference,* Greece, 414-417.

Nielsen, S. (1991). Distance perception in hearing. (Doctoral dissertation, Aalborg University, 1991).

O'Sullivan, C., Howlett, S., Morvan, Y., McDonnell, R. & O'Conor, K. (2004). Perceptually adaptive graphics. *Proceedings of Eurographics 2004*, State of the Art reports. September 2004.

Pfurtscheller, G., Neuper, C., & Birbaumer, N. (2005). Human brain-computer interface. In E. Vaadia & A. Riehle (Eds.), *Motor cortex in voluntary movements: a distributed system for distributed functions. Series: Methods and New Frontiers in Neuroscience* (pp. 367-401). CRC Press.

Pulkki, V. (1997). Virtual sound source positioning using vector base amplitude panning. *Journal of the Audio Engineering Society,45*(6), 456–466.

Radeau, M., & Bertelson, P. (1977). Adaptation to auditory-visual discordance and ventriloquism in semirealistic situations. *Perception and Psychophysics, 22,* 137–146.

Recanzone, G. H. (2003). Auditory influences on visual temporal rate perception. *Journal of Neurophysiology*, *89*, 1078–1093.

Rensink, R. A., O'Regan, J. K., and Clark, J. J. (1997). To see or not to see: The need for attention to perceive changes in scenes." *Psychological Science*, 8, 368-373.

Riecke, B., J. Schulte-Pelkum, F. Caniard & H. Bülthoff. (2005a) Influence of auditory cues on the visually-induced self-motion illusion (circular vection) in virtual reality. *Proceedings of 8th international workshop on Presence 2005, 49-57*

Riecke, B., Västfjäll, D., Larsson P., & J. Schulte-Pelkum. (2005b) Top-Down and Multi-Modal Influences on Self-Motion Perception in Virtual Reality. *International Conference on Human-Computer Interaction (HCI international 2005), 1-10*

Rumsey, F. (2001). *Spatial Audio*. Oxford; Boston: Focal Press.

Rumsey, F. (2002). Spatial quality evaluation for reproduced sound: Terminology, meaning, and a scene-based paradigm. *Journal of the Audio Engineering Society, 50*(9), 651-666.

Sanchez-Vives, M.V. & Slater, M. (2005) From presence to consciousness through virtual reality, *Nature Neuroscience*, 6, 332-339

Schulte-Pelkum, J., B.E. Riecke & H.H. Bülthoff. (2004) Vibrational cues enhance believability of ego-motion simulation. *International Multisensory Research Forum (IMRF)*

Shams, L., Kamitani, Y., & Shimojo, S. (2000). What you see is what you hear. *Nature, 408,* 788.

Shams, L., Kamitani, Y., & Shimojo, S. (2002). Visual illusion induced by sound. *Cognitive Brain Research, 14* (1), 147–152.

Shams, L., Kamitani, Y., Thompson, S., & Shimojo, S. (2001). Sound alters visual evoked potential in humans. *NeuroReport, 12* (17), 3849–3852.

Sheehan, P.W. (1967), A shortened form of Betts' Questionnaire Upon Mental Imagery, Journal of Clinical Psychology, vol. 23, pp. 386-389, 1967

Shilling, R.D., & Shinn-Cunningham, B.G. (2002). Virtual auditory displays. In K.M. Stanney (Ed.), *Handbook of virtual environments: Design, implementation, and applications* (pp. 65-92). Mahwah, NJ; London: Lawrence Erlbaum Associates.

Slater, M (2004). How colorful was your day? Why questionnaires cannot assess presence in virtual environments. *Presence - Teleoperators and Virtual Environments*, 13(4):484–493

Slater, M. (2002). Presence and the sixth sense. *Presence: Teleoperators and Virtual Environments, 11,* 435-439.

Snoek, C., & Worring, M. (2002). Multimodal video indexing: a review of the state-of-the-art. *Multimedia Tools and Applications, 25* (1), 5-35.

Stassen, B. (Director). (1999). *Alien Adventure 3D* [Motion picture]. nWave Pictures

Thomas, F., & Johnston, O. (1981). *Disney animation: The illusion of life*. New York: Abbeyville Press.

Urbantschitsch, V. (1897) "Über Störungen des Gleichgewichtes und Scheinbewegungen [On disturbances of the equilibrium and illusory motions]", Zeitschrift für Ohrenheilkunde, vol. 31, pp. 234-294

Väjamäe, A., Kohlrausch, A., van de Par, S., Västfjäll, D., Larsson, P., & Kleiner, M. (2006a). Audio-visual interaction and synergy effects: implications for cross-modal optimization of virtual and mixed reality applications. Submitted to Handbook of presence.

Väljamäe, A. (2003). A feasibility study regarding implementation of holographic audio rendering techniques over broadcast networks. (Master thesis, Chalmers Technical University, 2003).

Väljamäe, A., Larsson, P., Västfjäll, D., & Kleiner, M. (2005). Sonic self-avatar and self-motion in virtual environments. Submitted to *Presence: Teleoperators and Virtual Environments*

Väljamäe, A., Tajadura, A., Västjäll, D. & Larsson, P. (Manuscript in preparation). Experimental study on crossmodal optimization: still image trains combined with binaural sound.

Väljamäe, A., Västjäll, D., Larsson, P., & Kleiner, M. (2006b). Perceived sound in mediated environments. Submitted to Handbook of presence.

van der Zee, E., & van der Meulen, A. W. (1982). The influence of field repetition frequency on the visibility of flicker on displays. *IPO Annual Progress Report, 17,* 76–83.

Västfjäll, D., Larsson, P., & Kleiner, M. (2005). Visualization and auralization in merging: Influence of visual information on room acoustic perception. *CyberPsychology and Behavior*.

Västfjäll, D., Larsson, P., & Kleiner, M. (in press). Cross-modal interaction in Sound Quality evaluation: Experiment using the Virtual Aircraft. *Journal of Sound and Vibration*.

Virtual Environments Library (VElib). **http://velib.kyb.mpg.de/de/**

Vroomen, J., & De Gelder, B. (2004a). Perceptual effects of cross-modal stimulation: The cases of ventriloquism and the freezing phenomenon. In G. Calvert, C. Spence & B. E. Stein (Eds.), *Handbook of multisensory processes* (pp. 141-150). Cambridge: MIT Press.

Vroomen, J., & De Gelder, B. (2004b). Temporal ventriloquism: sound modulates the flash-lag effect. *Journal of Experimental Psychology: Human Perception and Performance, 30* (3), 513–518.

Vroomen, J., Keetels, M., de Gelder, B., & Bertelson, P. (2004). Recalibration of temporal order perception by exposure to audio-visual asynchrony. *Cognitive brain research, 22,* 32-35.

Wallace, M. (2004). The development of multisensory processes. *Cognitive processing, 5* (2), 69-83

Wallace, M.T., Roberson, G.E., Hairston, W.D., Stein, B.E., Vaughan, J. W., & Schirillo, J. A. (2004). Unifying multisensory signals across time and space. *Experimental Brain Research, 158*(2), 252 – 258.

Warren, D. H., Welch, R. B., & McCarthy, T. J. (1981). The role of visual-auditory "compellingness" in the ventriloquist effect: Implications for transitivity among the spatial senses. *Perception and Psychophysics, 30,* 557–564.

Watanabe, T., Nanez, J.E., & Sasaki, Y. (2001). Perceptual learning without perception. *Nature, 413,* 844–848.

Welch, R.B. (1999). Meaning, attention, and the unity assumption in the intersensory bias of spatial and temporal perceptions. In G. Aschersleben, T. Bachmann & J. Musseler (Eds.), *Cognitive contributions to the perception of spatial and temporal events* (pp. 371-387). Amsterdam: Elsevier.

Welch, R.B., & Warren, D.H. (1980). Immediate perceptual response to intersensory discrepancy. *Psychological Bulletin, 88,* 638-667.

Wenzel, E. M., Arruda, M., Kistler, D. J., & Wightman, F. L. (1993). Localization using nonindividualized head-related transfer functions. *Journal of the Acoustical Society of America, 94*(1), 111–123.

Witkin, H. A., Wapner, S., & Leventhal, T. (1952). Sound localization with conflicting visual and auditory cues. *Journal of Experimental Psychology, 43,* 58–67.

Witkin, H.A. & Goodenough, D.R. (1981). Cognitive styles: Essence and origins. International Universities Press, Inc. NY.

Zahorik, P. (2001). Estimating sound source distance with and without vision. *Optometry and vision science, 78,* 270-275.

Zetzsche, C., Röhrbein, F., Hofbauer, M., & Schill, K. (2002). Audio-visual sensory interactions and the statistical covariance of the natural environment. In Calvo-Manzano, A., Perez-Lopez, A., & Santiago, J. S. (Eds.) [CD ROM]. Sevilla: Proc Forum Acusticum.

Zimmermann, R., Kyriakakis, C., Shahabi, C., Papadopolous, C., Sawchuck, A.A., & Neumann, U. (2004). The remote media immersion system. *IEEE MultiMedia, 11*(2), 48-57.

Zotkin, D.N., Duraiswami, R., & Davis, L. S. (2004). Rendering localized spatial audio in a virtual auditory space. *IEEE Transactions on Multimedia, 6*(4), 553-564.

Zwiers, M.P., Van Opstal, A.J., & Paige, G.D. (2003). Plasticity in human sound localization induced by compressed spatial vision. *Nature Neuroscience, 6,* 175–181.

Paper A*

Auditory Presence, Individualized Head-Related Transfer Functions, and Illusory Ego-Motion in Virtual Environments

Aleksander Väljamäe, Pontus Larsson, Daniel Västfjäll and Mendel Kleiner

Published in
Proceedings of Seventh Annual Workshop Presence 2004,
Valencia, Spain, October 2004, pp. 141-147

*The layout has been changed to fit the overall thesis style

Auditory Presence, Individualized Head-Related Transfer Functions, and Illusory Ego-Motion in Virtual Environments

Aleksander Väljamäe[1], Pontus Larsson[2], Daniel Västfjäll[2,3] and Mendel Kleiner[4]
[1]Department of Signals and Systems, Chalmers University of Technology, Sweden
[2]Department of Applied Acoustics, Chalmers University of Technology, Sweden
[3]Department of Psychology, Göteborg University, Göteborg, Sweden
[4]Program in Architectural Acoustics, Rensselaer Polytechnic Institute, Troy, NY, USA
{aleksander.valjamae@s2.chalmers.se, pontus.larsson@ta.chalmers.se,
daniel.vastfjall@psy.gu.se, kleinm2@rpi.edu}

Abstract

It is likely that experiences of presence and self-motion elicited by binaurally simulated and reproduced rotating sound fields can be degraded by the artefacts caused by the use of generic Head-Related Transfer Functions (HRTFs). In this paper, an HRTF measurement system which allows for fast data collection is discussed. Furthermore, effects of generic vs. individualized HRTFs were investigated in an experiment. Results show a significant increase in presence ratings of individualized binaural stimuli compared to responses to stimuli processed with generic HRTFs. Additionally, differences in intensity and convincingness of illusory self-rotation ratings were found for sub-groups of subjects, formed on the basis of subjects' localization performance with the given HRTFs catalogues.

Keywords--- Self-motion perception, Auditory presence, Binaural reproduction, Individualized Head-Related Transfer Functions.

1. Introduction

Creating a sense of presence in the end-user of a Virtual Environment (VE) is one of the main goals of Virtual Reality (VR) technology. The feeling of presence is often described as a sensation of "being there", whereas several other definitions exist [1]. Being "spatially present" in a specific context provides a stable reference frame, which allows for a good spatial orientation

and spatial updating. Illusory ego- or self-motion (vection) can be described as a sensation of actual movement relative to a stable surrounding environment and has been shown to be closely related to spatial presence. For example, positive correlation between presence ratings and on-set times for induced illusory self-motion has been recently shown in experiments with visual stimuli [2].

A large body of research is concentrated on the illusory self-motion elicited by visual stimuli. On the contrary, research on auditory illusory self-motion received little attention until recently [3-6]. In our first experiments conducted within the European project POEMS (Perceptually Oriented Ego-Motion Simulation) [7], we focused on illusory self-rotation induced by sound-fields [8]. In this study, rotating sound sources were presented to subjects via headphones and the sensation of ego-motion in the opposite direction was expected.

These previous experiments were based on the ideas of ecological acoustics and it was hypothesized that the type of the sound source is an important parameter when studying auditory-induced illusory self-motion. Opposite to “artificial” sounds (e.g. pink noise), ecological sound sources can be classified by a listener into spatially “still” (e.g. a church bell) or “moving” (e.g. footsteps) categories. A major finding was that both presence ratings and experience of self-motion were highest for the sound fields containing sound sources from the “still” category. Speed of sounds’ rotation and number of sources also positively affected the ratings [8].

The stimuli in [8] were created using binaural technology, which is a two-channel spatial sound rendering technology where headphones are typically used for playback [9]. Pre-measured catalogues of Head-Related Transfer Functions (HRTFs) are used for binaural sound synthesis, where a non-spatialized (“dry”) sound is convolved with transfer functions corresponding to the desired spatial position of the source. Larsson et. al. [8] used a catalogue of non-individualized HRTFs provided for the CATT Acoustics auralization software [10].

However, binaural sound synthesized with non-individualized HRTFs often can be perceived as distorted because of the mismatch between the listener’s own and generic transfer functions. When generic HRTFs are used the most common problem is in-head localization (IHL), where sound sources are not externalized but are rather perceived as being inside the listener’s head [11]. Another known artifact is a high rate of reversals in perception of spatial positions of the virtual sources, where binaural localization cues are ambiguous (cone of confusion), e.g. front-back confusion [9]. Errors in elevation judgments can be also observed for stimuli processed with non-individualized HRTFs [12]. Applying individualized HRTFs for auditory VEs can significantly reduce the artifacts described above [11].

One of the goals of this study was to test the performance of the HRTFs measurement system designed by Chalmers Room Acoustic Group (CRAG). We decided to repeat some of the experiments on illusory self-rotation presented in [8], this time using individualized HRTFs. Responses from verbal probing in [8] showed that in some cases the participants experienced artifacts in presented sound scenes, such as non-circular trajectories and nonrealistic, very close distances to the moving sources. We believe that these artifacts may have been caused by the use of non-individualized HRTFs and that these artifacts in turn might have influenced ego-motion and presence ratings.

The main hypothesis for the current experiment stated that the higher presence ratings should be achieved with individualized HRTFs, since previous studies indicate that spatialization and localization may be linked to presence experiences [13-14]. A second hypothesis was that improved spatial quality of auditory scene might affect subjects' experience of ego-motion.

Auditory localization performance can depend on several factors in complex spatial sound scenes. Langendijk et. al. [15] studied the effect of localization of target sounds in the presence of one or two distracter sounds, which were interleaved but not overlapped with the target sound in time. They found that the localization performance was degraded as the number of distracters increased. In auditory VE with multiple ecological sounds target-distracter pairs can easily occur.

We decided to investigate how distracters added to the auditory VE can influence presence ratings. Our third hypothesis was that adding the auditory distracters, which are irrelevant to the sound scene, would decrease the localization accuracy (divided attention). This in turn might decrease the influence of non-individualized HRTFs artifacts on the sound scene perception in VE.

2. HRTF measurement system

The procedure of measuring catalogues of individualized HRTFs is a cumbersome and time-consuming procedure (for reviews see [9, 12, 16-17]). In line with POEMS project requirements, it was decided to develop a HRTFs measurement system which would allow for fast data collection in non-anechoic and somewhat noisy environments such as offices. In the CRAG HRTF measurement system the transfer functions are recorded for a grid of spatial positions on a virtual sphere, which center is aligned with a subject's head-related coordinate system.

2.1. Physical setup

In our laboratory setup we built an array of 32 loudspeakers as shown in Fig. 1, which can be seen as one sector or ”vertical slice” of a virtual sphere with a radius of 1.25 meters (far-field acoustical mode measurements). Loudspeakers were non-uniformly placed at 16 elevation angles, which guarantee a resolution of less than 10 degrees in the vertical plane. A test person must be shifted 19 times (20 sectors) for the full HRTF catalogue measurement resulting in less than 8 degree resolution in the horizontal plane. If higher frontal resolution is required, one additional sector can be measured, which results in 63 measured azimuth positions in the horizontal planes with elevation angles of 16, 6, -4 and -14 degrees.

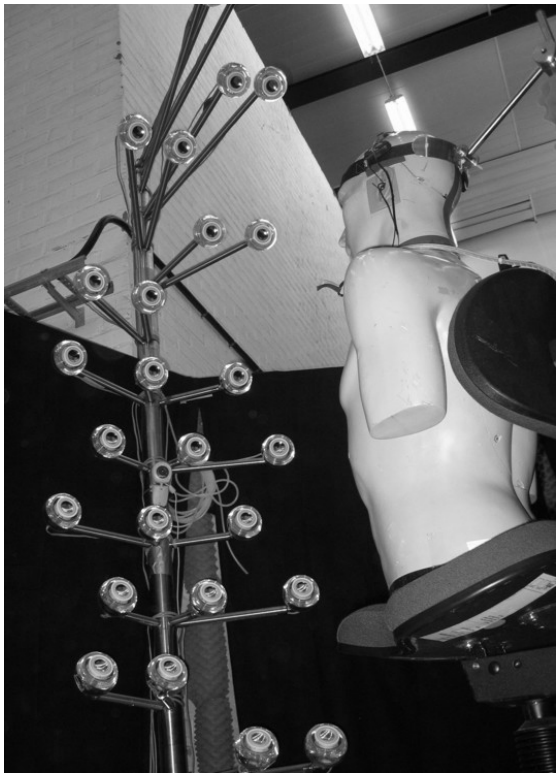


Figure 1: *HRTFs catalogue measurement system*

In our current setup, HRTFs catalogue measurements are conducted in a big room with a reverberation time of approximately 0.7s and noise floor of 30 dBA. The test subject is seated on a swivel chair and the head is fixed with a special headrest and an elastic band. The position of the measurement system in the room ensures that no early reflections arrive within the first few milliseconds after the direct sound. Additionally, the floor is covered with sound absorbing material.

Apple ProTM loudspeakers were chosen for the loudspeaker array because of their small size and spherical shape so as to reduce scattering of

sound by neighbouring loudspeakers. DPA 4060 miniature condenser microphones are placed at the entrance of the blocked-ear canal of the subject. This type of HRTF measurements gives results comparable to those measured with a microphone probe immediately inside the ear canal opening [18], but the procedure is faster and more convenient for the subject under measurement. An M-Audio Audiophile 2496 PCI-bus card is used for playback and recording of measurement signals because of its large dynamic range and the availability of driver routines for Linux. Specially written software for raw HRTFs data collection is used.

2.2. Post-processing of measured data

In our system we use frequency sweeps (quadratic chirp) as the deterministic stimulus for measurements of Head-Related Impulse Responses (HRIRs). This method was chosen due its immunity to harmonic distortions introduced by the measurement chain and the low crest factor of the signal (ratio between peak and root-mean-square in the voltage values) [19]. The latter property helps to ensure desired 20 dB signal-to-noise ratio for a desired frequency range. The chirp is band-limited from 0.1 to 15 kHz and its duration is 2048 samples (43 ms, fs = 48 kHz).

After the measurement procedure, the raw HRIRs are processed using MatlabTM software. The raw HRIRs are first deconvolved with the chirp signal and a half-Hanning window is applied to keep first 256 samples of the response. At this stage, the Interaural Time Delay (ITD) is estimated using the difference between onset times of the left and right HRIRs. This allows for optimizing in further processing steps by working only with the HRTFs magnitude responses. The ITD is added only to the final HRIR dataset. This decision was motivated by the assumption that it is possible to use linear-phase HRTFs for binaural sound synthesis without adding perceptually significant artifacts [20].

Furthermore, the raw HRIRs collected by the system have to be compensated for the artifacts coming from transducers limitations. For this purpose “baseline” or free-field responses are measured with the microphones placed at the center of the virtual sphere. Free-field corrected HRTF magnitude responses can then be obtained by division of the raw HRTF by the corresponding baseline TF data in frequency domain.

Interpolation is needed to obtain a final HRTFs catalogue with uniformly distributed spherical grid. For the current experiment, a 5-degree resolution in the horizontal plane was required for smooth rendering of moving sound sources. This is comparable with an average localization blur, which is 3 degrees for frontal positions and up to 20 degrees for peripheral and rear positions [11]. Spherical spline interpolation in the frequency domain is known to give best results compared to other interpolation types [21]. Before interpolation, magnitude responses are

smoothed using the procedure described in [22]. Instead of using a perceptually motivated reduction of magnitude response, smoothing is applied to all data points. After interpolation, HRTF magnitude responses are used for creating linear-phase FIR filters. A circular shift equal to the earlier estimated ITD is introduced within filter pairs, which represent HRTFs for certain spatial locations.

The processing steps presented above were used for creating the stimuli for the current experiment. One of the goals of this study was to test the performance of the CRAG HRTF measurement system and apply all necessary corrections at the post-processing stage if needed. At this stage we found good results for the HRTFs for the frequency range from 0.1 to 13 kHz. Deficiencies observed for frequencies above 13 kHz had no affect to the experiment as ecological sounds with frequencies from 0.1 up to 10 – 12 kHz were used.

3. Method

Twelve subjects (five male) with a mean age of 24 (SD 2.2) from the previous study described in [8] participated in the experiments. All subjects had normal hearing verified by a standard audiometric procedure [23]. After completing the experiment, subjects were debriefed, thanked and paid for their participation.

3.1. Measures

To assess auditory induced vection and subjective presence sensation, three direct measures were used in this experiment: presence, vection intensity and convincingness of vection.

Presence was defined in the questionnaire as “a sensation of being actually present in the virtual world”, which corresponded to the single perceptual dimension without any interaction with the VE. Vection intensity corresponded to the level of the subjective sensation when experiencing self-motion. On the Convincingness scale subjects had to report how convincing the sensation of self-motion was. It should be noted that the convincingness and intensity ratings often are highly correlated. Ratings of all three measures were given on a 0-100 scale.

Apart from the direct measures listed above, an indirect binary measure, reflecting the number of ego-motion experiences, was used (participants were asked to verbally indicate the direction of self-rotation). While on-set time for vection experience is often used in experiments with visual stimuli, the previous experiment [8] on auditory-induced vection indicated that this measure showed large inter-individual variance. The

present study measured onset time, but since no systematic effects were found, results from this measure will not be presented.

3.2. Stimuli

In the current experiment, rendering of acoustic environment was not considered as being important; therefore all stimuli were simulations of an anechoic environment rendered off-line. Three parameters varied in the presentation of the rotating sound field:

- Rotation velocity (20 or 60 degrees/second)
- Number of concurrent sound sources (1 or 3)
- Type of HRTFs catalogue used for stimuli synthesis (individualized or generic catalogue)

Since results from the previous experiment showed that the “still” type sound sources are the most instrumental in inducing ego-motion, only sounds from this category were used: “bus on idle”, “small fountain”, and “barking dog”. The stimuli duration was approximately 1 minute and consisted of the following parts: 3 seconds in stationary listening position, 4 seconds acceleration to maximum velocity, 60 seconds constant rotation speed and 4 seconds deceleration.

The stimuli were synthesized using one horizontal slice of HRTFs at -4 degree elevation. The stimuli synthesized with individualized HRTFs was contrasted with one processed with generic HRTFs catalogue, which resulted in 4 pairs of stimuli kept together for the verbal probing purposes. HTRFs measured from the KEMAR (Knowles Electronic Manikin for Acoustic Research) mannequin were used as the non-individual catalogue. Headphone equalization was applied to the final sound excerpts in order to prevent coloration artifacts.

For testing the effects of irrelevant auditory distracters, 20 clicks (6 kHz carrier, 4 ms duration) were added at random time moments to the two stimuli from the main set described above (1 and 3 sound sources, synthesis with generic HRTF, velocity of 60 degrees/second). Clicks were also convolved with KEMAR HRTFs and appeared at random positions in space. During the experiment stimulus with the clicks always followed the same stimulus without distracters hence creating two pairs for verbal probing.

Apart from the pair restrictions described above, all 10 stimuli were presented in the randomized order for proper statistical analysis.

3.3. Procedure

The experiment was conducted in a semi-anechoic room, where stimuli were played back with Beyerdynamic DT-990Pro circumaural headphones.

Participants were asked to report verbally the direction of their rotation – i.e. left or right, if they felt self-motion during the particular stimulus playback. Stimulus was stopped after the positive ego-motion response and subjects were asked to rate presence, intensity and convincingness.

In the current experiment presence is studied from the ego-motion perspective and rapid interruption, which could certainly influence the presence ratings, is acceptable. If the ego-motion sensation was not reported during the stimulus playback, only the presence rating was asked. Apart from the verbal responses to the questionnaire, verbal probing was done by the experiment leader.

Taking into account results by Lackner [3], special measures were taken in order to achieve auditory ego-motion. Participants were seated on an ordinary office chair, which was mounted on an electrically controllable turntable with a wooden base plate. Subjects were instructed that turntable would be used during some of the experimental trials. However, the turntable was still throughout the experimental trials. This manipulation was foremost used to make participants believe that they could actually move during the experiment. In addition, the turntable height prevented the subjects from having their legs any contact with the floor. In order to make the experimental setup look more convincing, four loudspeakers, placed around the experimental chair, were visible to participant as he/she entered the test room. The loudspeakers were never in use during the experimental trials. Finally, during the experiment participants were blindfolded.

3.4. HRTFs quality-test

In order to evaluate the quality of the individualized HRTFs a short test was performed before or after the main experiment. The purpose of this quality-test was to justify the improved localization performance compared with localization when using the generic (KEMAR) HRTFs catalogue. Instead of common strategies of defining absolute localization accuracy (e.g. [24]), we decided to use a simplified procedure, where a level of the most usual binaural rendering artifacts acts as indirect quality measure (it was important to keep the quality-test short in duration, as it had to be conducted together with the main experiment). The major parameters then are the front-back confusions rate and the externalization of perceived virtual sources.

It was decided to evaluate 6 positions on the horizontal slice at –4 degrees elevation: three in front (315, 0 and 45 degrees) and their reverse positions on rear (225, 180 and 135 degrees). A fountain sound of frequency range from 2 to 12 kHz was used as a sound source, since it is well known that, apart from small head movements, spectral differences in HRTFs around 5 and 9 kHz help to resolve front-back confusion [9]. Figure 2 shows how spectral differences can vary between the subjects.

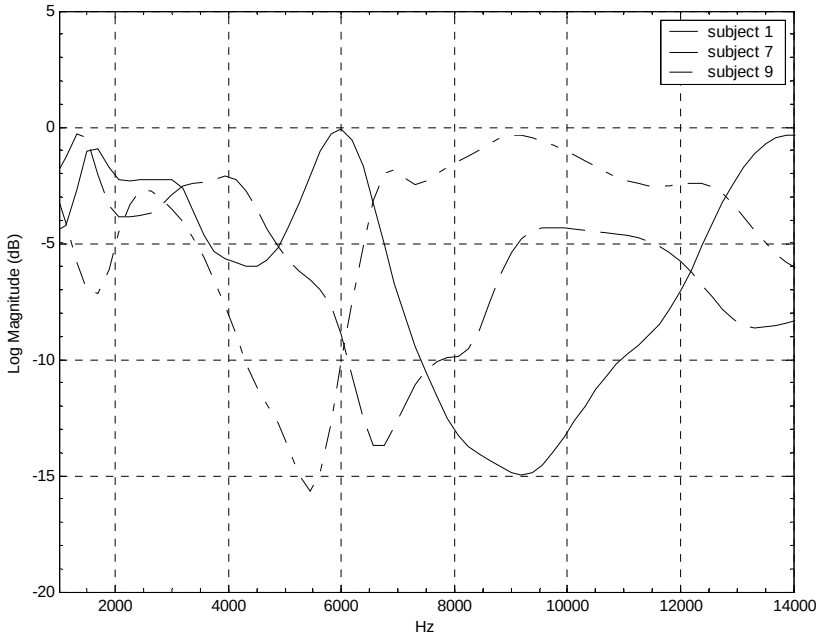


Figure 2: *Difference in spectra between two front-back source locations on a cone of confusion (45 and 135 degrees azimuth, -4 degrees elevation) for three representative subjects*

Four different measures were used for evaluating the HRTF quality: 1) the front-back confusion ratio, 2) the relative distance to the sound source, 3) errors in the elevation perception, and 4) responses from the verbal probing.

The front-back confusion ratio was based on the subjects' estimates of the spatial position of the sound source. In order to simplify this task, participants were asked to use a clock metaphor when verbally indicating one of the six positions listed above (e.g. 0 degrees azimuth corresponds to 12 o'clock).

For the distance evaluation, participants were asked to rate the distance to the source in meters. Later, answers were converted to the relative scale from 0 to 100, where 100 corresponded to the maximum perceived distance to the source from the stimuli pair processed with individualized or generic HRTFs.

For the elevation perception measure, subjects were asked to indicate the position of the source regarding the horizontal plane in the head-related coordinates system. Subjects were asked to indicate the height of the sound source relative to this plane.

As a last binary measure of the HRTF catalogues' quality, results from the verbal probing in the main experiment were used. Based on the subjects' comments on perceived trajectories of virtual sources, their distance and overall spatial scene consistency, decision was made regarding the

preference between stimuli processed by individualized or generic HRTF catalogues.

The quality-test stimuli consisted of three pairs of subsets synthesized by individualized and generic HRTFs catalogues. Each subset contained a fountain sound sequentially presented for 6 seconds from each of 6 spatial positions in random order. One pair was used for the elevation perception measure and two other pairs for collecting data on the front-back confusion ratio and the perceived source distance.

4. Results

4.1. Sub-groups based on the HRTFs quality-test

Results from the quality-test for all four measures highly varied among the subjects. However, when analyzing all 4 measures for each subject, these results could be combined into a final measure of preference between generic and individualized HRTFs used for the stimuli synthesis. Based on this binary measure, subjects were subdivided into two subgroups for further analysis: a G-group (better localization with individualized HRTFs) and a B-group (no clear preference between generic and own HRTFs). It has to be noted that localization performance could be degraded either by the HRTFs catalogues accuracy or due to individual localization abilities (see section 5 for a discussion).

For several subjects, the difference between the performance with individualized and generic HRTFs catalogues was not prominent but for a proper statistical analysis an equal number of participants was allocated to each group (i.e. median split). In this case the binary measure based on the verbal probing (see section 3.4) was used for the final decision.

Tables 1 and 2 show the average results for the three quality-test measures for all subjects and for the two subgroups formed. Table 1 shows average rates of front-back confusion, where responses from individualized and KEMAR HRTFs are compared. Means for all subjects showed an 11 % increase in the front-back confusion ratio for generic HRTFs. The sub-group analysis showed no effect for the B-group, but for the G-group 20 % improvement was found when individualized HRTFs were used.

Table 1: *Average rates (%) of front-back confusion for all subjects and the subgroups*

HRTFs	**All**	**G-group**	**B-group**
Ind.	33	32	35
KEMAR	44	53	35

Table 2 shows the responses for distance perception, where sub-group differences can be clearly seen. For the G-group, individualized HRTFs resulted in more distant percepts of the virtual source - 14 % improvement. For the B-group, the effect was reversed, resulting in small 8 % degradation for the stimuli processed with individualized HRTFs catalogues.

Table 2: *Average distance responses in relative scale (100 corresponds to the most distant percept)*

HRTFs	All	G-group	B-group
Ind.	61	72	50
KEMAR	58	58	58

The elevation perception measure was strongly biased by the high-frequency contents of the sound used for the quality-test. Sound with such characteristics is usually perceived as being located higher than its actual position [9]. This was clearly seen from the subject responses; more than 70 % of source positions were judged as being located above the horizontal plane. In general, both groups showed smaller deviations in the sound height judgments when individualized HRTFs were used. At the same time, the B-group showed roughly 3-times larger deviations in the answers to this measure compared to the G-group.

4.2. Main effects: Ego-motion and Presence

All dependent variables were submitted to separate 2(HRTF) x 2(Number of Sources) x 2(Velocity) ANOVAs. First, the analysis using the binary ego-motion measure yielded no statistically significant differences. The analyses of intensity and convincingness showed no significant main effects for HRTF's or velocity, but in both instances a significant main effect of number of sources was found. The means for the intensity measure were 20.5 (1 source) vs. 31.0 (3 sources), $F(1, 11) = 4{,}29$, $p<.05$. Similarly the means for convincingness were 22.2 (1 source) vs. 31.3 (3 sources), marginally significant for a $F(1,11) = 4.13$, $p =.06$. No other effects were significant for these measures.

For the presence ratings a significant main effect of HRTF was found where the individualized HRTFs yielded higher presence ($M = 61.8$) than did the generic KEMAR HRTFs ($M = 57.8$), $F(1,11) = 5.43$, $p<.05$. No other main effects or interaction effects reached significance.

Stimuli with auditory distracters affected only the presence ratings. While almost no effect was found for the stimuli containing multiple sources: $M = 58.3$ (with distracters) vs. $M = 59.2$ (without distracters), presence ratings for the single rotating sound source were higher for the

stimulus with distracters (M = 58.8) compared to non-distracter condition (M = 52.9). In addition verbal probing showed a clear difference in overall judgment of the presented sound scene and the trajectories of the virtual sound sources. In the presence of distracters less distorted trajectories were perceived.

4.3. Subgroup differences in ego-motion perception

HRTFs quality-test results presented in Tables 1 and 2 motivated a sub-group analysis of the data from the main experiment. Figure 3 shows the results from the statistical analysis of 3 direct measures used in the main experiment, where 4 pairs of stimuli were used for non- and individualized HRTFs catalogues comparison. However, since the sample size was too small to allow for parametric statistical analyses, only trends are reported here.

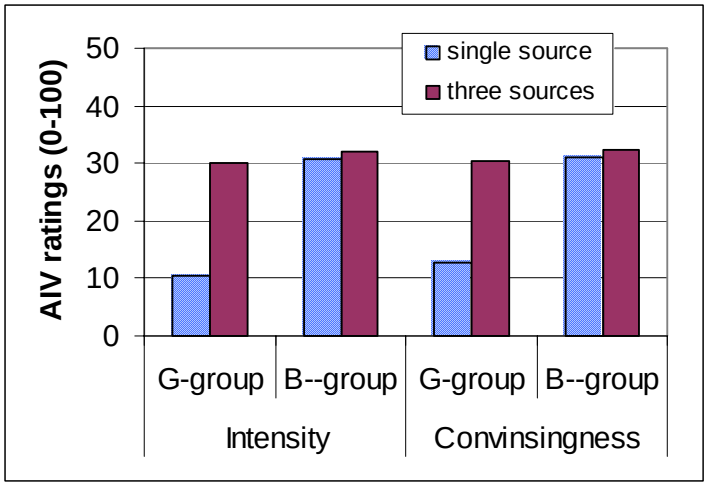


Figure 3: *Effects of concurrent sound sources number on intensity and convincingness average ratings in the sub-group analysis*

As was shown in [8], multiple sources positively affect presence ratings and this can be seen for both subgroups in the table. The same trend continued for intensity and convincingness of ego-motion, but in the B-group the difference is negligible. On the contrary, the G-group showed clear discrimination in ratings when the number of sources presented in stimuli were increased.

5. Discussion

The major finding in the present study was that stimuli processed with individualized HRTFs catalogues resulted in a significant increase of presence ratings as compared to stimuli processed with generic HRTFs. Several other lines of research have independently showed that

individualized HRTF increase spatial perception and spatial abilities [16, 24] or that more spatialized sound increase the sense of presence [8, 13-14]. However, to the best of our knowledge, this is the first study to show a direct link between individualized HRTFs and spatial auditory presence.

In addition, the results for ego-motion (intensity and convincingness ratings) showed consistency with the previous findings reported in [8] - a higher number of concurrently rotating sources and a higher rotational speed increased these ratings.

The sub-group differences shown for the HRTF quality-test (Tables 1 and 2) and the results of the main experiment (Figure 3) can be explained either by the errors that occurred during the individualized HRTFs measurement procedure or by the subjects' auditory localization abilities. Localization performance varies between the individuals and terms "poor" or "bad" localizers are used in the literature, e.g. [9, 25].

In general, the results from the HRTF quality-test were influenced by several factors. First, participants were not trained to perform localization tasks in previous experiments and the quality-test procedure did not include a training period. Second, utilization of an ecological sound as a stimulus might bias the judgments of the participants. More work has to be done for designing fast and reliable procedures for evaluation of HRTFs catalogues' quality.

Results presented in Table 1 showed that the front-back confusion ratio was significantly increased for G-group subjects when using non-individualized HRTFs. However, no such difference was found for B-group. It is known that the performance of good localizers degrades when using bad localizers' HRTFs [9]. The opposite effect, when bad localizers improve their abilities using other person HRTFs catalogues, has not been fully evaluated [9]. Larger deviations in rated source heights found in the B-group for both individualized and generic HRTFs catalogues suggest the influence of individual performance. Figure 3 presents the last evidence for difference in the sub-groups' performance: for intensity and convincingness ratings no discrimination between stimuli containing single or multiple sound sources was done by the B-group subjects.

When re-examining the data from the previous experiment [8] for the same participant sub-groups, the trends presented in Figure 3 were not found. Therefore it is more likely that sub-group difference is due to the errors occurred during the HRTFs measurements procedure. However since different stimuli synthesis procedure was used in [8] a direct comparison of the subject responses is not possible and further studies of this finding is needed.

Preliminary tests with 2 stimuli with added clicks supported the hypothesis that adding distracters to the auditory VE might influence overall perception of the sound scene and decrease influence of artifacts caused by non-individualized HRTFs utilization. Specially designed experiment

should shed further light on the effects of divided attention on quality judgments of VEs.

6. Conclusions and future work

There are several conclusions from this initial investigation. First, it was found that individualized HRTFs increase presence ratings. Second, the results were consistent with the previous results reported in [8], where the number of sound sources influenced both presence ratings and ego-motion experiences. Third, inter-group differences were found within the subjects, which were more likely caused by the errors occurred during the fast measurement of individualized HRTFs catalogues. Participants from the group with poor localization performance showed no discrimination in intensity and convincingness ratings for the number of presented sound sources. Finally, it is important to note that stimuli processed with generic (KEMAR) HRTFs also induced ego-motion regardless to the lowered rendering quality of spatial scene.

The authors are planning to test their findings in sub-group differences using a higher number of participants for reliable statistical analysis. Modification of existing methodology of both measuring and evaluating of individualized HRTFs catalogues is also planned for upcoming work. Influence of distracters on presence ratings and illusory ego-motion sensation is another topic for the follow up experiments.

Acknowledgements

This work was supported by the EU POEMS project (IST-2001-39223). The first author of this paper would like to thank for the support from Alfred Ots' scholarship foundation.

References

[1] M. Lombard, T. Ditton. At the heart of it all: the concept of presence. Journal of Computer Mediated Communication, 3(2), 1997.

[2] J. Schulte-Pelkum, B.E. Riecke, M. von der Heyde, H.H. Bülthoff. Circular vection is facilitated by a consistent photorealistic scene. In Proceedings of Presence 2003, Aalborg, Denmark. October 2003.

[3] J.R. Lackner. Induction of illusory self-rotation and nystagmus by a rotating sound-field. In Aviation, Space and Environmental Medicine, 48(2), 129-131. 1977.

[4] Y. Suzuki, S. Sakamoto, J. Gyoba. Effect of auditory information on self-motion perception. In Proceedings of ICA, Tokyo, Japan. 2004.

[5] S. Sakamoto, Y. Osada, Y. Suzuki, J. Gyoba. The effects of linearly moving sound images on self-motion perception. Acoust. Sci. & Tech, 25(1), 100-102. 2004.

[6] B. Kapralos, D. Zikovitz, M. Jenkin and L. R. Harris. Auditory cues in the perception of self-motion. In 116th AES convention, Berlin, Germany. May 2004.

[7] EU-Project POEMS (Perceptually Oriented Ego-Motion Simulation) URL: http://www.poems-project.info/

[8] P. Larsson, D. Västfjäll, M. Kleiner. Perception of self-motion and presence in auditory virtual environments. Submitted to Proceedings of Presence 2004, Valencia, Spain. October 2004.

[9] D.R. Begault. 3D sound for virtual reality and multimedia. Academic Press. 1994.

[10] CATT-Acoustics v 8.0 (Computer software), Gothenburg, Sweden. URL: http://www.catt.se

[11] J. Blauert. Spatial hearing. MIT Press. Cambridge, rev. edition. 1997.

[12] E. M. Wenzel, M. Arruda, D. J. Kistler, F. L. Wightman. Localization using nonindividualized head-related transfer functions. J. Acoust. Soc. Am., vol 94, 111–123. 1993.

[13] C. Hendrix, W. Barfield. The sense of presence within auditory virtual environments. In Presence – Teleoperators and Virtual Environments, 5(3), 290-301. 1996

[14] J. Freeman, J. Lessiter. Hear there & everywhere: the effects of multi-channel audio on presence. In Proceedings ICAD 2001. July 2001

[15] E.H.A. Langendijk, D.J. Kistler, F.L. Wightman. Sound localization if the presence of one or two distracters. J. Acoust. Soc. Am., vol. 109(5), 2123-2134. 2001.

[16] H. Møller. Fundamentals of binaural technology. Applied Acoustics, vol. 36, 121-218. 1992.

[17] D. Hammershøi, H. Møller. Methods for binaural recording and reproduction. Acta Acustica united with Acustica, 88 (3), 303-311. 2002.

[18] F.L. Wightman, D.J. Kistler, S.H. Foster, J. Abel. A comparison of head-related transfer functions measured deep in the ear canal and at the ear canal entrance. In Abstracts of the 17th Midwinter Meeting of the Association for Research in Otolaryngology, 71. 1995.

[19] S. Müller, P. Massarani. Transfer function measurement with sweeps. J. Audio Eng. Soc. vol. 49 (6), 443-471. 2001.

[20] A. Kulkarni, S.K. Isabelle, H.S. Colburn. Sensitivity of human subjects to head-related-transfer-function phase spectra. J. Acoust. Soc. Am., vol. 105(5), 2821-2840. 1999.

[21] K. Hartung, J. Braasch, S. Sterbing. Comparison of different methods for the interpolation of head-related transfer functions. In AES 16th Int. Conf. on Spatial Sound Reproduction, 319-329. 1999.

[22] J. Breebart. Modeling binaural signal detection. PhD thesis. Technische Universiteit Eindhoven. 2001.

[23] Home Audiometer (Hearing test software)
URL: http://www.esser.u-net.com/homeaudiometer.html

[24] T. Djelani, C. Pörschmann, J. Sahrhage, J. Blauert. An interactive virtual-environment generator for psychoacoustic research II: Collection of head-related impulse responses and evaluation of auditory localization. ACUSTICA acta acustica, vol. 86, 1046–1053. 2000.

[25] E.M. Wenzel, F.L. Wightman, D.J. Kistler, S.H. Foster. Acoustic origins of individual differences in sound localization behavior. J. Acoust. Soc. Am, vol. 84, S79. 1988.

Paper B*

Travelling without moving: Auditory scene cues for translational self-motion.

Aleksander Väljamäe, Pontus Larsson, Daniel Västfjäll and Mendel Kleiner

Published in
Proceedings of International Conference of Auditory Displays, ICAD 2005,
Limerick, Ireland, July 2005

---

*The layout has been changed to fit the overall thesis style

TRAVELLING WITHOUT MOVING: AUDITORY SCENE CUES FOR TRANSLATIONAL SELF-MOTION

Aleksander Väljamäe

Department of Signals and Systems,
Chalmers University of Technology,
SE-41296, Göteborg, Sweden.
sasha@s2.chalmers.se

Pontus Larsson, Daniel Västfjäll, Mendel Kleiner

Department of Applied Acoustics,
Chalmers University of Technology,
SE-41296, Göteborg, Sweden.
{pontus.larsson,daniel, mk}@ta.chalmers.se

Abstract

Creating a sense of illusory self-motion is crucial for many Virtual Reality applications and the auditory modality is an essential, but often neglected, component for such stimulations. In this paper, perceptual optimization of auditory-induced, translational self-motion (vection) simulation is studied using binaurally synthesized and reproduced sound fields. The results suggest that auditory scene consistency and ecologically validity makes a minimum set of acoustic cues sufficient for eliciting auditory-induced vection. Specifically, it was found that a focused attention task and sound objects' motion characteristics (approaching or receding) play an important role in self-motion perception. In addition, stronger sensations for auditory induced self-translation than for previously investigated self-rotation also suggest a strong ecological validity bias, as translation is the most common movement direction.

1. Introduction

Creating a sense of illusory self-motion for the end-user of a Virtual Environment (VE) is crucial for many Virtual Reality (VR) applications, e.g. various motion simulators. Illusory self-motion, also referred to as vection, can be described as a sensation of actual movement relative to a stable surrounding environment. While a large body of research has been focused on vection elicited by visual stimuli (e.g. [1] and references therein), research on auditory induced vection (AIV) has received little attention until recently [2] cf. [3], [4] and [5] cf. [1], [6-11].

Auditory induced self-motion can be elicited using moving sound fields, either real (e.g. loudspeaker array presentation) or virtual ones (typically headphone reproduction). Binaural technology provides the most flexible way of creating virtual sound environments, where non-spatialized ("dry") sounds are convolved with pre-measured Head-Related Transfer

Functions (HRTFs) of the corresponding spatial positions [12]. Binaural synthesis of moving soundfields is a computationally demanding task, where various factors have to be taken into account, e.g. spatial resolution of HRTFs catalogue, room acoustics model, and sound sources characteristics. Therefore, finding the auditory cues which are most instrumental in inducing vection is an important step towards perceptually optimized VR motion simulators.

In our previous experiments [10], [11] on rotational AIV, we decided to concentrate on the ideas of ecological acoustics which studies sound perception from the perspective of every-day listening experiences [13]. We hypothesized that the type of sound source is an important parameter when studying AIV. Unlike artificial sounds (e.g. pink noise), ecological sound sources can be classified by a listener into spatially "still" (e.g. a church bell) or "moving" (e.g. footsteps) categories. A major finding in our previous research is that the experience of self-motion is significantly higher for sound fields with sound sources from the "still" category representing clearly recognizable acoustic landmarks. The higher speed of the sounds' rotation and the larger number of sources also positively affect the AIV ratings [10], [11].

In the current study we address translational AIV and present our first experiment with artificial stimuli containing noise and tonal sound. Taking into account the findings in [8] and [9], we investigate how factors from ecological acoustics, motion metaphors (e.g. engine sound) and selected attention affect AIV. The results of this experiment serve a basis for the follow-up studies with ecological sound, which is reported in [14].

2. Auditory motion perception

Knowledge on auditory motion perception is essential for determining salient auditory cues contributing to auditory-induced vection. The mechanism of auditory motion perception is a complex phenomenon with many parameters involved and it remains to be an active area of research. Recent evidence from brain studies show that a specific "movement-sensitive" area in auditory cortex is most likely to exist (e.g. [15] and references therein) thus indicating separate mechanisms for stationary and moving sounds localization.

Three main cues for discrimination of auditory motion are intensity, binaural cues and the Doppler effect. Intensity cues arise from the changes in sound pressure level emitted by a moving sound source. Binaural cues reflect the interaural time and level differences (ITD and ILD) at listener's ears. The Doppler effect results in perceived frequency shifts in the case of motion between a sound source and a listener. Lufti and Wang [16] thoroughly examined these three cues and showed that for sound object

velocities below 10 m/s, intensity and binaural cues were the most instrumental in providing travelled distance information. The Doppler shifts were pre-dominant for sound object velocity and acceleration judgments. For a higher velocity (50 m/s), the Doppler shift tended to dominate in all discrimination tasks. It is important to note, however, that cue dominance depended not only on the task but also varied between tested individuals.

The intensity cue was found to be dominant for travelled distance perception in an earlier study by Rosenblum et al. [17], where this finding was explained from an ecological acoustics perspective. Recently it has been shown that continuous intensity changes can elicit illusion of pitch shift, which are roughly four times larger than the actual frequency shift caused by the Doppler shift [18]. It was also shown that continuous intensity changes in sound stimuli can solely lead to an illusory pitch shift. The authors concluded that Doppler shift perception in everyday listening is almost entirely driven by the intensity cue [18].

The intensity cue dynamics give rise to several secondary cues contributing to the auditory motion perception. When a sound source passes a listener, a "point of closest passage" is clearly marked by the highest intensity peak [18]. Intensity can be also used for a time-to-arrival estimation by tracking the intensity change rate called an acoustic tau [19], [20]. However, the acoustic tau and the auditory motion parallax (auditory equivalent to visual parallax) have a minor impact on the auditory motion perception compared to stronger cues as intensity and reverberation [21].

Studies on sound intensity perception revealed another interesting effect related to the perception of approaching or looming sound sources. In his recent study Neuhoff [22] showed asymmetry in perception of rising and falling intensity where continuous intensity increase resulted in a stronger perceived loudness change compared to the same intensity fall. This sound "looming" effect also resulted in a different perception of distances traveled by approaching or receding illusory sound sources, which were simulated by rising and falling intensity. Several concurrent studies corroborated the fact that looming sounds have perceptual and behavioural priority and that sounds perceived as approaching have greater biological salience than receding ones [23].

An alternative way of auditory motion perception mechanism was suggested in "snapshot hypothesis" by Grantham in [24], where he proposed that, instead of direct perception of sound objects' velocity, listeners base their judgement on the total distance travelled by these objects. Recent findings by [25] suggest that both direct perception of motion cues and displacement detection can take place. In this light, the effects of attenuation of high frequencies due to air absorption on sound distance perception [26] can play a role in sound motion judgements. Distance perception also depends on the type of sound source and on a listener's familiarity to it [12]. More accurate results in judgements on

travelled distance have been found for ecological sounds and sounds that are within listeners' reach [27].

The acoustic environment plays an important role in auditory distance perception, especially for indoor conditions where the ratio between direct and reflected sound is known to be one of the most salient cues [28]. Rosenblum et al. [29] uses the term "echolocation" for the human ability to track the echoic changes while moving in a reverberant environment. Knowledge on a sound source directivity pattern may also play a role in determine a source or self-motion [30].

To summarize, the presented information show that the perception of auditory motion can be influenced by the ecological context of the surrounding soundscape. It supports the suggestion by Popper and Fay [31] that the main function of the auditory localization mechanism may be to provide an input to the listener's perceptual model of the environment rather than exact estimates of sound sources' location and trajectories.

3. Working hypotheses

In this experiment, a context-free scenario based on artificial sounds (noises and tones) was used. The hypotheses listed below were tested in the experiment and the results were used for a refinement of the experimental methodology in the follow-up experiments with ecological sounds in [14].

H1 – **Looming sounds.** According to the looming effect, sounds with falling and rising intensity are perceived differently [22]. We hypothesize that if approaching sounds are biologically more salient than receding ones, simulation of moving towards a sound will give a stronger AIV sensation than scenarios where the listener is leaving the sound.

H2 – **Acceleration effect.** The ability of discriminating between the sound sources moving with a constant velocity or acceleration ones is an interesting but rarely addressed question in auditory motion research. Such ability was originally shown in [32] and more evidence was indirectly given by the Doppler effect study in [18], where illusory pitch shifts were found to be dependent on the intensity changes mimicking either accelerating or constant velocity sound source. As the acceleration is a necessary component for the human vestibular system in the perception of self-motion, we believe that the accelerating sounds will have a stronger effect on AIV than the sounds with a constant velocity.

H3 – **Focused Attention.** We hypothesize that in a focused attention task, where participants have to concentrate on intensity changes in one specific sound, the AIV experience will be negatively affected. It is known that monitoring the changes in bodily orientation from vestibular or visual information requires a significant degree of attention or cognitive load [3], [33], [34]. Therefore, distracting the listener from auditory streams which

are providing salient information for AIV can negatively affect the self-motion sensation.

H4 – **Auditory motion detection threshold.** We hypothesize that participants experiencing AIV will be slower in detecting a sound object motion. This argumentation was inspired by findings in [35], which showed elevated thresholds for object motion detection when experiencing visually induced vection. Similar experiments in auditory domain were suggested in [36] but, to the best of the authors' knowledge, where never reported before. Testing this hypothesis was done in conjunction with stimuli used for testing H3 and our aim was to get a first insight in developing a proper methodology for further studies on this effect in the auditory domain.

4. Method

Experiment was conducted using virtual auditory space and the stimuli were synthesized in MatlabTM using a catalogue of generic HRTFs. Binaural synthesis was used to simplify the experimental setup, which was intended to resemble an optimized, cost-effective VR motion simulator. Moreover, the authors were interested in how the AIV will be affected by imperfections in spatial sound rendering due to generic HRTFs.

The generic HRTFs catalogue was measured from KEMAR mannequin using the procedure described in [11]. For stimuli synthesis only one horizontal plane (-4 degree elevation) with a 5 degree resolution was used. In this experiment no acoustic environment rendering was applied.

4.1. Stimuli

In the stimuli synthesis two different initial distances to the sound sources were used - "distant", simulating approaching sound objects and "close", simulating the receding sound objects. Figure 1 illustrates 3 types of excerpts for eliciting translational AIV in the forward direction. Further in the text these 3 stimuli types will be referred as "distant" (approaching), "close" (receding) and "mixed" (one approaching one receding). These stimuli types contained two sound sources with additional "anchor" sound introduced in some conditions (see Table 1 for stimuli design).

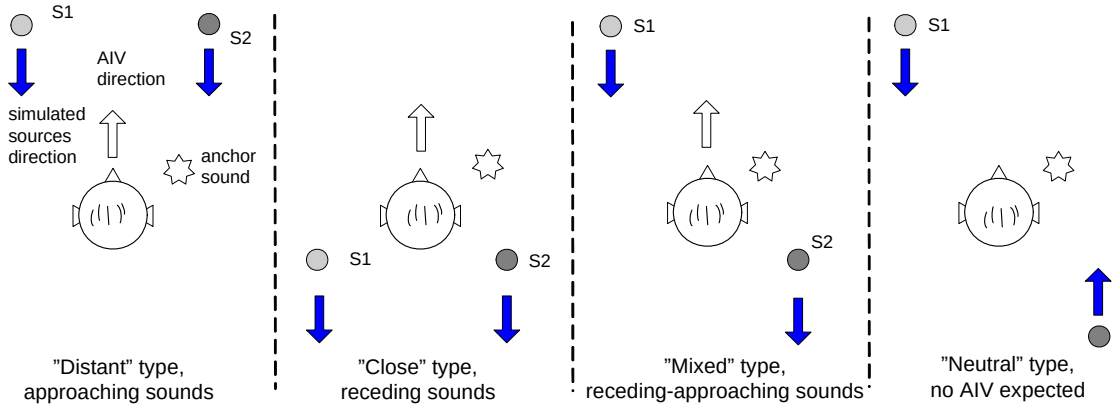


Figure 1: *Stimuli types used in experiment. The arrow in front of the listener indicates expected (AIV) direction, filled arrow indicates the motion direction of the virtual sound objects*.

In addition to the 3 stimuli types in Table 1, four excerpts of "neutral" stimuli type were used in order to provide a baseline for the detection threshold estimation task described in section 4.3. In the "neutral" stimulus the same two sounds objects were moving in the opposite directions (see Fig. 1), and therefore no AIV was expected. Neutral excerpts were always accompanied with the anchor sound.

Table 1: *Experimental design for distant, close and mixed type stimuli (expected AIV direction: forward or backward*

		No anchor		Anchor	
distance	*velocity*	*AIV direction*		*AIV direction*	
		forw.	backw.	forw.	backw.
Distant (approach.)	const. v	x	x	x	x
	accel	x		x	
Close (receding)	const. v	x	x	x	x
	accel.	x		x	
Mixed	const. v	x	x	x	x

In our definition, the "anchor" sound is a sound following moving listener and often caused by him (e.g. sounds of ones own breath, coins in the pocket, engine sound, etc.). The anchor sound was inspired by participants' verbal responses from [10], where footstep sounds elicited an illusion of moving with a crowd. In the case of AIV the anchor sound should be perceived as accompanying listener's illusory self-motion. The anchor sound can be seen as an auditory correspondence to the visual self-avatar, which has proven to be an important component for compelling VE experiences [37]. The effects of such sonic self-avatars on AIV were further studied in [14].

Apart from distance and direction, other parameters varied in stimuli design (Table 1) were: 1) sound sources velocity - constant (v = 1 m/s - pedestrian speed in the city) or increasing (a = 0.012 m/s2 – smallest value used in [9]); 2) direction of sound sources movement (forward and backward), which is opposite to expected AIV direction (see Fig. 1).

All stimuli were approximately 1 minute long, with a difference of several seconds between excerpts with accelerating and constant speed sounds. For the accelerating type, sources' "stationary" phase lasted the first 4 seconds and then the sounds started to accelerate. For the constant velocity type, after the same 4 second stationary phase and 3 second acceleration, sounds speed were constant. A Hann half-window of 0.5 second duration was applied to smooth stimuli on- and off-sets.

Bandlimited pink noise was used for the synthesis of two moving sound sources. Two regions from an idealized critical band filter bank [38] were used - 510-920 Hz (6th -8th band) and 1270-2000 (11th-13th band) – in order to maximize auditory stream separation. This frequency range was chosen to restrict distortions caused by the use of a generic HRTFs catalogue, i.e. a limited spatial resolution and HRTFs' mismatch to participants' ears. As ITD dominates spatial sound localization below 1600 Hz, this binaural cue was a main source for such distortions in this experiment and the effects of ILD and spectral cues mismatch were minimized [12].

The anchor sound was represented by a frequency modulated tone with 300 Hz carrier and 20 Hz modulation (modulation index = 0.5). In order to help participants to focus their attention on the anchor sound, it was appearing 3 seconds before the two moving sound sources playback. After 33 seconds of "still" period, the anchor sound intensity started to decrease mimicking a receding sound source with an acceleration of 0.1 m/s. This acceleration was specifically chosen for the detection threshold estimation, where a slow fading would spread detection times over a longer period and thus help timing data post-processing. The instructions for detection of intensity fall are described in the procedure section. The anchor sound was placed on the listener's right side (see Fig.1) with the intensity subjectively corresponding to a proximal position (1-2 meter range).

Intensities of moving sound sources were changing according to the inverse square law (6 dB level change per distance doubling). However, for the anechoic conditions the distance perception depends on various factors (type of sound source, listener expectations and knowledge, etc.) and sometimes higher intensity changing rates like 9 or 12 dB per distance doubling are used for a more subjectively adequate simulation [26]. In the experiment, the following assumptions about initial distances to the moving sound sources were used 1) distant type – 50 meters from a listener for constant velocity sounds and 20 meters for accelerating sounds (main motivation for this difference was to have the "point of closest passage" at a

similar time in all stimuli) 2) close type – 1 meter 3) mixed type - 5 meters for nearby source and 50 meters to distant (see Fig. 1).

4.2. Measures

To assess the AIV, two direct verbal measures were used in this experiment: vection intensity and convincingness of vection. Vection intensity corresponded to the level of subjective sensation when experiencing self-motion. On the convincingness scale participants had to report how convinced they were of having been moving in the direction of the experienced self-motion. It should be noted that the convincingness and intensity ratings are often highly correlated. Ratings of both measures were given on a 0-100 scale.

Apart from the direct measures listed above, an indirect binary measure, reflecting the number of ego-motion experiences, was used (participants were asked to verbally indicate the direction of self-translation). While an onset time for vection experience is often used in experiments with visual stimuli, in the present study the onset time was not measured since previous experiments [10] on auditory-induced vection indicated that this measure showed large inter-individual variance.

According to the hypothesis of change in detection threshold when experiencing AIV (see H4), the reaction times for intensity change in the anchor sound were monitored during the experiment (see procedure for further details).

4.3. Procedure

In the first experiment 24 naive participants (11 male) with a mean age of 24 (SD 3.8) took part. Before coming to the experiment all participants filled in two web-based questionnaires on mental imagery [39] and need for cognition [40].

The experiment was conducted in a special laboratory setup with black curtains surrounding the participant (see Fig. 2). Stimuli were played back with Beyerdynamic DT-990Pro circumaural headphones. Taking into account the experimental procedure in [6] and our previous experiments in [10] and [11], special measures have to be taken in order to amplify AIV sensation. During current experiment, participants were blindfolded and seated on a chair mounted on a wheeled platform coupled with a footrest as shown in Figure 2. The fact that participants knew that the platform could potentially move was intended to increase the convincingness of the simulation setup [10].

Figure 2: *Laboratory setup: a participant sitting on a chair mounted on a wheeled platform coupled with a footrest.*

Participants were instructed verbally and a short training session was performed before the experiment start (2 stimuli presented). For the anchor sound stimulus, participants were asked to concentrate on the tonal sound and to stop the playback verbally when they heard anchor sound fading or moving away. The reaction times for this detection task were monitored on the basis of listener's verbal response. After the stop in the stimulus playback, participants had to give ratings on intensity and convincingness of self-motion if such sensation was perceived. They also were asked about the direction of the perceived AIV. Sound excerpts without anchor sound were presented in full length and were followed by the same questionnaire as for the other excerpts. Stimuli were presented in randomized order with small breaks after each 6 excerpts. Apart from the verbal responses to the questionnaire, verbal probing was done by the experiment leader. After completing the experiment, participants were debriefed, thanked and paid for their participation.

5. Results

The experiment results from 2 (anchor) x 3 (distance) ANOVA showed no significant effects for the mixed type stimuli and in next subsections only 2 distance types (distant and close) were used for analysis. Results presented in the next subsections did not correlate neither with the need for cognition and mental imagery scores or gender of participants.

It is noteworthy to mention that 9 from 24 participants were sometimes reporting auditory-induced sensation of translation along vertical axis (e.g. forward-down).

5.1. Binary vection measure

One of the measures for the experienced AIV perception is a binary vection measure, which shows how many participants experienced vection for a particular stimulus type (Fig. 1). Results from this experiment showed that inducing self-translation sensation by purely auditory means is more successful compared to self-rotation experiments in [10] – current range of 33-79% (8-18 from 24) is higher compared to rotational binary vections range 23-50% (6-13 from 26).

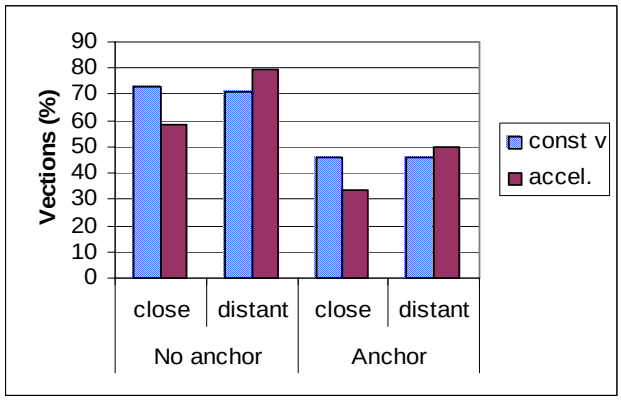


Figure 3: *Reported binary vections (100% correspond to stimuli where all 24 participants experienced vection). "Close" type corresponds to receding sounds and "distant" to approaching sounds*.

Figure 3 presents the percentage of experienced binary vections for 8 stimuli types where results for the constant velocity stimuli are averaged for backward and forward directions (see Table 1 and further discussion in section 5.3). It can be seen that higher amount of binary vections are experienced in the no anchor sound condition. Additionally, an asymmetric pattern emerges for the velocity parameter in close and distant conditions. While the stimuli with the constant velocity give almost the same ratings for binary vection - 73% (close) vs. 71% (distant), accelerating sounds show a large shift in favour of the distant, approaching type excerpts (58% vs. 79%, see also Table 3). This asymmetry might be accounted for the difference in the sound sources velocity for the period when sounds are in listeners' proximity, as previous findings in [10] showed that higher sound velocity positively affects AIV. However, this is only true for the close type condition, where receding sounds are much slower for the excerpts with accelerating stimulus, as in the distant type the accelerating sounds achieve speed of only 0.7 m/s when passing by the listener. Taken together, a first evidence for the difference in AIV sensation induced by accelerating or

constant velocity sounds was found and this effect was further investigated in the follow-up experiment reported in [14].

Furthermore, simple chi-square analysis shows that binary vection measures for both close and distant stimuli types in the no anchor condition differ significantly from what may be expected from chance, $\chi 2(1) = 4.54$, $p.<.05$ (close) and $\chi 2(1) = 8.91$, $p.<.01$. No such difference was apparent in the anchor condition. To further corroborate these results a Friedman rank-test was performed on the vection data (vection coded as 1, no vection coded as 0). The mean ranks were 2.16 (anchor, close), 2.25(anchor, distant), 2.70 (no anchor, close) and 2.89 (no anchor, distant). The test statistic for this analysis was highly significant, $\chi 2(3) = 18.51$, $p.<.001$.

5.2. Intensity and convincingness

Two separate ANOVAs with factorial design of 2 (anchor) x 2 (velocity) x 2 (distance) were conducted for vection intensity and convincingness. The same trends as for binary vection responses were found for anchor and distance parameters. For vection intensity the main effect of anchor was significant $F(1, 23) = 29.16$, $p.<.001$, means 18.6 (anchor) vs. 35.6 (no anchor). No other effect reached significance.

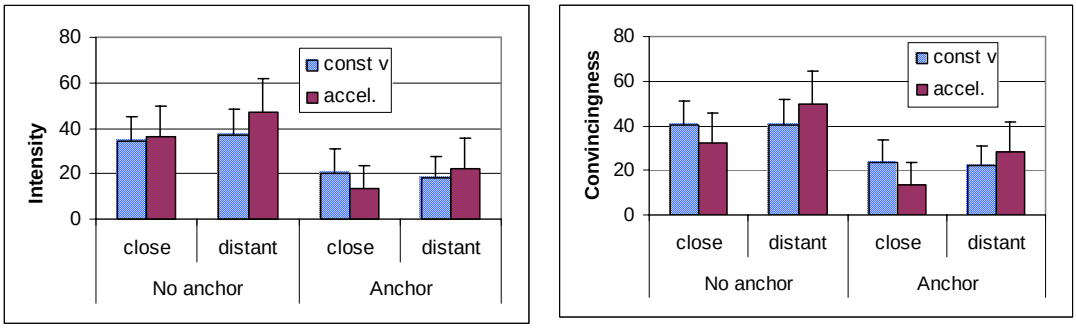


Figure 4: *Intensity (left) and convincingness (right) ratings with 95% confidence interval (upper bound). "Close" type corresponds to receding sounds and "distant" to approaching sounds*

Similar results were found for the convincingness data (Figure 4, right): the main effect of anchor was significant $F(1, 23) = 28.86$, $p.<.001$, mean 21.7 (anchor) vs. 40.6 (no anchor). In addition, the main effect of distance was significant, $F(1, 23) = 3.99$, $p.<.005$, with a higher mean (35.1) for the distant than the mean (27.3) for the close condition. No other effects or interactions reached significance.

5.3. Front-back reversals

The experiment results showed that the binaural stimuli designed specifically for AIV in forward/backward often did not give the expected self-motion direction sensation. The participants often misinterpreted two moving sounds direction, i.e. front-back reversals occurred [26] where created virtual sources were localized at the opposite to expected side. Due to the high rate of front-back confusions – approximately 30% of the total number of reported AIV, we decided to combine forward-backward stimuli pairs (see Table 1) into one and to discard the direction parameter in the results analysis.

Three listeners perceived the motion of the virtual sources only in the frontal hemisphere thus reversing some parts of perceived motion trajectories (see Fig 6a). This effect was observed from the sudden change in perceived AIV direction in the participants' reports. As no information about perceived movement of the sound objects was asked, one could assume that front-back confusions occurred also in the cases when no AIV was reported.

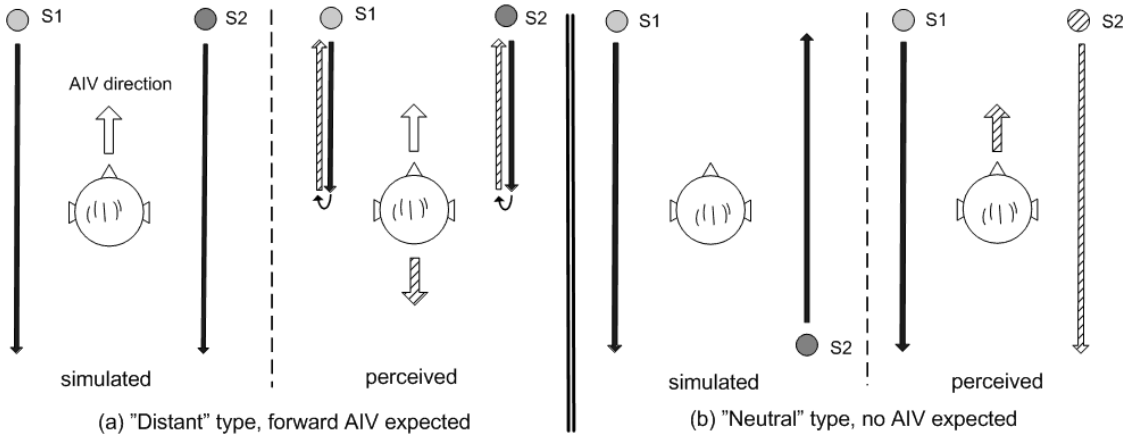


Figure 6: *Examples of back-front reversals (dashed lines represent perceived "reverse" trajectories and resulted AIV direction): a) sound objects motion is perceived only in front of the listener leading to switching of simulated forward AIV to backward direction; b) sound objects motion direction is coupled leading to undesired AIV responses.*

It is important to note, that usually there is a significant asymmetry between front-back (FB) and back-front (BF) reversals with stronger tendency for FB confusion (roughly 3 times higher than BF) [26]. On the contrary, in our experiment this difference was not observed and almost equal numbers of reversals in both directions were found for most of the participants. Moreover, on the average BF reversals were even more

frequent – 25% vs. 15%. Unfortunately, the neutral type stimuli were also often - 50% of the times on average - inducing AIV as shown on Fig 6b. Analysis of these "undesired" self-motions reports also showed equal occurrence of both reversal types.

5.4. Detection threshold estimation

Vection experiments in the visual domain showed that the self-motion sensation affects object motion perception thresholds [35]. The aim of the detection threshold task in this experiment was to take the first steps in the validation of this effect for the auditory domain. Unfortunately, the problem of front-back reversals also occurred with the neutral type stimuli (see Fig. 1), were two sound sources were perceived as moving in one common rather than in opposite directions, which in turn lead to unwanted AIV sensations. Some participants' responses (for roughly half of the participants) allowed for detection times comparison between cases with or without AIV experience. An indication of the threshold increase for reported self-motion cases was observed, however, this trend did not reach significance.

This part of the experiment gave us a first insight into the methodology for a new subjective corroborative measure of AIV, which might be based on auditory motion detection threshold estimation. In the refined experimental design several parameters have to be more carefully selected: 1) precise measurement of detection times 2) truly "neutral" stimuli not eliciting self-motion but resembling the spectral content of AIV oriented stimuli; 3) intensity change rate if it is used a motion cue for the detection task. Apart from the tasks involving distance and velocity judgments for auditory motion, sound localization tasks might be also a proper substitute to this methodology.

6. Discussion and conclusions

This study aimed at identifying the auditory cues important for translational self-motion sensation. One major finding was that a focused attention task significantly reduced AIV. Moreover, asymmetry in AIV was found between stimuli containing approaching or receding sound sources, where approaching sounds had marginally stronger impact on AIV. The current results also show that auditory induced self-motion is a more reliable phenomenon for simulated translational movements than for rotational movements [10], and that only a set of minimal acoustic cues is sufficient for successful translational AIV simulation.

The finding that a selected attention task reduces AIV ratings might be biased by the difference in the procedure for stimuli conditions with or

without the anchor sound. In the anchor condition stimuli playback was interrupted which could have affected participants' AIV ratings. However, continuous stimuli playback was used in our follow-up study [14] with the similar experimental methodology, and we found similar trends. We therefore suggest that the self-motion sensation can be negatively affected if the listener is distracted from the auditory stream which provides salient information for AIV. This reasoning is in line with findings from vestibular [33] and visual [34] self-motion research.

The result of asymmetry in translational AIV for the distant (approaching) and close (receding) type stimuli supports our hypothesis that looming sounds might increase translational AIV experience. The perception of looming objects is more biologically salient than for receding ones and evidence for this effect has been found both in auditory only [23] and audio-visual domain [41].

On the other hand, we have not found any systematic gender differences in the perception of the looming effect as was recently presented by [42]. Taking into account that the looming effect is significantly stronger for tonal than for noise sounds [22], the noise stimuli used in the present experiment could have prevented gender differences. As was suggested by [22], tonal sounds are more likely to represent separate sound objects or acoustic landmarks; on the contrary broad-band noise usually represents a surrounding environment with multiple sound sources. The looming effect perspective can bring to a new understanding of the results presented in [8], where approaching noise stimuli were found to be more instrumental for eliciting self-motion than the receding ones.

An alternative explanation for the asymmetry found in AIV responses for approaching and receding conditions is the influence of the "point of closest passage". In the distant condition the sound sources were passing by the listener, which was not the case for the receding sounds (see Fig. 1). The "point of closest passage" might be an important component for the self-motion sensation and follow-up translational experiments reported in [14] give further evidence to salience of this cue for AIV. When the experimental procedure was changed and participants had to stop the sound playback when experiencing AIV, most of the times the self-motion sensation was built-up at the time of the "point of closest passage".

In the current study we do not find a significant difference for the AIV between the sound objects moving with acceleration or constant velocity. The previous study in [10] showed that the higher sound objects' velocities and the higher number of sound sources were more instrumental for rotational AIV. Similar results can be predicted for translational AIV – the recent study in [9] showed that higher values of acceleration were more instrumental for auditory-induced self-motion. It is interesting to note that in the current experiment, the "mixed" type stimulus, where one receding source was separated in time from another approaching source (see Fig. 1),

can be seen as a single source stimulus compared to other conditions with two sounds moving together. Therefore, the lower AIV ratings trend for the mixed stimuli type can be accounted to the lower number of sound objects moving in one direction.

Difference in stimuli with accelerating or constant velocity sound could also cause asymmetry for distant and close types perception, however further study in [14] on this parameter suggests it might be closely related to the looming effect.

In general, the results from our studies show that translational AIV is more easily induced than rotational AIV, even with fewer special measures (blindfolding, wheeled platform but not special instructions, see [10] for more details) applied to achieve self-motion sensations. This is not surprising as translational movement is more common experience from an ecological perspective. Interestingly, even with artificial noises representing sound sources, participants tended to create a specific scene context, e.g. being in the train, metro or driving in a tunnel. Moreover, the almost equal percentage for front-back and back-front reversals suggests that an ecological context can influence the auditory scene perception. Sound localization in binaural synthesis systems has previously been found to be rather asymmetric in favour of sound appearance behind a listener [26]. The reason why such asymmetry was not found in the current results can be explained by the fact that people are simply more used to move in forward direction.

This experiment and the previous findings [10] suggest that the auditory scene consistency and ecological validity plays a crucial role in AIV. In the current experiment we deliberately used only the most salient acoustic cues (sound intensity and ITD) for moving sound fields simulation, and the quality of the rendering was determined by a generic HRTFs catalogue with a relatively low spatial resolution. The results suggest that this reduced level of details in spatial sound rendering can be sufficient for creating a self-motion sensation. This, in turn, would allow allocating sound processing resources for other tasks including, for example, low latency rendering of real-time interaction.

7. Future work

Two follow-up experiments on translational AIV have been conducted and the results will be reported in [14]. These findings suggest that reverberation and auditory scene “spaciousness” might play an important role in AIV, which will be investigated in the future experiments. Participants’ verbal responses show that auditory-induced sensation of translation along vertical axis (e.g. elevator sensation) can take place, which can be an interesting topic for future AIV studies. The refinement of detection threshold

estimation procedure as a corroborative measure of AIV is included in the currently conducted self-motion experiments.

Acknowledgements

The work presented in this paper was supported by the European Community under the FET Presence Research Initiative project POEMS (Perceptually Oriented Ego-Motion Simulation), IST-2001-39223 and the Swedish Research Council (VR project 40499601). The first author of this paper also would like to thank for the support from Alfred Ots' scholarship foundation.

References

[1] G.J. Andersen, "Perception of self-motion: Psychophysical and computational approaches", Psychol. Bull., vol. 99(1), pp. 52-65, 1986.
[2] V. Urbantschitsch, "Über Störungen des Gleichgewichtes und Scheinbewegungen [On disturbances of the equilibrium and illusory motions]", Zeitschrift für Ohrenheilkunde, vol. 31, pp. 234-294, 1897.
[3] T. Mergner and W. Becker, "Perception of horizontal self-rotation: multisensory and cognitive aspects", in Perception and Control of Self-motion: Resources for Ecological Psychology, R. Warren and A.H. Wertheim, Eds. Hillsdale, NJ: Erlbaum, 1990, pp. 219-263.
[4] S. yon. Stein, Schwindel (Autokinesis externa et interna) [Vertigo (External and internal self-motion)]. Leipzig: Lessier, 1910.
[5] R. Dodge, "Thresholds of rotation", J. Exp. Psych., vol. 6, pp. 107-137, 1923.
[6] J.R. Lackner, "Induction of illusory self-rotation and nystagmus by a rotating sound-field", Aviat. Space Environ. Med., vol. 48(2), pp. 129-131, 1977.
[7] Y. Suzuki, S. Sakamoto, and J. Gyoba, "Effect of auditory information on self-motion perception", in Proc. ICA, Tokyo, 2004.
[8] S. Sakamoto, Y. Osada, Y. Suzuki, and J. Gyoba, "The effects of linearly moving sound images on self-motion perception", Acoust. Sci. & Tech, vol. 25(1), pp. 100-102, 2004.
[9] B. Kapralos, D. Zikovitz, M. Jenkin, and L.R. Harris, "Auditory cues in the perception of self-motion", in 116th AES convention, Berlin, 2004.
[10] P. Larsson, D. Västfjäll, and M. Kleiner, "Perception of self-motion and presence in auditory virtual environments", in Proc. of Seventh Annual Workshop Presence 2004, Valencia, Spain, 2004, pp. 252-258
[11] A. Väljamäe, P. Larsson, D. Västfjäll, and M. Kleiner, "Auditory Presence, Individualized Head-Related Transfer Functions, and Illusory Ego-Motion in Virtual Environments" in in Proc. of Seventh Annual Workshop Presence 2004, Valencia, Spain, 2004, pp. 141-147
[12] J. Blauert, Spatial Hearing. MIT Press, Cambridge, rev. edition, 1997.
[13] W.W. Gaver "How do we hear in the world? Exploration of ecological acoustics", Ecol. Psychol., vol. 5(4), pp. 285-313, 1993.
[14] A. Väljamäe, P. Larsson, D. Västfjäll and M. Kleiner, "Sonic self-avatars and self-motion in virtual environments", manuscript in preparation

[15] J.D. Warren, B.A. Zielinski, G.R. Green, J. P. Rauschecker, and T. D. Griffiths, "Perception of sound source motion by the human brain", Neuron, vol. 34, pp. 139–148, March 2002.

[16] R. A. Lutfi and W. Wang, "Correlated analysis of acoustic cues for the discrimination of auditory motion", J. Acoust. Soc. Am., vol. 106(2), pp. 919–928, 1999.

[17] L.D. Rosenblum, C. Carello, and R.E. Pastore, "Relative effectiveness of three stimulus variables for locating a moving sound source", Perception, vol. 16, pp. 175–186, 1987.

[18] M.K. McBeath and J.G. Neuhoff, "The Doppler effect is not what you think it is: Dramatic pitch change due to dynamic intensity change", Psychon. Bull. Rev., vol. 9(2), pp. 306-313, 2002.

[19] D.N. Lee, "Getting around with light or sound", in Perception and Control of Self-motion: Resources for Ecological Psychology, R. Warren and A.H. Wertheim, Eds. Hillsdale, NJ: Erlbaum, 1990, pp. 487-505.

[20] B.K. Shaw, R.S. McGowan, and M.T. Turvey, "An acoustic variable specifying time to contact", Ecol. Psychol., vol. 3, pp. 253-261, 1991

[21] J.M. Speigle and J.M. Loomis, "Auditory distance perception by translating observers," in Proc. of the IEEE Symposium on Research Frontiers in Virtual Reality, 1993, pp.92-99.

[22] J.G. Neuhoff, "An adaptive bias in the perception of looming auditory motion", Ecol. Psychol., vol. 13(2), pp. 87-110, 2001.

[23] D.A. Hall and D.R. Moore, "Auditory neuroscience: the salience of looming sounds", Curr Biol., vol. 13(3), R91-3, 2003.

[24] D.W. Grantham, ''Detection and discrimination of simulated motion of auditory targets in the horizontal plane,'' J. Acoust. Soc. Am., vol. 79, pp. 1939–1949, 1986

[25] S. Carlile, and C. Best, C. ''Discrimination of sound velocity in human listeners,'' J. Acoust. Soc. Am., vol. 111, pp. 1026–1035, 2002

[26] D.R. Begault, 3D Sound for Virtual Reality and Multimedia. Academic Press, 1994.

[27] L.D. Rosenblum, A.P. Wuestefeld, and K.L. Anderson, "Auditory reachability: An affordance approach to the perception of sound source distance", Ecol. Psychol., vol. 8(1), pp. 1-24, 1996.

[28] W. Bronkhorst and T. Houtgast, "Auditory distance perception in rooms," Nature, vol. 397, pp. 517-520, 1999.

[29] L.D. Rosenblum, M.S. Gordon, and L. Jarquin, "Echolocating distance by moving and stationary listeners", Ecol. Psychol., vol. 12(3), pp. 181-206., 2000

[30] J.G. Neuhoff, M.A. Rodstrom, and T. Vaidya, "The audible facing angle", J. Acoust. Soc. Am., vol. 109(5), pp. 2376-2377, 2001.

[31] A.N. Popper and R.R. Fay, "Evolution of the ear and hearing: issues and questions", Brain Behav. Evol., vol. 50, pp. 213-220, 1997

[32] D.R. Perrott, B. Constantino, and J. Ball ''Discrimination of moving events which accelerate or decelerate over the listening interval'', J. Acoust. Soc. Am., vol. 93, pp. 1053–1057, 1993.

[33] L. Yardley, M. Gardner, N. Lavie, and M. Gresty, "Attentional demands of perception of passive self-motion in darkness", Neuropsychologia, vol. 37, pp. 1293–1301, 1999

[34] G. Rees, C.D. Frith, and N. Lavie, "Modulating irrelevant motion perception by varying attentional load in an unrelated task", Science, vol. 278, pp. 1616–19, 1997

[35] T. Probst, S. Krafczyk, T. Brandt, and E. Wist. 1984. “Interaction between perceived self-motion and object-motion impairs vehicle guidance”, Science, vol. 225, pp. 536-538, 1984.

[36] A.H. Wertheim, “Motion perception during self-motion: The direct versus inferential controversy revisited”, Behav. Brain Sci., vol. 17(2), pp. 293-355, 1994.

[37] M. Usoh et al., “Walking>virtual walking>flying, in virtual environments”, in Proc. of SIGGRAPH’99, 1999, pp. 359-364

[38] B. Scharf, "Critical bands," in Foundations of Modern Auditory Theory, J. V. Tobias, Ed., Academic, San Diego, 1970 pp. 159–202.

[39] P.W. Sheehan, “A shortened form of Betts' Questionnaire Upon Mental Imagery”, J. Clinical Psych., vol. 23, pp. 386-389, 1967

[40] J.T. Cacioppo, R.E. Petty, and C.F. Kao, “The efficient assessment of need for cognition,” J. Pers. Assess., vol. 48, pp. 306-307, 1984

[41] J.X Maier, J.G. Neuhoff, N.K. Logothetis and A.A. Ghazanfar, “Multisensory integration of looming signals by rhesus monkeys”, Neuron, vol. 43, pp. 177-181, 2004

[42] J.G. Neuhoff and T. Heckel, “Sex differences in perceiving loudness change and auditory looming”, in Proc. of ICAD’04, 2004

Paper C

Vibrotactile enhancement of auditory induced self-motion and presence

Aleksander Väljamäe, Pontus Larsson, Daniel Västfjäll and Mendel Kleiner

Submitted to
Journal of Audio Engineering Society

Vibrotactile enhancement of auditory induced self-motion and presence

Aleksander Väljamäe[1,2], Pontus Larsson[2], Daniel Västfjäll[2,3] and Mendel Kleiner[2]
[1]Division of Communication Systems, Chalmers University of Technology, Sweden
[2]Division of Applied Acoustics, Chalmers University of Technology, Sweden
[3]Department of Psychology, Göteborg University, Sweden
{aleksander.valjamae@s2.chalmers.se, pontus.larsson@ta.chalmers.se,
daniel.vastfjall@psy.gu.se, mendel.kleiner@ta.chalmers.se}

Abstract

The entertainment industry frequently uses vibro-acoustic stimulation, where chairs with embedded loudspeakers and shakers enhance the experience. Scientific investigations of the effect of such enhancers on illusory self-motion (vection) and presence are largely missing. The current study examined if auditory induced vection (AIV) may be further augmented by the simultaneous presentation of additional vibrotactile cues delivered via mechanical shakers and low frequency sound. We found that mechanically induced vibrations significantly increased AIV and presence responses. This cross-modal enhancement was stronger for stimuli containing an auditory-tactile simulation of a vehicle engine, demonstrating the benefits of the multisensory representation of virtual environments.

1. Introduction

Creating a sense of illusory self-motion for the end-user of a Virtual Environment (VE) is important to many Virtual Reality (VR) applications such as motion simulators. Illusory self-motion (also known as vection) can be described as a sensation of actual movement relative to a stable surrounding environment (Anderssen (1986) and references therein). A large body of research has shown that vection is easily elicited by visual stimuli; however, stimulation of other modalities is an essential, but often neglected, component for successful self-motion stimulation. For example, some of our recent studies on rotational and translational auditory-induced vection (AIV) suggest that sound also plays an important role in self-motion perception (see Väljamäe et al., 2005a).

The success of motion simulation can be quantified using human centered methodologies, where end user experiences serve as a basis for uni- and cross-modal optimization of the sensory inputs. Apart from AIV responses, self-reported presence was used to indicate the salience of audio-tactile cues in the presented experiment. The feeling of presence is often

described as a sensation of “being there”, however other definitions exist (see Lombard and Ditton, 1997 for a review). Being “spatially present” in a specific context provides a stable reference frame, and recently it has been shown that spatial presence can be closely related to reported self-motion sensation (e.g. Schulte-Pelkum et al., 2003, Riecke et al. 2005).

AIV can be elicited using moving sound fields, either real (e.g. loudspeaker array presentation) or virtual ones (typically headphone reproduction). Binaural technology provides the most flexible way of creating virtual sound environments, where non-spatialized (“dry”) sounds are convolved with pre-measured Head-Related Transfer Functions (HRTFs) of the corresponding spatial positions (Kleiner et al., 1993). Our previous studies have shown that AIV may reliably be induced by making use of these technologies. In the present paper we extend these findings to include multi-modal stimulation.

Advances in multi-modal self-motion simulator design create a need for a better understanding of how cross-modal interaction effects contribute to self-motion (Harris et al., 2002). Although motion platforms have been studied and used over the last three decades (Pausch et al., 1992), high cost, exploitation complexity and difficulty to mimic vestibular cues create the need for alternative solutions with minimal physical motion involved. A common problem for motion simulators is a sensory conflict between modalities, where mismatches between expected and actual movement information is likely to cause simulator sickness and cyber-sickness (Mccauley & Sharkey, 1992; see also Harris et al. (2002) and references therein). Integration of cross-modal information is one the open questions in the perception research community and situations when different modalities provide conflicting information about the surrounding environment can be seen from different theoretical viewpoints (see Stoffregen and Bardy (2001) for a discussion).

It is well known that our nervous system can adapt to the incongruence between different sensory inputs manifesting plasticity in perceptual organization (e.g. Bach-y-Rita, 1972; Wall & Weinberg 2003). For example, in the case of weightlessness, inadequate information from the vestibular system is suppressed and other senses are taking over its functions (Young & Shelhamer, 1990). Several balance prosthesis applications show that vibrotactile stimulation can effectively substitute an impaired vestibular system (Wall & Weinberg, 2003). Moreover, recent findings from research on multisensory integration demonstrate an inhibitory visual-vestibular interaction in the vestibular cortex area. For example, it has been shown that visually induced vection deactivates regions in the vestibular cortex area, which in turn suggests that the multisensory interaction patterns during self-motion are highly dependent on the inter-sensory stimulation conditions (see Dieterich et al., 2003).

In our view, a perceptually oriented design of a multi-sensory self-motion simulator should aim at creating the illusory percept which would override the vestibular cues signaling absence of physical motion. Increasing the simulator immersion capabilities by combining visual, auditory, somatosensory and tactile inputs together with additional cognitive factors (e.g. manipulation of users' attention, expectations etc.) might help to override the (lack of) input from the vestibular system. Recently, Schulte-Pelkum et al. (2004) showed that vibrotactile stimulation significantly increased the visually-induced self-motion responses. Seen from another perspective, small vibrations of the body may be strongly associated with self-motion in a vehicle and could lead to positive high-level cognitive bias on the self-motion sensation. Evidence of the effect of such motion metaphor cues is provided by Bürki-Cohen et al., (1998), where pilots' judgments on flight simulator performance were biased due to presence or absence of vibrotactile stimulation. Furthermore, the entertainment industry effectively uses vibro-acoustic stimulation, where *bodysonic* chairs with embedded loudspeakers are used to enhance the experience (Isono et al., 1996; Cohen, 2002).

In the present research we investigated how additional vibrotactile stimulation can facilitate AIV. In an experiment we coupled self-motion inducing naturalistic sound stimuli with vibration delivered via shakers and low frequency sound. In line with the previous findings from the experiments reported by Schulte-Pelkum et al., (2004) we expected to find a facilitation effect of the additional modalities on AIV.

2. Method

2.1. Stimuli

The experimental design contained 24 stimuli with following parameters varied: 1) **3 sound types**: a) non-spatialized engine sound (anchor sound) b) spatialized moving "auditory landmarks", and c) engine + landmarks; 2) **4 stimulation types:** a) sound scene only, b) vibrotactile stimulation induced by mechanical shakers - further referred to as MV, c) vibrotactile stimulation induced by subwoofer - further referred to as LFV, and d) combined LFV and MV stimulation - further referred to as LFMV) 3) **2 stimulation intensity levels** (lower and higher).

Sound types. Both the anchor and spatialized sounds were designed would be recognizable to the research participants. The "auditory landmarks"-scenes contained binaural spatializations of sound sources approaching the listener from a distance in front of him. The aim of using

these spatializations was to elicit translational self-motion illusions in the forward direction (Figure 1) in a similar manner as in previous research on auditory induced rotational vection (Larsson et al., 2004, Väljamäe et al., 2004). Hence, two ecological sounds - "bus" (sound of an idling bus) and "dog" (barking dog) previously found to be effective in inducing rotational self-motion illusions were used as input to the spatialization. The stimuli were spatialized in Matlab™ using a catalogue of generic HRTFs. This catalogue was measured from a KEMAR mannequin using the procedure described in Väljamäe et al., (2004). The frequency range of the spatialized sound ranged from 0.1 to 13 kHz. Headphone equalization was applied to in order to prevent coloration artifacts and to increase externalization. No room acoustical rendering was applied.

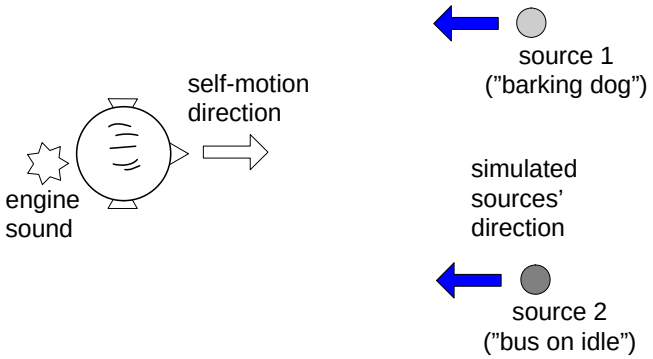


Figure 1: *Representation of the auditory scene used for creating translational auditory induced vection (AIV). The arrow in front of the listener indicates expected (AIV) direction, filled arrows indicate the motion direction of the virtual sound objects.*

Anchor sounds are sounds that are generated by a listener's movements or other sounds that are caused by the listener (e.g. engine sounds, footstep sounds, sounds of one's own breath, coins in the pocket, etc.; see Väljamäe et al., 2005b for discussion). The idea behind including such an anchor sound in the experiments was inspired by participants' verbal responses from Larsson et al., (2004), where the sounds of footsteps created an elicited an illusion of moving in a crowd. The anchor sound "follows" a listener and in the case of motion simulation it should be perceived as accompanying the direction of listeners' illusory self-motion. Such anchor sounds can be seen as an auditory correspondence to the visual self-avatar, which has proven to be an important component for compelling VE experiences (Usoh et al., 1999). In some recent experiments we showed that such "sonic self-avatars" may significantly increase translational AIV (Väljamäe et al., 2005b).

In the current experiments, an anchor sound intended to represent the sound of a small electrical vehicle - a "Cybercart" (such as an electrical

wheelchair) - was used. This sound was synthesized using software based on the Synthesis ToolKit-library (Cook & Scavone, 2004) and consisted of ten sinusoidal signals and a recorded noisy component mimicking the sound of a gearbox. The software developed to produce the engine sound also allows for varying the frequencies of the sine components in such way as to resemble an increase in rpm of an electrical engine. This feature was used in the experiment to create a sound mimicking the Cybercart engine acceleration. The frequency spectrum of the sound at constant velocity is shown in Figure 2 (left panel).

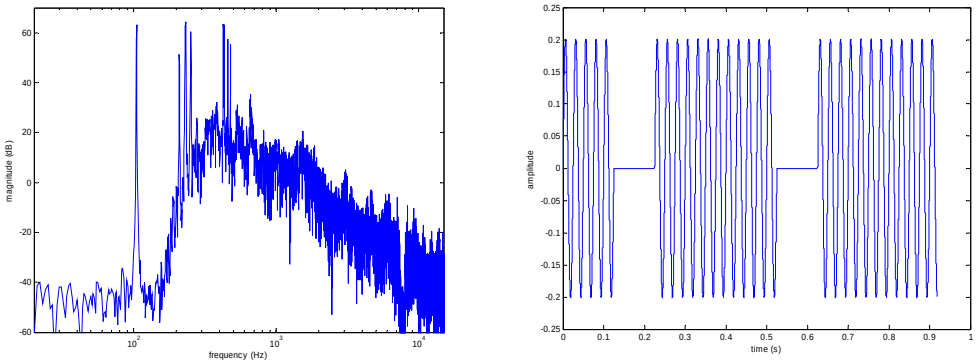


Figure 2. *(Left) Frequency spectrum of the engine sound at constant velocity. The fundamental frequency of the sinusoidal part of the sound is at about 105 Hz while the other sine components are in the range 210-670 Hz. All other, lower-magnitude parts of the spectrum are due to the noise-like "gearbox" sound. (Right) Temporal structure of the signals fed to the shakers. Each sine-burst is 0.3 s of duration while the silent parts are 0.1 s of duration.*

Mechanical vibrotactile stimulation by shakers (MV). In order to create the illusion of traveling over a rough surface (such as a cobblestone pavement), conditions were included where an intermittent sinusoidal signal, created in Matlab™, was fed to shakers mounted under the experimental chair. The stimulation was presented at two different levels – 2.6 cm/s^2 and 4.9 cm/s^2 – comparable to the threshold of the vestibular system needed for the self-motion direction perception (4.9 cm/s^2 (range 2.9-6.3) for step acceleration as reported in Gianna et. al., 1996). The temporal distribution of the 40 Hz sinusoidal bursts driving vibrators is shown in Figure 2 (right). An "acceleration phase" was also added to the signals in which silent period between the tone bursts driving the MV was gradually decreased.

Vibrotactile stimulation by subwoofer (LFV). A custom built subwoofer placed behind the experimental chair (see Figure 3) was used to present a 40 Hz stationary sinusoidal tone. This frequency was chosen since the subwoofer's lower cut-off frequency (slightly below 40 Hz) prevented satisfactory reproduction of lower-frequency tones. As with the shakers, two levels of stimulation were used: one at 60 dB and one at 66 dB

(measured at the head position). This stimulation is slightly above the threshold of hearing, which is about 50 dB at 40 Hz (Zwicker & Fastl, 1999). The low frequency stimulations were intentionally reproduced at these low levels in order not to mask the headphone-reproduced sound and to achieve only diffuse vibrotactile stimulation.

All sound excerpts were 67 seconds long, and the motion pattern of the sound sources followed a pre-rendered cycle: 4 second stationary phase, 3 second acceleration, 60 constant velocity phase. A Hann half-window of 0.5 second duration was applied to achieve smooth stimuli on- and off-sets. The engine sound started at the acceleration phase of the moving sound sources as did the vibrotactile stimulation.

2.2. Measures

To assess AIV and subjective presence sensations, three direct measures were used in this experiment: presence, vection intensity, and convincingness of vection.

Presence was defined in the questionnaire as a sensation of being actually present in the virtual world. Vection intensity corresponded to the level of the subjective sensation when experiencing self-motion. On the convincingness scale, subjects had to report how convincing the sensation of self-motion was for the reported direction. It should be noted that the convincingness and intensity ratings often are highly correlated. Ratings of all three measures were given on a 0-100 scale.

Apart from the direct measures listed above, an indirect binary measure, reflecting the number of ego-motion experiences, was used (participants were asked to verbally indicate the direction of self-translation). While vection onset time often is used in experiments with visual stimuli (Riecke et al. 2005), onset time was not measured in the present study since previous experiments Larsson et al., (2004) on auditory-induced vection indicated that this measure showed large inter-individual variance.

2.3. Procedure

Twenty-three participants (10 female) with a mean age of 24.5 (SD 4.8) took part. The experiment was conducted in a custom-made laboratory setup with black curtains surrounding the participant (see Fig. 3). Stimuli were played back to blindfolded participants using Beyerdynamic DT-990 Pro circumaural headphones. It should be noted that special measures were taken to enhance the auditory self-motion sensation – several studies (Larsson et al., 2005; Gustaviano et al., 2005) confirm that the visual

impression of the experimental setup (e.g. visible loudspeakers) can significantly bias participants' responses. In line with this reasoning, the subwoofer behind the chair (see Figure 3) was covered with black cloth.

Figure 3: *Laboratory setup: a participant sitting on a chair mounted on a wheeled platform connected to a wheeled footrest.*

We also believe that in the context of self-motion stimulation, seeing a completely immovable setup could bias the AIV responses. Therefore, participants were seated on a chair mounted on a wheeled platform connected to a wheeled footrest as shown on Figure 3. Participants were not explicitly told that the chair could physically move, but the facts that the wheels were visible and that the chair moved slightly when they sat down should act to insure them that the simulator setup potentially could make use of motion cues.

Participants were briefed verbally about the test procedure and a short training session was performed before the experiment started in which two stimuli were presented. During the playback of the sound excerpts participants were asked to verbally report about the direction of their movement, if a sensation of self-motion was perceived. Although the simulated auditory scene shown in Figure 1 intended to elicit only forward AIV, binaural synthesis with non-individualized HRTFs can lead to mislocalization of sound sources positions and trajectories and result in different percepts of the direction of self-motion (i.e. backwards, forward-left etc., see Väljamäe et al., 2005a for further discussion on binaural

synthesis artifacts in translational AIV simulation). Another reason why participants were asked to report the direction of self-motion was to attempt to mask the aim of experiment and thus reduce the related cognitive bias.

The experiment leader stopped the stimuli playback when a direction of self-motion was reported and then participants were asked to give ratings on intensity, convincingness of self-motion, and the overall presence sensation. If no self-motion was experienced during the sound excerpt presentation, participants had to rate the overall presence sensation at the end of the stimulus playback. Stimuli were presented in randomized order. Apart from the verbal responses to the questionnaire, verbal probing was done by the experiment manager. After completing the experiment, participants were debriefed, thanked and paid for their participation.

3. Results

All ratings were analyzed using separate repeated-measures ANOVAs. From these analyses the present article focuses on three aspects 1) main effect of vibrotactile stimulation (sections 3.1 and 3.2), 2) the interaction between the sound types and vibrotactile stimulation (section 3.3).

3.1. Binary vection measure

One of the measures for experienced AIV is a binary vection measure, which shows how many participants experienced vection for a particular stimulus type. One can also convert the number of vection instances per stimulus into a percentage where 100% will mean that all participants felt self-motion for this particular stimulus type. The results in Figure 5 show that additional MV vibrotactile stimulation resulted in higher binary vection limits 57-93% (13-22 from 23 participants) compared to the other experiment results reported in (Väljamäe et al., 2005b), where the same ecological sounds were used 50-83% (12-20 from 24).

Several interesting trends can be seen from the binary vection ratios depicted in Figure 4, which are also reflected in the statistical analysis of the other measures given below. One can see that mechanically induced vibrations (MV) are the most effective in enhancing AIV which is further corroborated by ANOVA analysis for the other measures presented in the next subsection. Additionally, one can also see an interaction between sound type and MV stimulation can be seen (cf. section 3.3 for discussion).

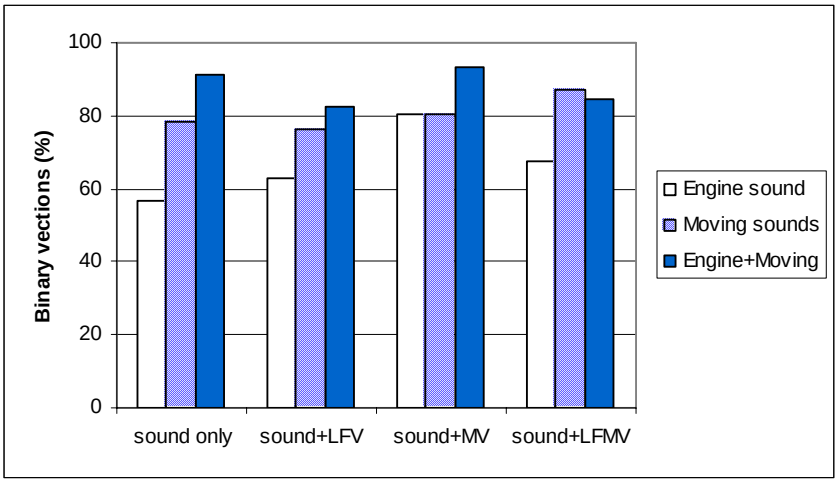


Figure 4: *Binary vection measures for different vibrotactile stimulation types: LFV – low frequency induced vibration, VB – mechanically induced vibration, LFMV – low frequency and mechanical induced vibration. Sound types: engine sound – non-spatialized engine sound metaphor, moving sounds – spatialized acoustic landmarks ("idling bus" and "barking dog").*

3.2. Main effects of cross-modal stimulation

To examine the main effects of different types of vibrotactile stimulation, ratings for stimuli with the same type of vibrotactile stimulation were averaged over two different stimulation levels. The resulting factorial design for repeated-measures ANOVA was 3 (sounds types) x 4 (stimulation type: sound only, sound with low frequency induced vibrations (LFV), sound with mechanically induced vibrations (MV), sound with low frequency and mechanical vibrations (LFMV).

For the intensity ratings a main effect of cross-modal stimulation was significant at $F(3,66) = 4.33$, $p = .013$, with the means 38.7 (sound only), 38 (sound + LFV), 44.9 (sound + MV) and 43.2 (sound + LFMV), see Figure 5. Bonferroni-corrected pairwise comparisons revealed that mechanical vibration (MV) type was marginally significant ($p = .09$) compared to the condition without vibratory stimulation. For the convincingness ratings no such effect was observed (cf. section 3.3).

For the presence ratings, a main effect of cross-modal stimulation reached significance at $F(3,66) = 2.90$, $p < .05$, with the means 48.2 (sound only), 51.1 (sound + LFV), 52.1 (sound + MV) and 53.2 (sound + LFMV). Bonferroni-corrected pairwise comparisons showed that only the difference between "sound only" and "sound with LFMV" stimulation was significant ($p<.05$). It is important to note the difference between the vection and presence ratings seen in Figure 5, where both AIV intensity and convincingness ratings are highest for MV condition but for self-reported

presence a gradual increase can be seen with the maximum for LFMV condition. This is in line with Sanders and Scorgie (2002) finding which showed that low frequency sound positively affects presence responses.

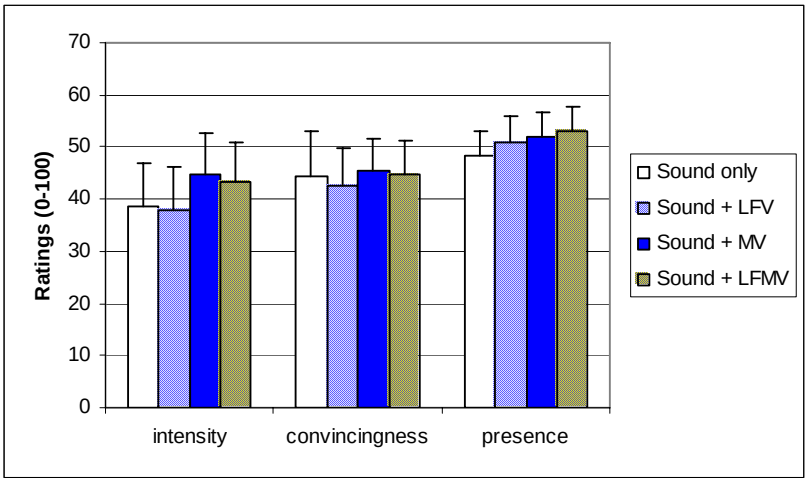


Figure 5: *Main effect of vibrotactile stimulation types: LFV – low frequency induced vibration, MV – mechanically induced vibration, LFMV – low frequency and mechanical induced vibration; (upper bound of 95% confidence interval is indicated).*

3.3. Interaction between sound type and vibrations

As described above, an engine sound (anchor) has been added to some of the stimuli, which coupled in synchrony with the MV stimulation produced an interesting interaction reported here.

The binary vection measure is presented in Figure 6 (top left panel). From a fist glance, it is surprising that almost half of the participants (48%) reported AIV with only the non-spatialized engine sound. However, this finding corroborates results presented in Väljamäe et al., (2005b). Figure 6 also shows that the engine sound coupled with MV results in same binary vection ratings (69 %) as the auditory scene type with moving sounds.

The intensity ratings showed the same marginal trend for the interaction between sound and cross-modal stimulation types, $F(6,132) = 2.02$, $p=.09$. As can be seen in Figure 6 (bottom right panel), this effect was primarily caused by the interaction between MV stimulation and the engine sound. Paired-samples T-test confirmed the significance of this interaction ($p=0.006$, $t=-3.034$).

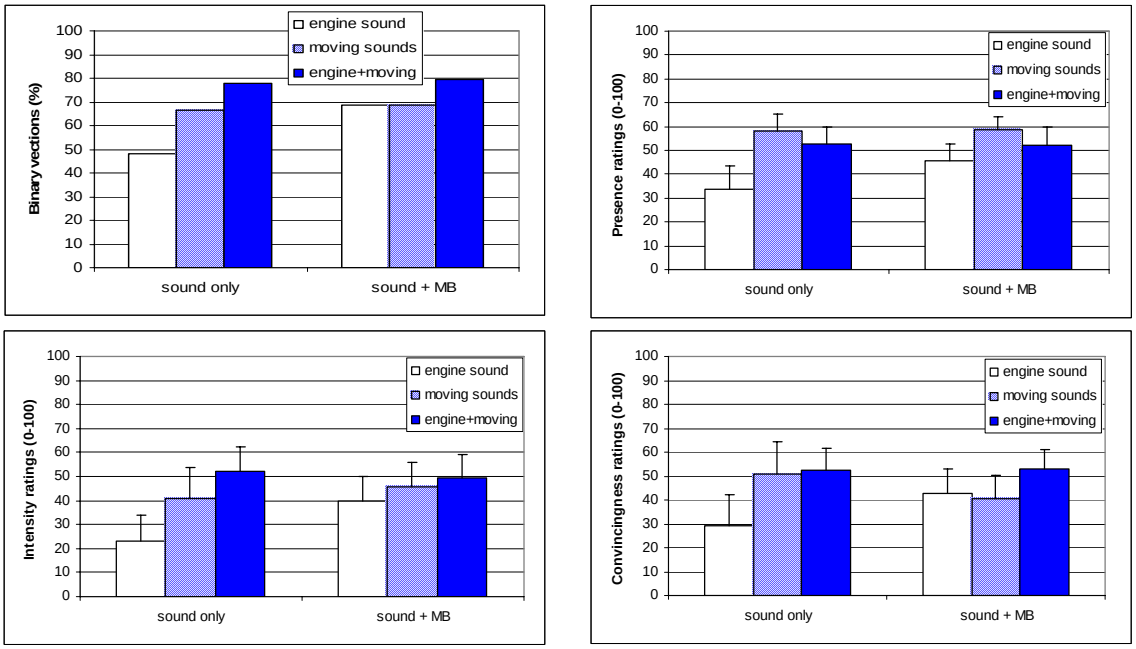


Figure 6: *Sound type and vibration interaction: (left top) binary vections (%), (left bottom) intensity ratings, (right bottom) convincingness ratings, (right top) presence ratings. Upper bound of 95% confidence interval is indicated for the ratings.*

For the convincingness ratings (Figure 6, bottom right panel), the interaction between sound type and MV also reached significance, $F(6,132)=2.83$, $p<.03$, due to two effects. First, as for the intensity ratings, MV stimulation increased AIV ratings for the engine sound condition ($p=0.046$, $t=-2.112$). Second, convincingness of AIV for the moving sounds condition decreased with the addition of MB stimulation (marginally significant, $p=0.06$, $t=1.981$).

Similar to the intensity and convincingness ratings, the analysis of the presence ratings showed a marginally significant interaction between cross-modal stimulation and sound type $F(6,132) = 2.22$, $p=.07$ (Figure 6, top right panel). As for intensity and convincingness, MV stimulation and engine sound interaction caused this effect ($p=0.015$, $t=-2.636$).

4. Discussion and conclusions

This study aimed to show that vibrotactile stimulation can enhance illusory self-motion and presence responses. The main effect of vibrotactile facilitation of translational AIV responses is consistent with the similar finding by Schulte-Pelkum et al. (2004) where vibrations enhanced visually induced self-motion. Mechanically induced vibrations, but not the low frequency sound conditions, significantly increased the self-motion intensity and convincingness ratings. Absence of AIV facilitation by low

frequency sound could be accounted for the low intensity level applied. In addition, the highest presence ratings were reported for conditions including both mechanical vibrations and low frequency stimulation. This finding corroborates results from Sanders and Scorgie (2002) where low frequency sound significantly increased presence ratings for presented auditory scenes, both in questionnaire-based responses and physiological measures. Thus, it is plausible that a more systematic variation of low frequency content would show facilitating effects on AIV. Future research should further investigate this issue.

A main finding was that vibrotactile facilitation of AIV interacted with sound sources type. Three sound conditions were applied containing only moving sources, only a motion metaphor (engine sound) or both types combined. Significant interaction between engine sound and vibrations was observed both for self-motion and presence responses. First, stimuli containing only the non-spatialized engine sound became as effective as stimuli with only moving sound fields when combined with vibratory stimulation. Second, the ratings for the convincingness of self-motion direction decrease when vibrations are added to the stimuli containing only moving sound fields. Both effects suggest that the observed effects are likely to depend on participant's expectations and their prior experience, as auditory-vibrotactile metaphor of an engine has a high ecological validity (i.e. the cyberkart sounds and shakes). Consequently, shaking an image could serve as a visual counterpart of such multi-modal metaphor representing a ride on a virtual vehicle.

It should be noted that the temporal aspects of the multi-modal engine metaphor might be a crucial component of the observed interaction effect. The starting time of the engine sound was synchronized with the vibrations initial burst producing a clearly noticeable perceptual event. In another study on vibrotactile facilitation on visually induced vection by Riecke et al. (2005), several participants reported a mismatch between the visual motion velocity profile and the expected vibratory pattern which resulted in the decreased self-motion and presence responses. One could speculate that this synchronous auditory-vibrotactile imitation of the engine start might work as a virtual "jerk" – the illusory counterpart of a small initial physical movement which has been found to enhance visually induced vection (Wong & Frost 1981). Taking into account the recent findings on inhibitory effects in the vestibular cortex area caused by visually induced vection (see Dieterich et al. 2003), one might expect to find corresponding neurophysiological patterns underlying auditory-vibrotactile interaction effects in self-motion simulations. Hopefully, a future research will provide us with a better understanding of the brain processes underlying the effects presented in this paper.

In our AIV simulation we applied only a minimal set of acoustic cues, which were the most instrumental for auditory motion simulation – intensity level and binaural cues (Väljamäe et al., 2005a). For binaural synthesis, a generic HRTFs catalogue was used, which resulted in some perceptual artifacts of the auditory scene (e.g. distortions in the perceived trajectories of sources' motion) both for sound-only and bimodal conditions. However, the observed self-motion facilitation via vibrotactile stimulation can be seen as a way to compensate for reduced quality of spatial sound rendering. Although vibratory stimulation via a steering wheel or a game pad is often implemented in commercially available products, it is likely that multisensory representations of virtual objects and events (e.g. audio-tactile engine simulation) might lead to the highest self-motion and presence responses. Future studies should further investigate the benefits of cross-modal optimization of different sensory information which will be used for self-motion simulators and other mediated environment applications.

Acknowledgements

The work presented in this paper was supported by the European Community under the FET Presence Research Initiative project POEMS (Perceptually Oriented Ego-Motion Simulation), IST-2001-39223 and the Swedish Research Council (VR project 40499601). The first author of this paper also would like to thank for the support from Alfred Ots' scholarship foundation.

References

Andersen, G.J., (1986). Perception of self-motion: Psychophysical and computational approaches, *Psychological Bulletin, vol. 99(1),* 52-65.

Bach-y-Rita, P. (1972). *Brain mechanisms in sensory substitution*. New York: Academic Press.

Begault, D.R. (1994). *3D Sound for Virtual Reality and Multimedia*. London: Academic Press Professional.

Bürki-Cohen, J., Soja, N.N., & Longridge T. (1998). Simulator Fidelity Requirements: The Case of Platform Motion. *Proceedings of 9th ITEC International Training & Education Conference*, Lausanne, Switzerland, 216- 231

Cohen, M. (2002) A survey of emerging and exotic auditory interfaces. *Proceedings of International Conference on Auditory Displays*, Kyoto, Japan

Colebatch, J.G., Halmagyi, G.M., & Skuse, N.F. (1994) Myogenic potentials generated by a click-evoked vestibulocollic reflex. *Journal of Neurologies and Neurosurgical Psychiatry, 57,* 190–197.

Cook, P. R. & Scavone G. P. (2004). The Synthesis ToolKit in C++ (STK), http://ccrma.stanford.edu/software/stk/

Dieterich, M., Bense, S., Stephan, T., Yousry, T.A., Brandt, T. (2003). fMRI signal increases and decreases in cortical areas during small-field optokinetic stimulation and central fixation. *Experimental Brain Research, 148(1),* 117-27

Gaver, W. W. (1993). How Do We Hear in the World? Explorations in Ecological Acoustics. Ecological Psychology, 5, 285–313.

Gianna C., Heimbrand, S., & Gresty, M. (1996). Thresholds for diction of motion direction during passive lateral whole-body acceleration in normal subjects and patients with bilateral loss of labyrinthe function. *Brain Research Bulletin, 40,* 435–439.

Guastavino, C., Katz, B., Polack, J-D., Levitin, D., & Dubois, D. (2005). Ecological validity of soundscape reproduction. *Acustica united with Acta Acustica, 91(2),* 333-341.

Harris, L.R., Jenkin, M., Zikovitz, D., Redlick, F., Jaekl, P., Jasiobedzka U. et al. (2002) Simulating self motion I: cues for the perception of motion. *Virtual Reality 6(2),* 75–85

Isono H., Komiyama, S., & Tamegaya, H. (1996) An autostereoscopic 3-D HDTV display system with reality and presence, *SID Digest,* 135-138.

Kleiner M., Dalenbäck B-I. & Svensson P., "Auralization - An Overview, *J. Audio Engineering Soc.*, Vol. 41 (11), pp 861-875, 1993

Lackner J.R., & Graybiel A. (1974). Elicitation of vestibular side effects by regional vibration of the head. *Aerospace Medicine, 45,* 1267–1272.

Larsson, P., Västfjäll, D., & Kleiner, M. (2004). Perception of self-motion and presence in auditory virtual environments. *Proceedings of the Seventh Annual Workshop of Presence,* Valencia, Spain, 252-258.

Larsson, P., Västfjäll, D., & Kleiner, M. (2005). Auditory-visual interaction in concert halls. *To appear in Acustica/Acta Acustica.*

Lombard, M. & Ditton, T. At the heart of it all: the concept of presence. *Journal of Computer Mediated Communication*, 3(2), 1997.

Mccauley, M.E. & Sharkey T.J. (1992). Cybersickness: Perception of self-motion in virtual environments *Presence-Teleoperators and Virtual Environments, 1(3),* 311–318.

Pausch, R., Crea, T., & Conway, M. (1992). A literature survey for virtual environments: Military flight simulator visual systems and simulator sickness. *Presence-Teleoperators and Virtual Environments,1(3),* 344-363.

Riecke B.E., Schulte-Pelkum, J., Caniard, F., & Bülthoff H.H. (2005) Towards lean and elegant self-motion simulation in virtual reality. *In Proceedings of IEEE VR2005,* Bonn, Germany, 131–138

Sanders, R.D. Jr. & Scorgie, M.A. (2002). The effect of sound delivery methods on a user's sense of presence in a virtual environment. (Master thesis Naval Postgraduate School, Monterey, CA).

Schulte-Pelkum J., Riecke B.E., von der Heyde, M., & H.H. Bülthoff. Circular vection is facilitated by a consistent photorealistic scene. *Proceedings of Presence 2003,* Aalborg, Denmark. October 2003.

Schulte-Pelkum, J., Riecke B.E., & Bülthoff H.H. (2004) Vibrational cues enhance believability of ego-motion simulation. *International Multisensory Research Forum (IMRF'04),* Barcelona, Spain

Schulte-Pelkum, J. et al. Vibrotactile interaction effects in visually and auditory induced self-motion (in preparation)

Stoffregen, T.A. & Bardy, B.G. (2001). On specification and the senses. *Behavioral and Brain Sciences, 24,* 195- 261

Usoh M et al., "Walking>virtual walking>flying, in virtual environments", *Proceedings of SIGGRAPH'99*, 1999, pp. 359-364

Väljamäe, A., Larsson, P., Västfjäll, D., & Kleiner, M. (2004). Auditory presence, individualized head-related transfer functions and illusory ego-motion in virtual environments. *Proceedings of the Presence 2004*, Valencia, Spain, 141-147.

Väljamäe, A., Larsson, P., Västfjäll, D., & Kleiner, M. (2005a). Travelling without moving: Auditory scene cues for translational self-motion. *Proceedings of ICAD'05*, Limerick, Ireland.

Väljamäe, A., Larsson, P., Västfjäll, D., & Kleiner, M. (2005b). Sonic self-avatar and self-motion in virtual environments. (in preparation)

Wall III, C., & Weinberg M.S. (2003), Balance prostheses for postural control. *IEEE Engineering in Medicine and Biology Magazine, 2(22)*, 84-90.

Wong, S.C.P., & Frost B.J. (1981) The effect of visual-vestibular conflict on the latency of steady-state visually induced subjective rotation. *Perception & Psychophysics, 30(3)*, 228–236.

Young, L.R., & Shelhamer, M. (1990). Weightlessness enhances the relative contribution of visually-induced self-motion. In R. Warren and A.H. Wertheim (Eds.), *Perception and Control of Self-motion: Resources for Ecological Psychology* (pp. 219-263). Hillsdale, NJ: Erlbaum.

Zwicker E., Fastl H. (1999). *Psychoacoustics: Facts and models*. 2nd ed. Springer Verlag, Berlin.

Paper D

Sound can compensate for a restricted field-of-view in self-motion simulators

Aleksander Väljamäe, Daniel Västfjäll, Pontus Larsson and Mendel Kleiner

Extended version is to be submitted to
IEEE Transactions on Multimedia

Sound can compensate for restricted field-of-view in self-motion simulators

Aleksander Väljamäe[1,2], Pontus Larsson[2], Daniel Västfjäll[2,3] and Mendel Kleiner[2]
[1]Division of Communication Systems, Chalmers University of Technology, Sweden
[2]Division of Applied Acoustics, Chalmers University of Technology, Sweden
[3]Department of Psychology, Göteborg University, Sweden
{aleksander.valjamae@s2.chalmers.se, pontus.larsson@ta.chalmers.se,
daniel.vastfjall@psy.gu.se, mendel.kleiner@ta.chalmers.se}

Abstract

The sensation of illusory self-motion and presence can be reliably elicited in motion simulators with large visual displays. However, many low-cost multimedia applications use smaller screens and could benefit from multisensory enhancement of user experiences. In the current study we examine how rotating visual scenes with a restricted field-of-view can be enhanced by the simultaneous presentation of additional spatial sound. In an experiment, we found that binaurally synthesized rotating auditory scenes significantly increased the perception of self-motion and presence. Additionally, the results showed that even a limited spatial sound resolution was sufficient for substantially enhancing the experience. This finding may be put to practical use in commercially available multimedia technologies (e.g. computer games, low-cost motion simulators).

Keywords*-Audio-visual interaction, multisensory optimization, auditory scene synthesis, illusory self-motion*

1. Introduction

In the recent years, the field of multisensory research has witnessed a substantial growth. A consequence of this is that knowledge on cross-modal interaction effects has spread into other disciplines, including Virtual and Augmented Reality (AR/VR) applications. Cost-effective motion simulators might greatly benefit from multisensory enhancement of end-user experience (e.g. self-motion and presence sensation). In particular, creating a sensation of self-motion and presence ("a sense of being present in a remote or virtual environment") might be highly appreciated in low-cost multimedia applications where relatively small visual displays are typically used. In addition, a wide field-of-view (FOV) is often believed to be one of the contributing factors to simulator sickness or cybersickness (Duh et al. 2002).

Illusory self-motion (also known as vection) phenomenon implies a perception of one's own actual movement relative to the surrounding environment (Anderssen, 1986). Historically, research has been largely focused on vection elicited by visual stimuli and has often neglected other modalities contributing to self-motion. Our recent studies on rotational and translational auditory-induced vection (AIV) suggest that sound also plays an important role in the self-motion perception (Larsson et. al (2004), Väljamäe et al. (2005a)). These studies have shown that AIV may reliably be induced by moving virtual sound fields, synthesized using binaural technology (Kleiner et al., 1993). Vibrotactile enhancement of bisensory self-motion stimulation has been demonstrated for visual (Schulte-Pelkum et al., 2004) and auditory stimuli (Väljamäe et al., 2005b).

Psychophysical investigations on multisensory enhancement of motion simulators can greatly benefit from the assessment of the "perceptual illusion of non-mediation", as presence has been often defined (see Lombard and Dittion, 1997). Being "spatially present" in a specific context provides a stable reference frame, which allows for a good spatial orientation. Consequently, illusory self-motion ratings have been shown to be closely related to spatial presence (e.g. Schulte-Pelkum et al., 2003, Riecke et al. 2005).

The current study investigated whether additional spatial auditory cues can be utilized to enhance visually induced self-motion. A recent study by Riecke et al. (2005) showed that adding spatial sound to a rotating visual scene had a small but significant effect on self-motion and presence responses. The magnitude of the observed effect suggested a ceiling effect due to the relatively wide FOV used (54˚×45˚). To get further insight into this cross-modal interaction effect in audio-visually induced self-motion, the current experiment utilized a FOV restricted to 20˚×15˚ and 10˚×7.5˚, and used different spatial resolutions for the synthesis of the accompanying rotating auditory scenes.

2. Method

2.1 Participants and Apparatus

Twenty-four participants (nine female) with a mean age of 25.1 (SD 4.2) took part in this experiment. The experiment was conducted in a custom-made laboratory setup with black curtains surrounding the participant (see Figure 1). Participants were comfortably seated at a distance of 1.7m from a curved projection screen (2m curvature radius) on which the rotating visual stimulus was displayed (see Fig. 1, right panels). Visual stimuli were rendered in real time using the VElib software (VElib, 2005). Auditory

stimuli were spatialized in real-time via Lake Huron DSP system (Lake Technology, 2005) and played back with Beyerdynamic DT-990Pro circumaural headphones.

Figure 1. *Left: Participants were seated at a distance of about 1.7 m from a curved projection screen displaying a view of the market place. Right top: trial with small FOV (10˚×7.5˚) Right bottom: fragment of 360˚ roundshot of the Tübingen market place used as visual stimuli.*

2.2. Stimuli and Design

Each participant was exposed to 24 audio-visual excerpts following the 2x2x3x2 experimental design: 1) two alternating rotation directions - left/right; 2) two simulated FOVs - small (10˚×7.5˚) and medium (20˚×15˚), which matched the physical FOV under which the projection screen was seen by the participants; 3) three sound rendering conditions with different spatial characteristics – a) mono, b) reduced 3D sound (BinScape synthesis mode in Lake Huron which uses generic HRTFs set at a very scarce, 60° resolution comparable to the 5-channel loudspeaker reproduction and c) 3D sound with full resolution (HeadScape synthesis mode in Lake Huron where generic HRTFs used at maximal, 5˚ resolution; 4) 2 spatial sound conditions where in one of them an additional ambient sound recorded binaurally on a market square was added to the Lake-rendered spatial sound (this ambient sound contained typical small town square sounds which are difficult to localize, e.g. distant speech, sparrow's tweets, etc.).

The visual stimulus consisted of a photorealistic view of a city marketplace (Tübingen market place; Riecke et al., 2005) that was generated by wrapping a 360˚ roundshot (4096 × 1024 pixel) around a virtual cylinder and adding a black mask to get the desired FOV (see Fig. 1, right panels). Spatial sound scenes consisted of scenes with binaural spatializations of sound sources corresponding to "auditory landmarks"-ecological sound sources that can be classified by a listener as spatially "still" (Larsson et al., 2004). Hence, two ecological sounds – an "idling bus" (sound of an idling bus) and a "fountain" previously found to be effective in inducing rotational self-motion illusions (Larsson et al., 2004) were used binaural synthesis. The frequency range of the spatialized sound ranged from 0.1 to 13 kHz. Headphone equalization was applied in order to prevent coloration artifacts and to increase externalization. No room acoustical rendering was applied. The temporal structure of rotation (still, acceleration to 30 deg/s, constant velocity, deceleration) was kept the same.

2.4. Measures

To assess the AIV, two direct verbal measures were used in this experiment: vection intensity and presence. Vection intensity corresponded to the level of subjective sensation when experiencing self-motion. Presence was defined in the questionnaire as "a sensation of being actually present in the virtual world". Ratings of both measures were given on a 0-100 scale where 100 corresponded to sensation equal to that of a real event (physical motion or "being there" sensation for presence).

2.3. Procedure

Participants were instructed verbally about the experimental procedure and a short training session was performed before the experiment start (2 stimuli presented). Participants were instructed to stop the stimuli playback by using a joystick in the case of experiencing self-motion illusion, or to wait until the end of the stimulus playback. After that, verbal ratings of presence and self-motion intensity were reported to the experiment leader. Stimuli were presented in 3 blocks containing different sound rendering qualities due to the need to restart the Lake System each time a new sound rendering engine from Lake Huron was applied (other factors were randomly presented within each block). The order of these 3 blocks was randomly changed between subjects. Apart from the verbal responses to the questionnaire, verbal probing was done by. After completing the experiment, participants were debriefed, thanked and paid for their participation.

3. Results

3.1. Main effects

All ratings were submitted to parametric Analyses of variance (ANOVAs) and the ratings means are shown on Figure 2. For self-motion intensity ratings, a main effect of FOV reached significance, $F(1, 23) = 12$, $p < 0.005$ with means of 46.1 (medium FOV) vs. 36.0 (small FOV). For rendering quality, intensity ratings also reached significance $F(2,46) = 4.82$, $p < 0.05$ with means 35.8 (mono), 43.2 (reduced spatial sound) and 44.3 (spatial sound). LSD-corrected pairwise comparisons showed a significant ($p < 0.05$) difference between mono and spatial sound conditions.

For presence ratings a main effect of FOV reached significance, $F(1,23) = 27.96$, $p < 0.001$, with means of 38.6 (medium FOV) and 27.3 (small FOV). For rendering quality presence ratings showed a marginal trend ($p=0.067$) with means of 29.4 (mono), 33.5 (reduced spatial sound), and 35.9 (spatial sound). LSD-corrected pairwise comparisons revealed that the difference between mono and spatial sound condition was significant ($p < 0.05$).

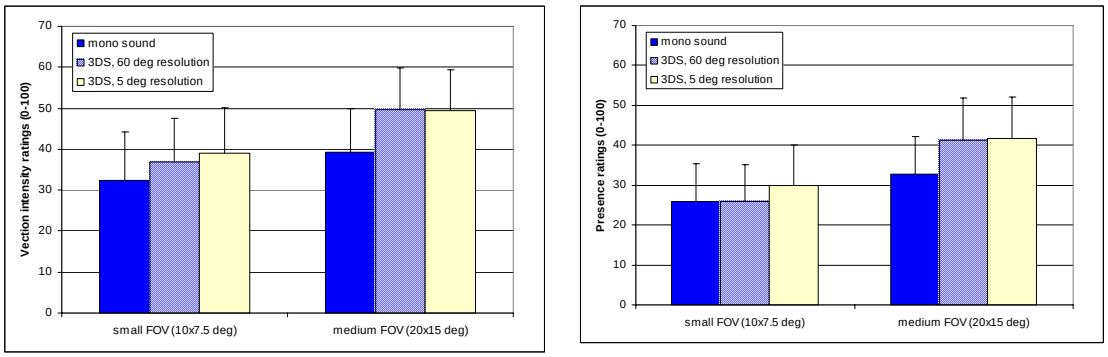


Figure 2: *Vection intensity ratings (left panel) and presence ratings (right panel) with 95% confidence intervals (upper bound indicated).*

3.2. Audio-visual interaction

The interaction between FOV and rendering type also reached significance both for intensity $F(2,46)=3.67$, $p<0.05$ and presence $F(2,46)=4.54$, $p<0.05$ ratings (see Fig 2, left panel). As may be seen, this effect is caused by different patterns of audio-visual interaction for different screen sizes. Auditory facilitation of the reported AIV and presence reached significance only for trials with medium FOV and between the mono and the reduced

3DS conditions. Paired-samples T-test results for these two conditions were p=0.004 (t=-3.211) for intensity, at p=0.028 (t=-2.343) for presence.

4. Discussion

The present study aimed to investigate if sound could compensate for a limited FOV in self-motion simulation scenarios. As predicted, rotating auditory scenes significantly enhanced self-motion intensity and presence ratings. Furthermore, it was found that the sound condition with reduced spatial resolution was sufficient for the enhancement of the visual effect and no further enhancement was obtained by high-resolution binaural synthesis. One should note, that spatial sound synthesis with a reduced, 60 deg resolution (Lake Huron Binscape mode) is comparable with the conventional 5-channel loudspeaker reproduction systems. This finding suggests that there may be a ceiling effect for how much spatial sound quality contributes to motion simulation when it accompanied by a visual scene.

This study also showed an audio-visual interaction effect between spatial sound quality and presented field-of-view. As expected, a twice larger FOV resulted in a significant increase both in the vection and presence responses. However, the auditory enhancement of self-motion and presence ratings was observed only for the medium FOV condition. One possible explanation for this finding is the large multisensory inconsistence between auditory and visual inputs, as the small FOV condition resulted in a very scarce representation of the content of the visual scene (market place). Since the sounds were coupled to the visual content, it may be that spatialized sound represented a larger improvement for the medium FOV (where visual objects were seen and heard), than for the small FOV (where objects sometimes only were heard). An interesting finding though is that when adding the spatialized sound to the smaller FOV, vection and presence ratings become comparable to the mono condition for the medium FOV.

It should be noted, that the medium FOV condition used in this study is very limited compared to the study by Riecke et al. (2005) - 54˚×45˚ or, other studies, where typically horizontal FOV of 60˚ to 80˚ is used for eliciting self-motion (Duh et al., 2002). The main aim of this study was to investigate the efficiency of small visual displays for eliciting self-motion sensation. Table 1 present the FOV sizes for several commercially available visual displays (computer monitors, TV and other small portable devices). As one can see, the tested medium FOV condition is even smaller that the typical FOV of a computer user (e.g. when playing a computer game).

Table 1: *Applications with a small FOV compared to experimental FOV conditions*

Application (diagonal in cm)	Distance to the viewer [m]	Screen size [m]	FOV [deg]
TV set (3:4 ratio, 72 cm)	4.0	0.58 x 0.43	8.3˚×6.2˚
Small FOV			**10˚×7.5˚**
Portable device	0.4	0.10 x 0.07	14.3˚×10˚
TV set (3:4 ratio, 72 cm)	2.0	0.58 x 0.43	16.5˚×12.3˚
Medium FOV			**20˚×15˚**
Monitor (43.2 cm)	0.7	0.34 x 0.27	26.9˚×21.8˚
Monitor (43.2 cm)	0.4	0.34 x 0.27	45.4˚×37.3˚

5. Conclusions

The present study showed that spatialized sound can enhance illusory self-motion and presence ratings when presented simultaneously with visual stimuli providing a limited field-of-view (20˚×15˚). This effect was achieved by presenting binaurally synthesized rotating auditory scenes with low spatial resolution comparable to the resolution of conventional 5-channel loudspeaker reproduction. These findings suggest that it might be feasible to use commercially available multimedia audio-visual systems for applications delivering self-motion and presence experiences. The success of these cost-effective displays, however, will be largely dependant on the research determining the perceptually optimal multisensory combinations.

Acknowledgements

The work presented in this paper was supported by the European Community under the FET Presence Research Initiative project POEMS (Perceptually Oriented Ego-Motion Simulation), IST-2001-39223 and the Swedish Research Council (VR project 40499601).

References

Andersen, G.J., (1986). Perception of self-motion: Psychophysical and computational approaches, *Psychological Bulletin, vol. 99(1)*, 52-65.

Duh, H., Lin, J., Kenyon, R., Parker, D., and Furness, T. (2002). Effects of characteristics of image quality in an immersive environment. Presence, 11(3), pp. 324–332

Kleiner M., Dalenbäck B-I. & Svensson P., "Auralization - An Overview, *J. Audio Engineering Soc.*, Vol. 41 (11), pp 861-875, 1993

Lake technology http://www.lake.com.au/lake_huron.htm

Lombard, M. & Ditton, T. At the heart of it all: the concept of presence. *Journal of Computer Mediated Communication*, 3(2), 1997.

Riecke B.E., Schulte-Pelkum, J., Caniard, F., & Bülthoff H.H. (2005) Towards lean and elegant self-motion simulation in virtual reality. *In Proceedings of IEEE VR2005*, Bonn, Germany, 131–138

Schulte-Pelkum J., Riecke B.E., von der Heyde, M., & H.H. Bülthoff. Circular vection is facilitated by a consistent photorealistic scene. *Proceedings of Presence 2003*, Aalborg, Denmark. October 2003.

Schulte-Pelkum J., Riecke B.E., von der Heyde, M., & H.H. Bülthoff. Circular vection is facilitated by a consistent photorealistic scene. *Proceedings of Presence 2003*, Aalborg, Denmark. October 2003.

Schulte-Pelkum, J., Riecke B.E., & Bülthoff H.H. (2004) Vibrational cues enhance believability of ego-motion simulation. In International Multisensory Research Forum (IMRF), Barcelona, Spain

Väljamäe, A., Larsson, P., Västfjäll, D., & Kleiner, M. (2005a). Travelling without moving: Auditory scene cues for translational self-motion. *Proceedings of ICAD'05*, Limerick, Ireland.

Väljamäe, A., Larsson, P., Västfjäll, D., & Kleiner, M. (2005b). Vibrotactile enhancement of auditory induced self-motion and presence. *Submitted to Journal of Audio Engineering Society*

VElib Virtual Environments Library (veLib) http://velib.kyb.mpg.de/de/