Olli Rummukainen

Audiovisual Reproduction in Surrounding Display: Effect of Spatial Width of Audio and Video

School of Electrical Engineering

Thesis submitted for examination for the degree of Master of Science in Technology.

Espoo 28.12.2011

Thesis supervisor and instructor:

D.Sc. (Tech.) Ville Pulkki

Aalto University
School of Electrical
Engineering

AALTO UNIVERSITY
SCHOOL OF ELECTRICAL ENGINEERING

ABSTRACT OF THE
MASTER'S THESIS

Author: Olli Rummukainen

Title: Audiovisual Reproduction in Surrounding Display: Effect of Spatial Width of Audio and Video

Date: 28.12.2011 Language: English Number of pages:10+73

Department of Signal Processing and Acoustics

Professorship: Acoustics and Audio Signal Processing Code: S-89

Supervisor and instructor: D.Sc. (Tech.) Ville Pulkki

Multimodal perception strives to integrate information from multiple sensorial channels into a unified experience, that contains more information than just the sum of the separate unimodal percepts. As a result, traditional quality metrics for unimodal services cannot reflect the perceived quality in multimodal situations, and new quality estimation methods are needed.

In this work, audiovisual perception was studied with an immersive audiovisual display. The audiovisual display consisted of a video screen with field of view of 226° and 3D sound reproduction with 20 loudspeakers. The aim of the study was to observe the cross-modal interaction of auditory and visual modalities, when the spatial widths of audio and video reproduction were limited. A subjective study was organized, where the overall perceived degradation of the stimuli was evaluated with Degradation Category Rating in four different types of audiovisual content. In addition, free descriptions of the most prominent degrading factors were collected. The participants' individual tendencies to experience immersion were screened prior to the experiment with a questionnaire.

The results show that video width is the dominant element in defining the degradation of a stimulus. Also audio width had an impact when the video width was at maximum. Individual tendency to experience immersion was not found to have significant impact on perceived degradation in this study. Slight content effects were observed. Constrained correspondence analysis of the free description data suggests the reasons for highest perceived degradation to be caused by wrong audio direction, reduced video width and missing essential content.

Keywords: Quality of Experience, multimodal perception, presence, immersion, audiovisual display

AALTO-YLIOPISTO
SÄHKÖTEKNIIKAN KORKEAKOULU

DIPLOMITYÖN
TIIVISTELMÄ

Tekijä: Olli Rummukainen

Työn nimi: Havaitsijaa ympäröivän äänen- ja kuvantoisto: Äänen ja kuvan leveyden vaikutus kokemuksen laatuun

Päivämäärä: 28.12.2011 Kieli: Englanti Sivumäärä:10+73

Signaalinkäsittelyn ja akustiikan laitos

Professuuri: Akustiikka ja äänenkäsittely Koodi: S-89

Valvoja ja ohjaaja: TkT Ville Pulkki

Moniaistinen havaitseminen perustuu informaation yhdistämiseen eri aistikanavista siten, että yhdistetty aistimus tuottaa enemmän tietoa ympäröivästä maailmasta kuin aistimusten käsitteleminen erillisinä. Tämän seurauksena vanhat laatumittarit yhteen aistiin perustuville järjestelmille eivät toimi arvioitaessa monimutkaisempia audiovisuaalisia järjestelmiä, ja uusien laatumittareiden kehittäminen on tarpeellista.
Tässä työssä audiovisuaalista havaitsemista tutkittiin immersiivisen audiovisuaalisen näytön avulla. Näyttö koostui 226° laajasta videokuvasta ja 20 kaiuttimella toteutetusta 3D äänentoistosta. Tutkimuksen tavoite oli tarkkailla kuulon ja näön vuorovaikutusta, kun kuvan- ja äänentoiston avaruudellista laajuutta rajoitettiin. Subjektiivinen laatuarviointi toteutettiin käyttäen diskreettiä näytteenhuonontumaskaalaa (DCR) havaitun laadun heikkenemisen arviointiin neljän eri videosisällön kanssa, kun äänen- ja kuvantoiston leveyttä rajoitettiin. Tämän lisäksi osallistujilta kerättiin vapaita kuvauksia heidän antamiinsa laatuarviointeihin vaikuttaneista seikoista. Osallistujien yksilölliset taipumukset kokea uppoutumista arvioitiin ennen koetta kyselylomakkeen avulla.
Tulokset osoittavat, että videon leveys on määräävä tekijä arvioitaessa havaittua laadun heikkenemistä. Myös äänenleveydellä oli merkitystä, kun videonleveys oli suurimmillaan. Taipumus kokea uppoutumista ei ollut merkittävä tekijä laadun kannalta tässä tutkimuksessa. Videosisällön merkitys oli vähäinen. Vapaille kuvauksille suoritettu rajoitettu korrespondenssianalyysi ehdottaa huonoon havaittuun laatuun vaikuttaviksi tekijöiksi äänen väärän tulosuunnan, rajoitetun videonleveyden ja puuttuvan tärkeän sisällön.

Avainsanat: Kokemuksen laatu, moniaistinen havaitseminen, läsnäolon tunne, uppoutuminen, audiovisuaalinen näyttö

Preface

This master's thesis was carried out in the Department of Signal Processing and Acoustics of Aalto University School of Electrical Engineering during the year 2011. The research leading to these results has received funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement no. [240453].

I wish to thank D.Sc. Ville Pulkki for giving me the opportunity to work on this topic and for his guidance and support. His professional expertise has helped me to avoid many pitfalls and hazards lurking in the land of science. My gratitude also goes to my co-workers in the Acoustics Lab for maintaining a stimulating work environment, and for helping me overcome every problem I encountered during this work.

I would like to thank Toni Virtanen and Timo Säämänen from Psychology of Evolving Media and Technology research group at University of Helsinki. Their comments and suggestions were invaluable in designing the experimental part of this thesis and analyzing the results. They are also responsible for introducing me to the more humane side of technology.

Finally, I wish to thank my friends and family, especially my beloved Anna, for their support all through my studies. Special thanks to my parents for teaching me the importance of education.

Helsinki, December 28, 2011

Olli Rummukainen

Contents

List of Abbreviations

2D	two-dimensional
3D	three-dimensional
A1	Primary Auditory Cortex
ACR	Absolute Category Rating
CAVE	CAVE Automatic Virtual Environment
CSF	Contrast Sensitivity Function
DCR	Degradation Category Rating
DirAC	Directional Audio Coding
ERF	Egocentric Reference Frame
FOV	Field of View
HD	High-Definition
HMD	Head Mounted Display
HRTF	Head Related Transfer Function
IBQ	Interpretation-based Quality
IC	Inferior Colliculus
ILD	Interaural Level Difference
IMAX	Image Maximum
ITD	Interaural Time Difference
ITQ	Immersive Tendencies Questionnaire
ITU	International Telecommunication Union
LGN	Lateral Geniculate Nucleus
LSO	Lateral Superior Olivary Nucleus
MGN	Medial Geniculate Nucleus

MOS	Mean Opinion Score
MSO	Medial Superior Olivary Nucleus
OPQ	Open Profiling of Quality
PC	Pair Comparison
PQ	Presence Questionnaire
QoE	Quality of Experience
QoS	Quality of Service
SNR	Signal-to-Noise Ratio
SPL	Sound Pressure Level
SSCQE	Single Stimulus Continuous Quality Evaluation
SSM	Spatial Situation Model
SUS	Slater-Usoh-Steed Questionnaire
THD	Total Harmonic Distortion
UHDTV	Ultra High-Definition Television
V1	Primary Visual Cortex
VBAP	Vector-base Amplitude Panning
VE	Virtual Environment
WFS	Wave Field Synthesis

List of Figures

Chapter 1

Introduction

1.1 Motivation

Current signal processing techniques and affordable hardware facilitate the reproduction of lifelike spatial audio and high-definition video even in private home theaters. Apart from home theaters, virtual reality and immersive audiovisual systems have matured during two decades since the birth of the first CAVE Automatic Virtual Environment in 1991 [14] to a point not too far from Star Trek-like holodecks. Similarly, multimedia services are growing in popularity due to the evolution of digital communication systems. High-definition video can be streamed to our living rooms at low price and teleconferences can be arranged with participants from all over the world.

This progress raises important questions on how to evaluate the quality of these new services and installations in order to offer optimized experiences for customers while simultaneously saving money and bandwidth. The quality we estimate our systems are able to produce in terms of technical quantities, such as signal to noise ratio and total harmonic distortion, does not directly translate to the quality we are perceiving as users of the services [34]. On the other hand, evolving audiovisual technology facilitates studying the human perceptual system and behavior like never before in a controlled environment.

A lot of effort has been put into understanding the perceptual mechanisms affecting subjective quality experience with single modality services, *e.g.* audio only systems like the radio. This has resulted in objective quality metrics based on the human perceptual system for either audio or video. The downside with these metrics, regarding more complex systems, is that human is a multimodal being by nature and unimodal quality metrics cannot accurately reflect the true perceived quality of experience in a service offering information for multiple senses [26, 98, 101].

Our various senses can simultaneously sample different regions of space around us [16]. We can, for example, hear seagulls fly above our head without seeing them while we are eating an ice cream by the sea. This scenario involves auditory, vision, olfactory, tactile and gustatory senses, and our perception of the world depends on the integration of information from all of them. In addition, attention plays a key role in guiding our perception. We can choose to pay more attention to the sky in order to protect the ice

cream from attacking birds, while failing to observe the birds we could end up losing our ice cream.

1.2 Objective

Objective of this study is to investigate the relative importance of the spatial width of audio and video reproduction on the perceived quality of experience in a realistic immersive audiovisual environment. The goal is to gain psychophysical knowledge of the multisensory integration system of the human brain and to identify the factors contributing to the perceived audiovisual quality in a naturalistic setting. The quality of experience is evaluated with subjective, perceptual-based methods in an immersive audiovisual environment with a 3D auditory display and a video screen with 226° field of view.

1.3 Organization

This thesis is divided into six chapter, first one being this introduction. The principles of audiovisual perception and attention are overviewed in Chapter 2. In addition, the neural basis for auditory and visual perception are briefly discussed here. Immersive audiovisual systems and enabling technologies are presented in Chapter 3. Chapter 4 reviews the previous work related to evaluating objectively and subjectively the combined audiovisual quality perception in order to find the state of the art in the field. In Chapter 5 the experimental setup and procedure that are part of this thesis are presented with the obtained results. Finally, conclusions and future work directions are given in Chapter 6.

Chapter 2

Human perception

In this Chapter, the anatomy of the auditory and visual sensory organs, and the related neural processing enabling the conscious perception, are presented. Furthermore, higher cognitive functions, such as attention and multisensory effects, affecting the human perception are elaborated.

2.1 Auditory system

In this Section, the anatomy of the human ear is reviewed with an emphasis on the inner ear. Also the neural processing steps along the central auditory nervous system are presented, and aspects of the auditory perception of space are discussed.

2.1.1 Anatomy of the ear

The auditory system begins with the sensory organ sensitive to air pressure variations, the ear. The ear can be divided into three sections: outer, middle and inner ear. Each section has its own function in transforming the pressure waves from the surrounding environment into neural activity, and finally to a sound percept by the central nervous system. The structure of the ear is depicted in Figure 2.1.

Outer and middle ear

Outer ear collects pressure waves from the surrounding air and directs them to the middle ear. The visible part of the outer ear is called the *pinna*, which is formed mainly of cartilage without any useful muscles. The pinna's surface has many bumps and grooves that have a special function in filtering the incoming audio signal. Center portion of the pinna is called the *concha* which is a bowl-like starting point of the external *auditory canal* leading to the *ear drum* or the *tympanic membrane*. The ear drum is the boundary to the middle ear. [100]

The outer ear causes an increase of approximately 10 to 15 dB in the sound pressure level

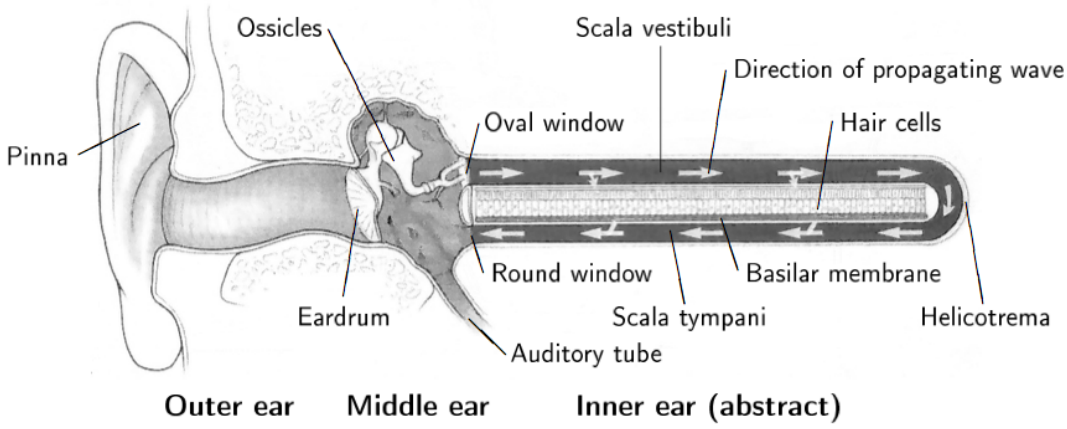

Figure 2.1: The outer, middle and inner ear. *From [69]*

(SPL) in a frequency range of about 1,5 kHz to 7 kHz. This is an essential frequency range for human speech. The increase is due to the resonances of the concha and the ear canal and ear drum. Another effect is the spectral coloration caused by the upper torso, head and the outer ear. The colorations can be described with a *head-related transfer function* (HRTF) , which is elevation and azimuth dependent, and also different for each individual. Essentially, it describes how an audio signal from a certain point in space is transformed by the physiology of the human before it reaches the ear drum. [100]

The ear drum is set to vibration as the pressure waves travel along the ear canal. The mechanical vibration is conducted to the three ear bones, or *ossicles*, called *malleus*, *incus*, and *stapes*. The bones are located in the middle ear cavity and their main function is to transmit the ear drum vibration to the fluid of the inner ear through the oval window membrane. The middle ear cavity is a closed space whose pressure can be adjusted through the *auditory tube* or the *eustachian tube* to adapt to atmospheric pressure changes.

An important aspect of the ossicular chain is the impedance matching it provides between the air in the ear canal and the fluid in the inner ear. The fluid needs more pressure to vibrate than air can provide by itself. This impedance matching is provided by the ratio of the oval window surface to the ear drum surface and also by the lever mechanism of the ossicles, which amplifies the force at the ear drum. As a consequence, the pressure at the oval window is about 20 times greater than at the ear drum and the sensitivity of hearing is very good. The middle ear has also a protective function, due to the *acoustic reflex*, that causes muscles attached to the ossicles contract and attenuate the sensitivity of hearing mainly at low frequencies. The attenuation is 12-15 dB at 500 Hz and 0 dB at 2000 Hz. The reflex occurs with high sound pressure levels when the inner ear neurons would saturate, and the sound would possibly damage the hearing. [35, 100]

Inner ear

The inner ear is not only related to hearing but also to the sense of balance. The *vestibular system* contains the *semicircular canals* and the *vestibule* that help to maintain the body's

equilibrium. However, the vestibular system is not discussed here further, and only the hearing related *cochlea* of the inner ear is considered in this section.

The cochlea is a spiral shaped hollow tube made of bone and it plays a key role in transforming the mechanical vibrations into neural signals. The length of the tube is about 32 mm and diameter about 2 mm, and it is divided into three chambers. The *scala vestibuli* and *scala tympani* are connected through an opening at the apex of the cochlea, called the *helicotrema*. The *scala media* is separated from the other chambers by *Reissner's membrane* and *basilar membrane*. Basilar membrane also holds the *organ of Corti*, which contains the auditory receptor neurons.

At the beginning of the tube there are two membranes: an oval window and a round window. The stapes is attached to the oval window and together they work like a piston setting the fluid in the scala vestibuli and scala tympani, the *perilymph*, into vibration. This creates a propagating wave in the two chambers towards the round window at the base of the scala tympani, as shown in Figure 2.1. This wave bends the flexible basilar membrane and causes the organ of Corti to respond and send impulses to the cochlear nerve.

The basilar membrane has a non-uniform structure that determines how it responds to the propagating wave. First, the membrane widens as it approaches the apex, and second, the stiffness decreases from the base to the apex. Thus, the membrane is narrow and stiff at the base of the cochlea and wide and flexible at the apex. In consequence, the propagating wave in the cochlear fluid causes a traveling wave along the basilar membrane. The distance the wave travels depends on the frequency, so that high frequencies cannot travel far because the stiff base dissipates most of the energy. Low frequencies generate traveling waves that can reach the flexible and wide end of the membrane at the apex.

In effect, the basilar membrane creates a frequency map, or a place code, of the incoming sound wave so that different frequencies cause maximal vibration at different regions of the membrane. This place code is the basis for the neural coding of pitch and can also explain the frequency selective behavior of the auditory system. Frequency selectivity can be described with *critical bands*. Width of the critical bands depends on frequency and they get wider as frequency is increased. Tones within one critical band are thought to be analyzed together by the central nervous system. [4, 35]

The auditory receptors in the organ of Corti are called *hair cells*, named like this because of the hair-like *stereocilia* extending from the top of the cell. The hair cells are divided into two groups along the organ of Corti, inner (about 3500 cells in one row) and outer hair cells (about 15000-20000 cells in three rows). The bending basilar membrane also causes the stereocilia to bend and depending on the type of movement, the corresponding hair cells will respond by firing action potentials.

The hair cell axons converge to the *spiral ganglion*, whose ganglion cells are responsible for sending the auditory information to the brain. The inner hair cells are mostly responsible for producing the input to the spiral ganglion, despite the fact that the outer hair cells outnumber the inner hair cells by a factor of 3 to 1. About 95 % of the spiral ganglion neurons receive their input from the inner hair cells. This means that approximately one inner hair cell feeds ten ganglion cell's neurites. The situation is the opposite for the outer hair cells, where numerous outer hair cells feed one spiral ganglion cell. [4]

The function of the outer hair cells is different than that of the inner hair cells. They work as a *cochlear amplifier* that amplifies the movement of the basilar membrane. The cochlea receives feedback from the brain through efferent fibers that synapse onto the outer hair cells and activate motor proteins that can change the length of the hair cell. In this manner the ear not only receives sound but also creates it. The amplification increases the movement of the basilar membrane 100-fold larger than it would be without it, and greatly enhance the sensitivity of hearing.

2.1.2 Auditory pathway

The axons of the cells in the spiral ganglion form the auditory part of the XIII cranial nerve, or the auditory-vestibular nerve. Before the auditory information reaches the cortex several intermediate nuclei process and relay the information to one another in a parallel manner. The left and right auditory pathways are also interconnected in many points during the processing chain. The central auditory pathway is depicted in Figure 2.2.

The sound information is presented in a *tonotopic* mapping meaning that the sound information is organized according to the frequency content of the stimulus. The tonotopic mapping is preserved throughout the auditory pathway from the basilar membrane to the *primary auditory cortex* (A1). The auditory path beyond the cochlea begins from the dorsal and ventral *cochlear nuclei* in the brain stem. The auditory nerve innervates both the nuclei ipsilateral[1] to the cochlea where they originated from. The cochlear nucleus is thought to preprocess and route the spectral information coming from the cochlea. Henceforth, there are multiple possible, parallel paths through the auditory system and all the connections aren't yet well understood. [100]

One particularly important path is described here beginning from the *superior olivary complex*, that receives axons from the ventral cochlear nuclei on both sides of the brain stem. Superior olive is further divided into *lateral* (LSO) and *medial superior olivary nuclei* (MSO), with both nuclei performing low level signal analysis on the signals arriving from the two ears. The LSO is considered to be active in detecting interaural level differences (ILD) and the MSO in detecting interaural time differences (ITD). Both ILD and ITD values are essential in the human binaural hearing and in detecting the location of a sound source in a space. [100]

The next major nucleus in the pathway is the *inferior colliculus* (IC) situated in the midbrain. The IC receives inputs from the superior olivary complex but also monaural inputs from the cochlear nuclei. The function of the IC is thought to be to combine interaural information and spectral information obtained from the lower level processes in the brainstem. This could facilitate more refined processing of sound in two or three dimensions. Furthermore, the IC is situated right below the *superior colliculus*, which is an area involved in processing of visual stimuli. The colliculus structure is a place, where multimodal information is partly integrated and routed. [25]

The IC projects to the *medial geniculate nucleus* (MGN) located in the thalamus. MGN is the gateway to the primary auditory cortex located in the temporal lobe. The projection from the MGN towards A1 is called the *acoustic radiation*. The information is projected

[1] The same side of the body (*vs.* contralateral)

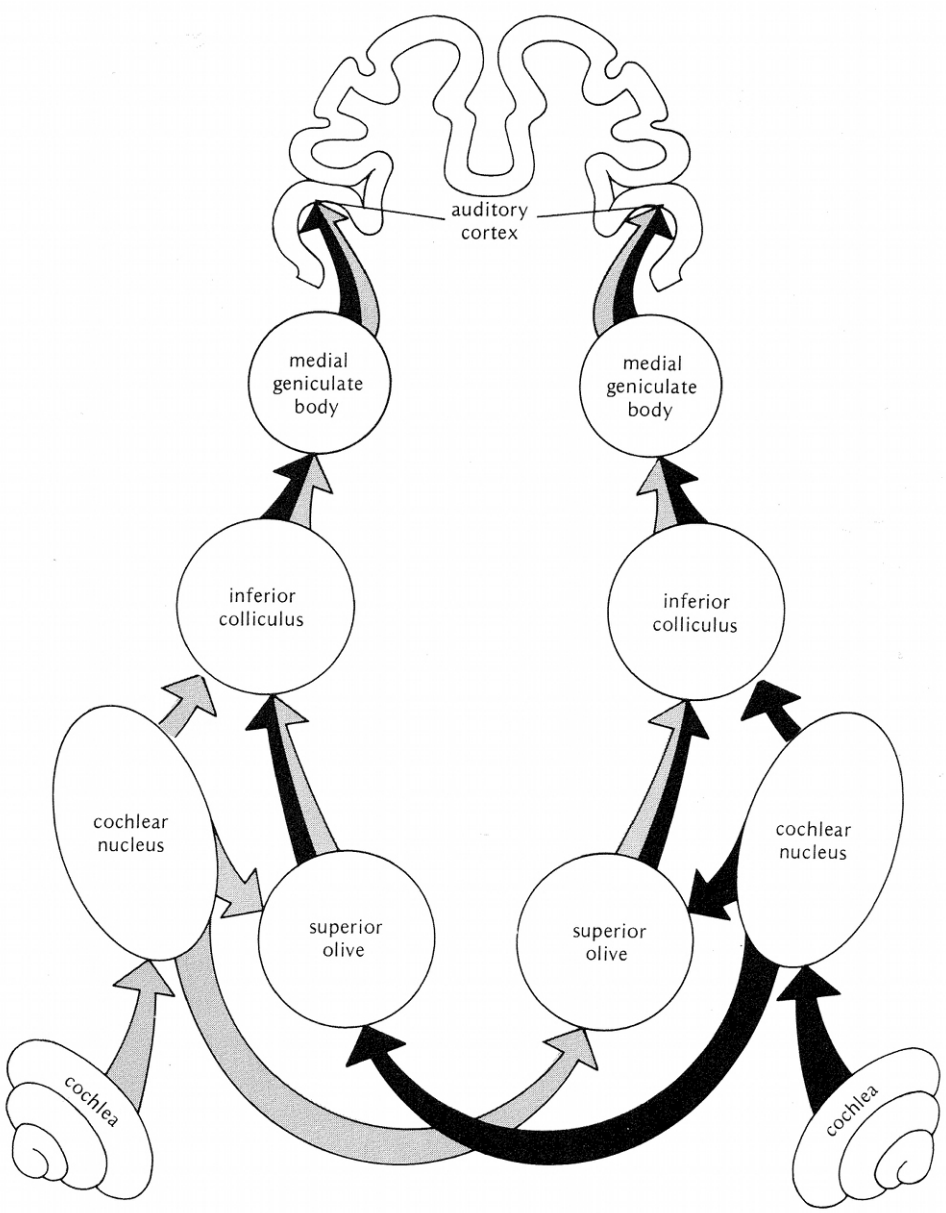

Figure 2.2: Central auditory pathway. *Modified from [100].*

bilaterally meaning that both hemispheres receive information from both cochleas. The tonotopic mapping is still preserved in the A1, where distinct *isofrequency bands* containing neurons with fairly similar characteristic frequencies can be identified. Apart from the characteristic frequency tuning, some neurons in the auditory cortex are intensity tuned or tuned to respond to a particular type of sound, *e.g.* clicks, noise bursts or more complex frequency-modulated sounds. The receptive fields in the A1 are still not well understood, but different temporal structures can be identified with neurons responding with a transient activation to a short sound stimulus and others having a sustained response. [4, 100]

Similarly to the visual pathway, the cortical processing of auditory stimuli extend beyond the primary auditory cortex. The dual-stream hypothesis of information processing is

the most popular current theory of processing done in higher cortical areas. After area A1 the processing is split into postero-dorsal "where" path towards the parietal lobe and antero-ventral "what" path towards the temporal lobe. The "what" path is hierarchically organized progressing from simple low-level feature analysis to selectively identifying complex sounds based on the low-level features. The hierarchical structure facilitates auditory pattern recognition and object identification in the "what" path.

The "where" path is critical in perceiving the auditory space and motion, but also in the processing of speech and language. Therefore, the dual-stream theory can be insufficient to fully describe the cortical auditory processing and more streams may exist. In addition, the auditory processing is asymmetric in the human brain with the two hemispheres specializing in different aspects of sound. Especially, speech perception and production are left-lateralized in contrast to spatial processing that is right-lateralized. Finally, the information from the two streams is projected to the frontal cortex where the streams eventually converge together and form a unified percept. [58, 59]

2.1.3 Spatial hearing

Spatial hearing, the ability to localize sound sources in a space, is an essential feature of the human auditory system. Although the spatial accuracy of the auditory system is not nearly as good as the spatial accuracy of the visual system, it contributes valuable information to our perception of space.

In the horizontal plane (azimuth) the most prominent cues of localization are the interaural time difference (ITD) and interaural level difference (ILD). Both the ITD and ILD are maximized when the sound is arriving directly from left or right side of the head, and decrease as the sound source is moved towards the median plane. If the sound source is moved in the median plane (elevation), neither ILD or ITD change. The importance of ITD and ILD to sound source localization is frequency dependent. ITD is active in low frequencies in the range 20 - 2000 Hz because the wavelength in this frequency range is larger than the diameter of the human head and meaningful phase differences can be observed. ILD is active in the range 2000 - 20000 Hz because at lower frequencies the sound waves diffract around the head and no significant level differences can be observed.

More accurate elevation cues are obtained by observing how the human torso, head and the outer ear filter the incoming sound signal. The filtering effect is described by the head-related transfer function (HRTF), mentioned briefly in Section 2.1.1, that tells how the sound signal is transformed by the human physiology before it reaches the ear drum. The HRTF is azimuth and elevation dependent and also different for each individual. HRTFs can be measured ideally in a free field environment by placing a microphone near the ear drum and moving the sound source to various angles. In reality, placing the microphone to the ear drum is difficult and the effect of the ear canal has to be approximated or an artificial dummy head used. [10, 35]

In addition to direction information, also distance of the sound source can be perceived to some extent. In anechoic conditions estimating the distance is difficult and the estimate is based only on sound intensity. In these conditions the sound sources are localized to maximally about 10 m away from the listening position as the so called *acoustic horizon* is

reached. If reflections and late reverberation are present, the distance evaluation is easier by evaluating the direct-to-reverberant energy ratio. This ability of the human auditory system is largely based on learning and on experiences from different sound scenes. [35]

The spatial cues can be exploited in designing sound reproduction methods or creating virtual environments with realistic 3D audio. The sound can be reproduced either with multiple loudspeakers producing the desired sound field, or relevant parts of it, in a space or with headphones producing the sound field to the ear canals, as it would be in a real environment. With headphones it is necessary to simulate the sound source, acoustics of the space, and also the physiology of the listener. Headphone setups are usually easier to implement because the acoustics of the listening space can be neglected. Nevertheless, with headphones the sound can be easily localized inside the head, for example with non-individual HRTFs, and realistic spatialization is difficult to achieve. Loudspeaker setups remove the need for taking human physiology into account. [35]

In real world, localization of sound sources is not solely dependent on auditory cues. Knowledge of different sound sources helps in placing them approximately to correct directions (*e.g.* airplanes fly in the sky). Also visual cues influence the localization of sound sources, and possibly the externalization of sounds in headphone listening is enhanced by an accompanying visual stream. Audiovisual perception is further discussed in Section 2.4. [69]

2.2 Visual system

In this Section, the basic anatomy of the human eye and structure of the retina are reviewed along with an overview of the related neural processing. At the end of the section visual perception of space is discussed.

2.2.1 Ocular anatomy

Human visual system starts from the light sensitive sensory organ, the eye. Eye's function is to collect emitted or refracted light rays from the environment and focus them on the retina to form images. Cross-section of the eye is shown in Figure 2.3.

As light rays enter the eye, they pass first the glassy external surface of the eye, called *cornea*. The cornea is an integral part of the rest of the eyeball wall, *sclera*, which is the white matter of the eye seen from outside. The sclera is connected to three pairs of *extraocular muscles* moving the eyeball in its eye socket in the skull. Behind the cornea is a chamber filled with fluid known as *aqueous humor*, whose duty is to nourish the cornea, which has no blood vessels.

The cornea has most of the refractive power, or the focusing power, of the eye. The refractive power of the cornea is about 40 diopters, which is two thirds of the refractive power of the whole eye. After the cornea and the fluid chamber lie the *pupil* and the *iris*. The pupil is the opening through which light enters the eye and the iris controls the size of the opening determining how much light is allowed to reach the retina. After the iris, there is a transparent lens suspended by ligaments attached to *ciliary muscles*.

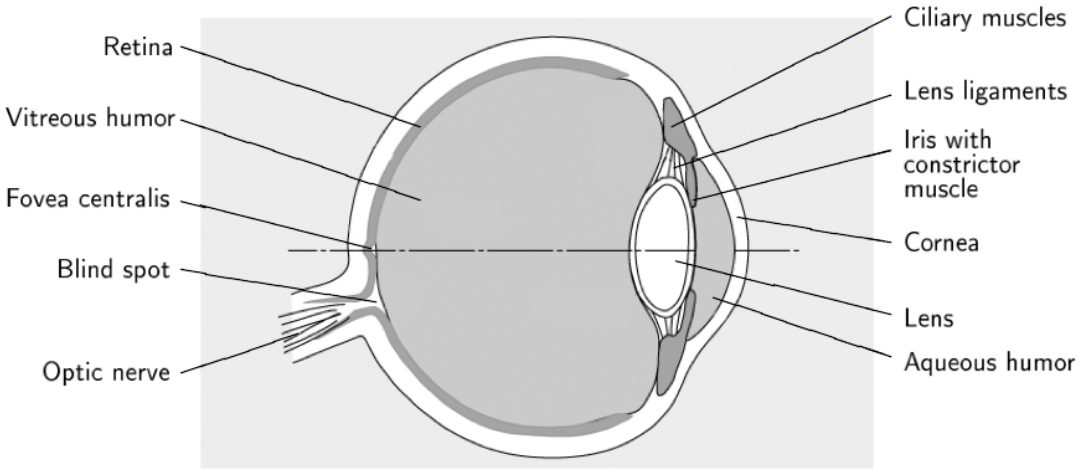

Figure 2.3: Anatomy of the human eye. *From [69].*

These muscles control the shape of the lens and, consequently, the amount of refraction by the lens. The lens contributes the last one third (20 diopters) of the refractive power of the eye. Changes in the shape of the lens enable the eye to focus on different viewing distances. The space between the lens and the retina is filled with jellylike, viscous fluid *vitreous humor* whose pressure keeps the eyeball spherical. [4, 80]

2.2.2 Retinal anatomy

The most important part of the eye for visual perception is the nervous structure called *retina*. The retina occupies back part of the eyeball, and this is where the optical image is focused on. The retina essentially converses light energy into neural activity by sending electrical impulses into the optic nerve. The visual information pathway from the retina to the optic nerve goes through three distinct layers of neurons, namely, *photoreceptors*, *bipolar cells* and *ganglion cells*. Besides the direct path there are also lateral connections in the retina provided by horizontal cells and amacrine cells. Through these connections the neurons can laterally affect surrounding cells. [4]

Photoreceptors are the only light sensitive cells of the retina. There are two types of cells in this layer, *rods* and *cones*, with different appearance and functionality. Rods are about 20 times more numerous in the retina than cones, and they are responsible for vision in dim light. Rods don't respond to bright light in contrast to cones, who are responsible for the ability to see fine detail and colors in bright lighting situations. Rods and cones are unevenly distributed through the retina. The peripheral retina has higher number of rods than cones, whereas the cones are concentrated in the *fovea*, the center point of the retina about 2 mm in diameter specialized in high-resolution vision. The high number of rods in the peripheral parts of the retina makes the peripheral vision more sensitive to light for example in night time. On the other hand, peripheral vision is poor at resolving fine details in daylight. Area where the optic nerve exits the retina is called the *optic disk*. There are no photoreceptors in this area and, in consequence, no sensation of light can occur. The visual world is still continuous despite the blind spot because the brain fills in

the perception of this area. [4, 27]

The retina has about 125 million photoreceptive cells but only one million ganglion cells producing the output from it. The bipolar cell and ganglion cell layers must efficiently compress the information received from the photoreceptors. The compression ratio changes progressively when moving from the center of the fovea towards peripheral regions. In and near the fovea, each cone feeds one bipolar cell and each bipolar cell feeds one ganglion cell. Moving away from the fovea this rule changes, so that more and more photoreceptors converge to one bipolar cell and multiple bipolar cells converge to one ganglion cell. As a consequence, only the most important aspects of an image are extracted by the retina, and the compression ratio of 125:1 is possible to achieve without dramatic loss of visual acuity. The axons leaving the eye carry visual cues, such as spatial contrast and temporal frequency, that are far more sophisticated and better suited for further analysis by rest of the central nervous system than the raw data acquired through the photoreceptor cells. [27, 47]

2.2.3 Visual pathway

The ganglion cells are grossly divided into two groups with differing functionalities. There are *magnocellular* ganglion cells that are sensitive to motion and have thick axons meaning they can transmit information fast. The second group are the *parvocellular* ganglion cells, that are sensitive to form and color and have thinner axons. The ganglion cell axons exit the eye through the optic disk and together form the second cranial nerve, or the optic nerve. The optic nerves and the optic tracts towards the visual cortex are shown in Figure 2.4.

The optic nerves from the two eyes meet each other at the *optic chiasm* and half of the information from each eye crosses sides. As a result, the left visual hemifield is projected to the right optic tract and vice versa. Target of most of the axons in the optic tracts are the right and left *lateral geniculate nucleus* (LGN) situated in the dorsal thalamus. Both the magnocellular and parvocellular paths converge to the LGN, but the streams are kept separate. LGN is a gateway to the primary visual cortex (V1) situated in the occipital lobe of the human brain. V1 is the first cortical area involved in the processing towards the conscious visual perception.

The projection from the LGN to the visual cortex is called the *optic radiation*. However, LGN is not simply a relay station on the visual pathway. About 80 % of the input to the LGN comes as feedback from the visual cortex, and this feedback is considered to significantly alter the behavior of the LGN. There are also incoming connections from the brainstem whose activity is related to alertness and attentiveness. In result, LGN is the first site in the visual pathway, where visual perception is affected by emotions.

The organization of the central visual system is retinotopic, meaning that optic nerve fibers from a particular region of the retina go first to a particular region of the LGN and then V1. Essentially this means, that the surface of the retina is mapped onto the surface of the primary visual cortex. The mapping is however distorted, because the photoreceptors are not uniformly distributed in the retina and also the compression of information results in overlapping *receptive fields*, that activate larger areas of the V1 than one-to-one mapping

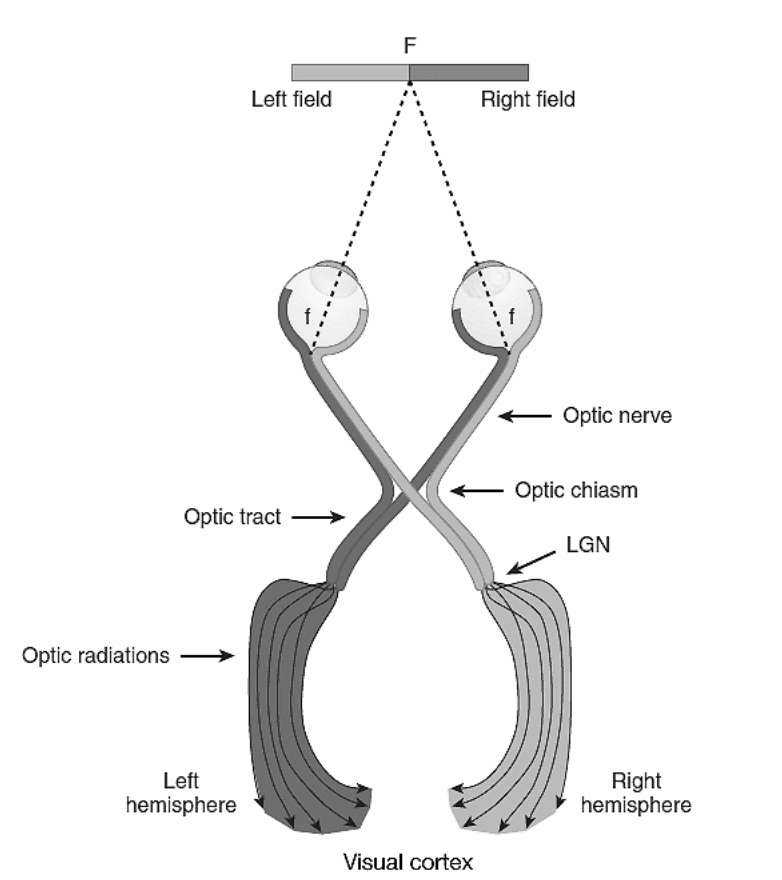

Figure 2.4: Postretinal pathways. *From [80]*

of the visual image would assume. A receptive field is an area in visual space where changes in lighting cause changes in neuronal activity.

V1 extracts more complex information, such as orientation, direction and color, than the retina. Cells in the V1 are categorized to simple and complex cells depending on the structure of their respective receptive fields. Simple cells might respond to elementary *on* or *off* stimulus excelling at detecting lines and edges, whereas complex cells would respond to stimuli oriented and directed in a particular manner. The receptive field of a simple cell is thought to consist of the output of several LGN neurons, with distinct regions of excitatory and inhibitory properties. It has been suggested that the receptive field of a complex cell is constructed from the output of multiple simple cells, but no definite resolution exist.

Primary visual cortex is the first cortical area to receive information from the LGN, but the cortical processing extends well beyond the V1. There are at least 20 other distinct areas that contain a retinotopic representation of the visual world, similarly as V1, but specialize in analyzing different aspects of the image. Two major streams originating from the V1 can be identified. First stream, *ventral stream*, is called the "what" path because it is critical for object recognition. The second stream, *dorsal stream*, is called the "where" path because it is important in motion perception and localization in visual space.

The streams extend from the primary visual cortex towards parietal and temporal lobes, respectively, as shown in Figure 2.5. Information from these two streams must finally be combined with memory and integrated to form a unified percept. [4, 27, 80]

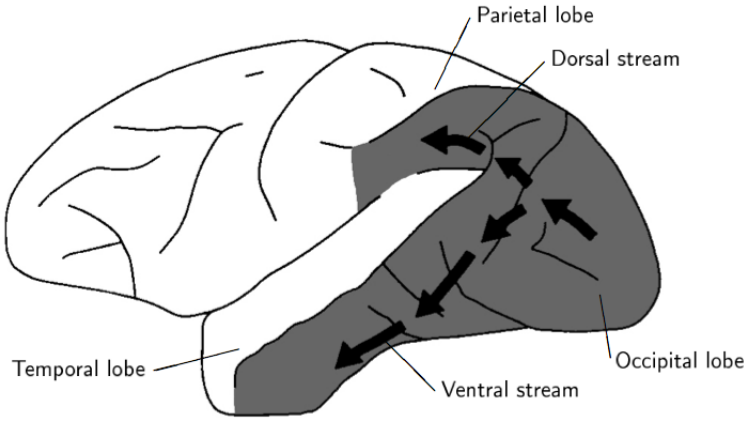

Figure 2.5: Dorsal and ventral visual streams originating from the primary visual cortex. *From [69]*

The human visual system transmits more information to the brain than any other sensory system. In consequence, a relatively large proportion of the human brain is devoted to visual perception. The processing of visual information is done in a parallel manner. Beginning from the magnocellular and parvocellular paths, from the retina to the thalamus, visual information is divided into separate data streams. Divergence of the data is maximized when it is passed to higher cortical areas after the primary visual cortex. Similarly to auditory processing, a topographic mapping of the sensory surface is retained throughout the processing chain. How, exactly, are the separate data streams integrated and where is the conscious perception of the visual world formed, remain as open questions. [4, 93]

2.2.4 Visual perception of space

Spatial resolution

Spatial resolution, or visual acuity, describes the visual system's ability to detect and resolve various size and contrast stimuli defined by luminance. The corresponding psychophysical measure is the contrast sensitivity function (CSF) depicted in Figure 2.6.

The function is a band-pass type filter with maximum sensitivity at about 4 cycles/degree of spatial frequency for black and white lines. This means, that humans can detect spatial frequency of 4 cycles/degree at lowest contrast and more contrast is required for other spatial frequencies. The arrow in the Figure denotes the high-frequency cutoff value, where higher frequency details cannot be resolved even with 100 % contrast. The high-frequency cutoff is typically from 40 to 60 cycles/degree. [80]

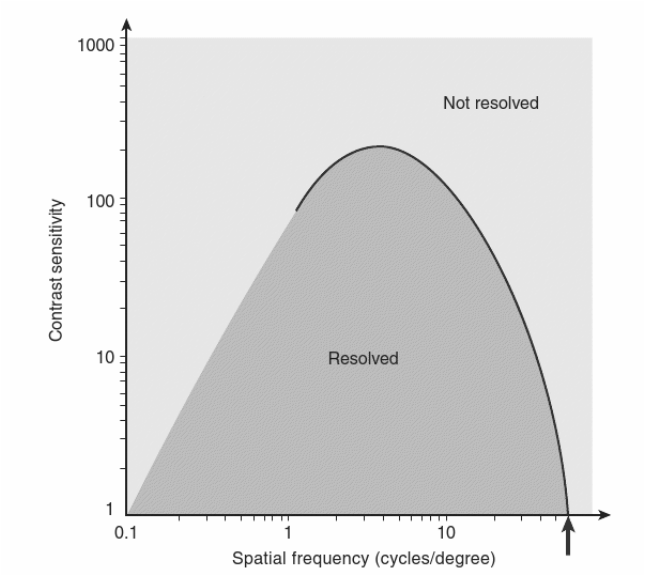

Figure 2.6: A typical adult contrast sensitivity function (CSF). The contrast decreases as Contrast sensitivity increases. *From [80]*

Depth cues

Humans can perceive depth even with one eye shut, monocularly. Most prominent monocular depth cues are pictorial in a way that they can be presented in a 2D representation, for example in a painting.

- *Size* - Relative size cue is important when viewing unknown objects whose sizes can be compared. Larger objects are perceived as closer. Familiar size cue is used with known objects whose size can be compared to the size of a typical example of the same object. If the observed object is larger than the assumed object, the object is perceived as being close.
- *Interposition* - Interposition occurs when a scene is partially occluded by an object. The occluding object is perceived as being closer than the background.
- *Linear perspective* - Due to the properties of the lens, parallel lines seem to converge. For example railroad tracks in a photograph seem to approach a point in the horizon creating an illusion of depth.
- *Accommodation* is the process when the lens is focused to a certain distance by the ciliary muscles. Accommodation can provide monocular cues of depth by sending information about the varying muscle tension.
- Other monocular depth cues are *texture*, *clarity*, and *lighting and shadows*. All these cues give information about the objects' relative position, direction and extension of their surfaces.

Monocular visual field is about 170° wide. In humans, the two eyes are facing the same direction resulting in overlapping fields of vision from slightly different vantage points.

Full field of view extends 200° horizontal, of which the central 120° is an area of binocular overlap. The different monocular views in this area lead to *stereopsis*, that is a binocular form of depth perception. Main cues of stereopsis are *convergence* and *disparity*. In convergence, the two eyes fixate on an object at a close distance and turn inwards. The distance of the object can then be estimated by the central nervous system from the deviation angle of the eyes. In disparity, both eyes are fixated on a point in given distance, and its image is projected on the fovea of each eye. Then, another point located nearer or farther away is projected to differing points of the two retinas. This disparity can provide cues for distance in relation to the fixation point.

2.3 Attention in perception

Processing capacity of the human brain is limited, so it must be allocated in an efficient way. Attention is the controlling force guiding our perception. Also, attention is not directed only to external stimuli, but sometimes internal processes like planning ahead or trying to remember some detail require our full attention [41]. In this Section, the main theories of selective attention and perception are reviewed.

2.3.1 Perceptual cycle

Neisser describes human perception as a continuous process in his model, called the *perceptual cycle* [48]. Neisser's perceptual cycle is depicted in Figure 2.7. The cycle shows how *schema*, *exploration* and *object* influence each other in a circular process. In Neisser's model, the schema represents our knowledge about the environment based on previous experience. Schema directs further exploration of the environment by creating expectations and emotions that steer our attention to certain objects and events available in the environment. The environment is sampled and available information picked up to be compared against the existing schema, or, the previously known information about our environment. If the stimuli are recognized, they are given a meaning according to the schema. Unrecognized stimuli will modify the schema and, consequently, direct the environment exploration further.

2.3.2 Selective attention

A classic example of attention guiding our perception is the cocktail party phenomenon. In a crowded and noisy environment, most people are able to concentrate on the current conversation, without being distracted by others too much [11]. However, if someone says their name in an other discussion, people will sometimes notice it and redirect their attention from the current discussion, to find out what others are speaking about them behind their backs [97].

This scenario poses multiple challenging tasks to the brain. First, the relevant information must be analyzed and the distracting noise suppressed. Here cross-modal attention helps by picking up the relevant speech sounds from the auditory input and matching them with the corresponding lip-movements from the visual stream entering the eyes [16]. Second, the

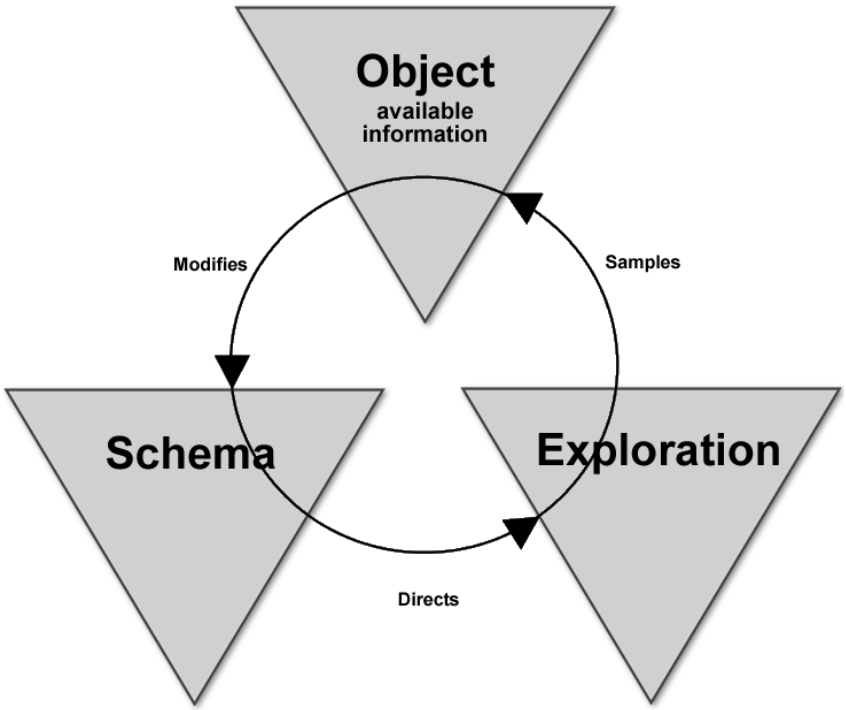

Figure 2.7: Neisser's perceptual cycle. *From [18]*

surrounding noise has to be analyzed to some extent, in order to recognize the individual's name outside the current field of attention.

Traditionally, theories of focused attention try to explain this phenomenon by two different approaches: early selection and late selection. In the early selection model, all the incoming stimuli are preprocessed and filtered by a selective filter on the basis of their physical characteristics. The selected stimuli pass along to further semantic analysis and eventually they become conscious perceptions. The stimuli that are filtered out are not analyzed for meaning and do not reach consciousness. In this model, the fusion of separate physical attributes takes place at the end of the information processing chain and no meaning is given to the physical attributes before that. This model cannot, however, explain how a person's own name can be heard from a distant conversation outside the field of attention.

The late selection model takes this into account by stating that all the physical attributes entering our senses are processed to some extent, whether attention is paid to them or not. A semantic meaning is given to each perceptual object that results from the integration of a set of attributes. The semantic importance of the objects is then weighted against each other and only the most important ones are allowed to enter consciousness. In this model, the fusion of attributes takes place at an early level of the information processing, and meaning is given to objects before the end of the chain. [15]

Information flows in two directions in the brain: bottom-up and top-down. Bottom-up flow means parsing of the information entering our sensory system. Information is processed beginning from the individual physical attributes of the input, and progressing towards higher levels of cognition through feature analysis. Abrupt appearance of, for example,

a visual event in the visual field can capture attention and redirect it towards the event [99]. Automatic allocation of selective attention is called exogenous attention, since it is triggered stimulus-driven by some event in the space surrounding the human [16].

Other mode of information flow, top-down, is related to endogenous attention, since the attention is directed by decisions originating from higher cognitive levels of the mind. Expectations and predictions can voluntarily direct attention towards a particular event that is considered to be important for the individual.

2.4 Audiovisual perception

In this Section, known effects regarding audiovisual perception are presented. Also, the cross-modal effects of multimodal perception on the perceived audiovisual quality are discussed.

2.4.1 Multisensory integration and conflicts

Multisensory integration, or the binding of a pair of auditory and visual stimuli, is dependent on a few preconditions that can be divided to structural and cognitive factors. Structural factors include the spatial and temporal coincidence, meaning that the events in the two modalities have to be timed and located within certain thresholds, and temporal correlation between the stimuli. Cognitive factors are more top-down constructs such as an assumption of unity or semantic congruence of the stimuli. The cognitive and structural factors are not, however, easily distinguished from each other. The bottom-up integration of structural variables also enhances the formation of cognitive assumption of unity, for example. [86]

A classic example of the interaction between hearing and vision is the McGurk effect [42], where visual perception induces an auditory illusion. The effect was first noticed when the influence of vision on speech perception was studied. In the test, a video of a talking head pronouncing the syllable */ga/* was dubbed with the syllable */ba/* and shown to the test subject. As a result normal adults reported hearing */da/*. The effect also proved to be very robust in a sense that knowledge of the illusion does not help in recognizing the syllables correctly. Closing the eyes reverts the previously heard */da/* into */ba/*, but when opening the eyes the syllable is again reverted back to */da/*.

Another multimodal integration effect is *intersensory bias*, or *ventriloquism* with audiovisual stimuli, where the perceived location of the non-dominant modality shifts towards the other. In ventriloquism, visual perception steers the localization of an auditory event towards visually anticipated direction, when the visual and auditory cues conflict [8]. However, recently Alais and Burr [1] pointed out that also sound can capture vision perception when the visual stimulus is severely blurred. With less blurred stimuli, neither of the modalities was found to be dominant and localization followed the mean position. Also, the combined bimodal localization accuracy was found to be superior compared to either visual or auditory localization alone.

Besides spatial mismatch, audiovisual cues can conflict also temporally. The asynchrony

detection is not symmetric across the modalities, as audio lead is detected more easily than audio lag [8]. This asymmetry can be explained by considering the speed difference of sound and light propagation in real world environments. Auditory perception can never lead visual perception in the real world, and such a situation feels very unnatural when reproduced artificially. Also, the stimulus content has an effect on the detection thresholds. Multiple studies suggest different numeric values for the thresholds, and, according to the most conservative estimates, audio lead of 75 ms and lag of 90 ms are detected [40]. ITU (International Telecommunication Union) has published a recommendation concerning synchronization thresholds for audio and video components in television signal [61]. The recommendation states that maximum tolerated audio lead is 20 ms and lag 40 ms.

2.4.2 Impact of cross-modal interaction on perceived quality

A number of studies have concentrated on researching the influence of cross-modal interaction on perceived audio or visual quality. The interaction has been found to substantially affect the quality perception. Rimell and Hollier [75] found that cross-modal interaction is present for all multimedia content. Especially talking head material was found to have high interaction between auditory and visual modalities, and to be sensitive to distortions. Ensuring that distortions do not occur in the region around the eyes and mouth, would help to improve the perceived quality. They also made an overall conclusion that: *"the quality of one mode affects the perceived quality of the other mode and a single mode should not be considered in isolation"*. Similar results were achieved by Beerends and De Caluwe [7], who also noted that video quality is the dominant element in overall quality perception.

The impact of content on the perceived quality was observed by Korhonen *et al.* [36] when they studied how asymmetric distortion in auditory or visual streams affect the overall perceived quality. They found that the relative importance of audio and video quality is dependent on the content. The authors encouraged further studies on content classification, as they found that current descriptors are insufficient to explain the content dependent cross-modal effects they encountered.

Rimell and Owen [76] studied the effect of focused attention on quality estimation. The results showed, that if the perceiver's attention is focused on a particular modality, their ability to detect errors in the other modality is greatly impaired, when compared to a case of equally distributed attention. Also, they tested how the cross-modal interaction behaves when the attention is on one modality and the question is presented on the other, *i.e* visual question and auditory attention, for example. They found that cross-modal interaction is enhanced for cross-modality attention-question situations and almost eliminated for intra-modality situations. Similarly, Zielinski *et al.* [102] found that focusing on a visual task affects the perceived audio quality while playing a computer game. The effect was, however, listener-specific and, in this study, tested only with one type of audio degradation. Focused audiovisual attention can be exploited in multimedia codec design as proposed by Lee *et al.* [39].

Joly *et al.* [32] evaluated the mutual influence of audio and video on continuous perceived quality of television programmes. The results showed significant impact of a good quality soundtrack on how video impairments were perceived. With good sound quality video

impairments were less annoying. On the contrary, good video quality was noticed to have no impact on the perceived audio quality, when impairments were present in the audio signal. For long time analysis, video quality was observed to be the main factor on the overall perceived audiovisual quality.

Valente and Braasch [90] studied the evaluation of acoustic parameters based on visual cues of a real environment. The participants adjusted the direct-to-reverberant (D/R) ratio of stimulus sounds to match an accompanying visual stimulus. They found that the participants constantly overestimated the amount of direct sound energy, when compared to the measured ratio. The authors conclude that excluding the visual cues can have implications on the perceived quality when evaluating acoustic conditions of a space.

2.4.3 Effect of spatial width of audiovisual reproduction

Latest research concerning the effect of spatial width of audio and video reproduction on the perceived quality is presented here for the experimental part of this thesis. A number of studies have looked into the effect of audio and video reproduction setup separately, but only a few consider them jointly. Prothero and Hoffman [54] studied the effect of widening the field of view on sense of presence. They found significantly higher presence for 105° FOV than for 60° FOV . Their study, however, did not include sound perception. Bech *et al.* [6] studied audiovisual interaction in home theater systems. According to their results, screen size affects the relative importance of audio and video, and also the perceived overall quality. Also, they studied varying the base width of stereo audio reproduction between narrow and wide setting, limited by the width of the television screen and 60°, respectively. Effects were observed only with largest changes in the base width, *i.e.* with narrowest width (smallest 17" television) compared to the 60° width. Finally, they identified auditory impression of "space" to be important for general quality of the system.

Similar results of the effect of stereophonic width was obtained by Bech [5], when using a standard television set. He found that increasing audio reproduction width increases the quality of reproduction of space. Also, listening position had a significant influence on stereophonic listening as the quality decreases as the listening position moves off-center.

Hamasaki *et al.* [24] have studied the advantages of using a 22.2 multichannel sound system with an Ultra High-Definition TV (UHDTV) when compared to 2.0 and 5.1 sound systems. As a result, the 22.2 channel setup was found to produce better sensations of spatial quality and presence than 2D sound systems. However, the authors remark that the evaluation seems to be dependent on the content, and interaction between audio and visual cues should be investigated further. In the present study audio quality attributes such as reality, transparency, and gaudiness were not rated better for the 22.2 setup than for the 5.1 setup. The authors conclude that this might be due to the influence of the picture on the audio impression.

Also Strohmeier and Jumisko-Pyykkö [87], and Reiter [68] studied the impact of loudspeaker setups in audiovisual applications. Strohmeier and Jumisko-Pyykkö experimented with 4 and 5 loudspeaker at two distances with a 15" autostereocopic screen. They found that optimal perceived quality is achieved with 4 loudspeakers at 1 m distance. Reiter, on the other hand, used a large projection screen (2.72 m wide) and different numbers

of loudspeaker channels to examine sound source localization accuracy. He found, that optimum number of loudspeaker channels with a large screen is five. Strohmeier and Jumisko-Pyykkö conclude, that the perceived audiovisual quality is strongly impacted by different audio and visual presentation modes and devices, and future research is needed to examine the impact of different setups on experienced multimedia quality.

Chapter 3

Immersive audiovisual environments

In this Chapter the basic cognitive concepts related to immersive audiovisual environments are first presented and their underlying psychological mechanisms discussed. Next, the enabling technologies for audiovisual immersion are reviewed along with descriptions of a few recent immersive environments. At the end of the Chapter, the audiovisual system deployed in this study is presented.

3.1 Related concepts

Two concepts tightly related to audiovisual setups are immersion and presence. Presence is normally used to somehow examine and measure the sense of "being there", the illusion of being in another place than a mediated environment, or a virtual environment (VE), although consciously being aware that you are not there. Immersion, on the other hand, is something that presence builds on.

3.1.1 Immersion

A person can feel immersion with various types of media. Movies and video games are effective in evoking feelings of immersion. In addition, an interesting book can immerse the reader so deeply he forgets his surroundings and blocks outside distractions. The feeling of immersion is also a highly individual experience and some are more prone to experience it than others.

Witmer and Singer [96] studied immersion as a precondition for presence and theorized that the overall feeling of immersion can be measured on three subscales: involvement, focus and tendency to play video games. They created a questionnaire to assess individuals on these subscales, in order to evaluate their overall tendency to experience immersion. Later, Weibel *et al.* [94] studied the effects of personality traits on immersion and used the questionnaire Witmer and Singer had created. With factor analysis, they identified only two factors from the questionnaire data they collected: emotional involvement and

absorption. The original theoretical division into three subscales could not be supported and immersion appeared to correlate well with tendency to react emotionally during media usage, as well as with ability to focus on a task and block out external distractions.

Furthermore, Weibel *et al.* studied the relation between immersion and the Big Five personality traits. The Big Five is a descriptive model that tries to conceptualize the dimensions of human personality. The Big Five includes *extraversion*, *agreeableness*, *conscientiousness*, *neuroticism* and *openness*, which can be measured through questionnaires. They found that openness to experience, neuroticism and extraversion were positively related to tendency to feel immersed. Openness to experience is an expected result, but neuroticism and extraversion are negatively correlated with each other. The authors hypothesized that emotional involvement can refer to both positive and negative reactions to stimuli and, therefore, unpleasant stimuli can easily elicit strong negative feelings, and eventually immersion, in persons scoring high on neuroticism.

3.1.2 Presence

Slater [82] argues that presence research consists of essentially experimenting with different factors that make up immersion. An equation for presence could be formed with presence on the lefthand side and factors of immersion and individual psychological differences on the righthand side. IJsselsteijn *et al.* [29] have collected the factors thought to underlie the perception of presence:

1. *Extent and fidelity of sensory information* - Technology's ability to produce relevant and accurate sensorial information, for example field of view and spatialization of audio.
2. *Match between sensors and the display* - User's actions should result in corresponding real-time changes in the virtual environment.
3. *Content factors* - How objects and events are represented by the medium and the ability to interact with the content. Also social elements, such as reactions of other users in a shared virtual environment.
4. *User characteristics* - Sensorial acuity and cognitive abilities. Also expectations towards virtual environments and previous experience of such systems.

Presence is thought to consist of spatial and social components [29]. Factors of social presence are mostly dealing with communicating and acting with someone in a shared mediated environment. Understanding social presence is not, however, considered essential regarding this thesis but the focus is on spatial presence which is presented here in greater detail.

Wirth *et al.* [95] define spatial presence as: *"Spatial presence is a binary experience, during which perceived self-location and, in most cases, perceived action possibilities are connected to a mediated spatial environment, and mental capacities are bound by the mediated environment instead of reality"*. This definition is in accordance with an earlier

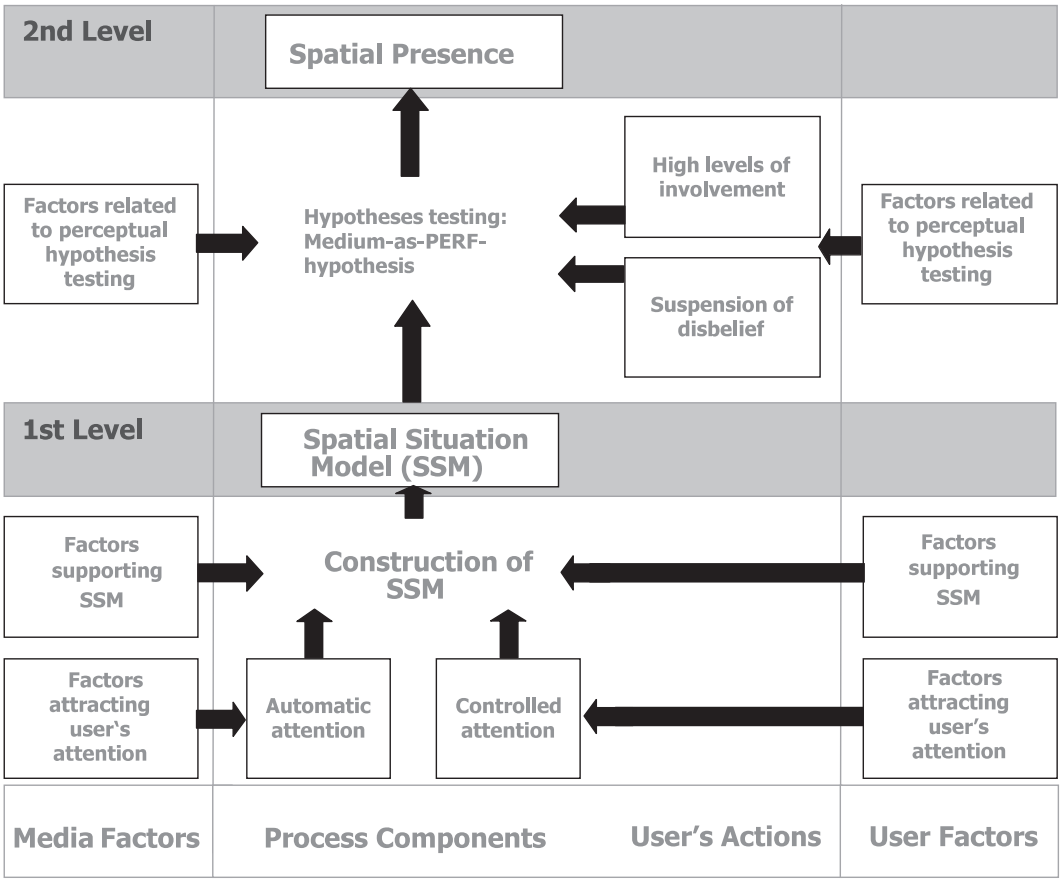

Figure 3.1: Two-level model of spatial presence. *From [95].*

conceptualization by IJsselstein *et al.* [29]. Wirth *et al.* also describe a two-level process model of the formation of spatial presence. The model is presented in Figure 3.1.

First, a mental model of the mediated situation is created. In this stage the user might deal with questions on whether the stimulus is a space or a room, and what kind of a space is it. The allocation of attention in the mediated environment is the basis for a *spatial situation model* and a precondition for spatial presence to occur. Spatial situation model is a mental model the user constructs based on the perceived spatial cues combined with personal spatial memories and cognitions. Also, it is an individual construct, meaning that it may be constructed differently for the same space by different receivers.

The second level of the model describes the formation of spatial presence from the spatial situation model. The process involves testing perceptual hypothesis to find out whether the user perceives him/herself located in the space. The authors use the term *egocentric reference frame* to represent the mental model of the perceived world around an individual. It contains at least the immediate surroundings and is constantly updated according to the sensory input from the environment. In effect, an egocentric reference frame tells the individual where he is in a spatial environment.

In addition to the real world, virtual worlds can give rise to an egocentric reference frame that is different from the user's real-world egocentric reference frame. The media-bound reference frame can only occur if the conditions for a spatial situation model are first fulfilled, *i.e.* the user has a perception of some kind of a space or a room induced by the virtual environment. In such situation, the media-bound egocentric reference frame begins to compete with the real world counterpart because different modalities may offer contradicting information about the surroundings and confuse the perceiver. Eventually, the perceiver must decide which of the reference frames to consider the *primary ego reference frame* in order to reduce confusion and to be able to act. The authors state that spatial presence can only occur if the perceiver accepts the media-bound egocentric reference frame as the primary egocentric reference frame over the real-world reference frame.

Presence is usually measured with retrospective questionnaires and interviews, and also with physiological measurements during exposure to the environment. These methods are presented in Section 4.3. However, the whole concept of presence is somewhat controversial. Criticism towards the concept has been presented stating that there is no real evidence for the phenomenon and it could exist only because questions are being asked about it [82, 83].

3.2 Technology for immersive audiovisual systems

In this Section, the enabling technologies for immersive audiovisual environments are first reviewed. Next, a few recent immersive virtual environments are presented with their benefits and drawbacks, and finally, the audiovisual system deployed in this study is presented in greater detail.

3.2.1 Immersive visual display technologies

Immersive visual display technologies can be divided into three groups defined by the scale of implementation. First group of displays are intended to be used by one individual at a time and comprise Head Mounted Displays (HMD) and desktop stereoscopic displays. Second group of displays are setups for medium-scale collaborative use such as CAVE Automatic Virtual Environments (CAVE). Third group are large-scale displays designed for group immersion, such as IMAX (Image MAXimum) theaters and various simulators. Here the focus is on small- and medium-scale displays.

An HMD is a display mounted on the user's head via a helmet or eye-glasses. A generic HMD system includes the display device, optics to deliver the imagery to the eyes of the user and a head tracking system to conform the imagery in response to the user's movement. HMDs can have the ability to display stereoscopic imagery, where both eyes receive different images, and enhance the user's depth perception. Other design parameters to take into consideration are field of view and resolution of the display. Modern professional equipment achieve field of views in the range $60° - 150°$, but the resolution gets poorer as the FOV is increased because same pixels are mapped to a larger display area. [51]

HMDs have certain benefits over projector- and panel-based solutions. With an HMD, the display is always in front of the user and, consequently, the virtual world will be visible

regardless of where the user is facing. Another benefit is that the user is isolated from the real visual environment. This can be useful in some cases, where the real environment has visual distractions that could break the immersion. Downsides of HMDs are that the helmet or vizor has to be connected via cumbersome wires that can disturb the user. Also, with an HMD the user cannot see his own body or interact easily with real-world objects. These have to be modeled in the virtual world and the user must, for example, wear gloves that provide haptic feedback. In addition, HMDs are limited to be used by only one person at a time. [30]

Different projector-based setups, on the other hand, can immerse multiple people at one time and allow collaboration in an immersive space. Traditionally, front- or back-projection has been used to create large display areas by tiling multiple projectors side by side. Current mainstream projector technology offers High-Definition (HD) resolution of 1920x1080 pixels at an affordable price. This has been found inadequate in CAVE installations where the pixel sizes become as big as 3 mm with the HD resolution and result in blurred imagery. Projector technology is advancing and the first 4K projectors capable of producing 4096x2160 pixels are already on the market. The price point with 4K projectors is still high, but the prices are likely to drop as the technology becomes more popular. [14]

Stereoscopic imagery is achievable also with projectors, but with higher costs than for HMDs. Active and passive stereoscopy require the users to wear special eyewear that is either shutter synchronized with the projector's frame sequential frequency for the active stereo, or polarized differently for each eye with passive stereoscopy. Passive stereoscopy also usually requires two projectors to display the differently polarized images. Third method, autostereoscopy, does not require the users to wear any special 3D glasses. The technology is based on partially blocking the view to the screen by thin vertical lines or cylindrical lenses so that each eye sees different part of the screen due to the differing vantage points. The problem is that the user should be positioned in a small sweet spot and not much movement is tolerated in current systems. [14]

Immersive visual displays can be built also by tiling flat panels rather than projectors. Projection-based systems have troubles with acuity and brightness as the projection screens get larger and the same pixels are spread to a larger area. Also, when using back-projection, the space occupied by the installation is substantial. Superior acuity and brightness, and smaller footprint can be achieved for example with multiple tiled 4-megapixel 2560x1440 30" flat panels inexpensively available today. Tiling panels has some downsides as well. The frame, or bezel, around the panel creates visible seams in the imagery for an array of panels, and current panel technology produces large amounts of heat that has to be ventilated. In addition, synchronizing and controlling dozens of panels requires expensive hardware. However, panel-based solutions have become a popular choice in current virtual reality installations and the advancing panel technology and decreasing cost of the hardware are likely to further increase their popularity. [14]

3.2.2 Spatial sound reproduction

During the last three decades various three-dimensional sound spatialization techniques have been introduced and in this section three of them are presented. The first two,

Ambisonics and Wave Field Synthesis (WFS), aim at reconstructing the original sound field as accurately as possible, while the third method, Directional Audio Coding (DirAC), aims to reproduce perceptually relevant features of the sound field. The DirAC method is explained in greater detail since it is the spatialization technique used in the experimental part of this thesis.

Wave Field Synthesis

Wave Field Synthesis (WFS) [9] uses a large array of loudspeakers to reconstruct the physical properties of a sound field. Virtual sound sources can be positioned freely both in front and behind of the loudspeaker array. The reproduction is based on the Huygens' principle stating that any wave front can be constructed as a superposition of elementary spherical waves. The elementary waves are generated by densely placed loudspeakers controlled by a computer, that determines a suitable delay for each loudspeaker, in order to achieve a particular wave field.

The listening position is not limited to any sweet spot, instead, the listener can move freely in the space with no changes in the localization of virtual sources. The only limitation is that the listener cannot stand on the line between the virtual source and the loudspeaker array. Downsides of WFS are the high cost of building a dense enough loudspeaker array, probably consisting of hundreds of loudspeakers, and the need to minimize the room effect. WFS aims to reconstruct the whole sound field and, in effect, anechoic chamber would be needed for optimal reproduction.

Ambisonics

Ambisonics [20] is a two- or three-dimensional sound reproduction technique based on coincident microphone signals. The basic idea of Ambisonics is to reproduce sound field recorded in B-format by creating virtual directional microphones corresponding to the deployed loudspeaker layout. The B-format microphone setup has signal from an omnidirectional microphone capsule w and three signals from figure-of-eight microphones pointing to x, y and z directions in Cartesian coordinates. Virtual microphone patterns can be created by appropriately summing these signals.

First-order Ambisonics reproduction uses four loudspeakers placed in tetrahedral layout. All four loudspeakers are used to reproduce one particular virtual source, and consequently, the loudspeaker signals are usually highly coherent. This results in undesired effects like comb filtering of the spectrum and blurred localization of sound sources. The effects are even more pronounced for higher order versions of Ambisonics, because the density of loudspeakers with coherent signals increases.

Directional Audio Coding

Directional Audio Coding [56, 91] is a spatial sound processing technique motivated by the human auditory resolution. It strives to reproduce the most salient features of a

sound field regarding human sound perception from coincident microphone signals. The processing is based on a set of assumptions about spatial hearing:

- Interaural cues, and spectral and temporal properties of the sound form the spatial auditory perception
- The interaural cues are affected by the direction of sound propagation, diffuseness of the sound field, and effect of listener on the sound field
- Based on the two previous assumptions, the direction, diffuseness, and spectral and temporal properties of the sound field determine the spatial auditory perception in one position

The processing is divided into analysis and synthesis stages. The aim of the DirAC analysis stage is, based on the assumptions above, to measure the direction and diffuseness of the sound field in frequency bands. Figure 3.2 displays a block diagram of the DirAC analysis stage. The input to the system are four microphone signals from a first-order B-format setup, similarly as in Ambisonics, where w is the signal from an omnidirectional microphone, and x, y and z are the signals from figure-of-eight microphones pointing to corresponding Cartesian directions. The signals are divided into frequency bands approximating the frequency selectivity of human auditory system.

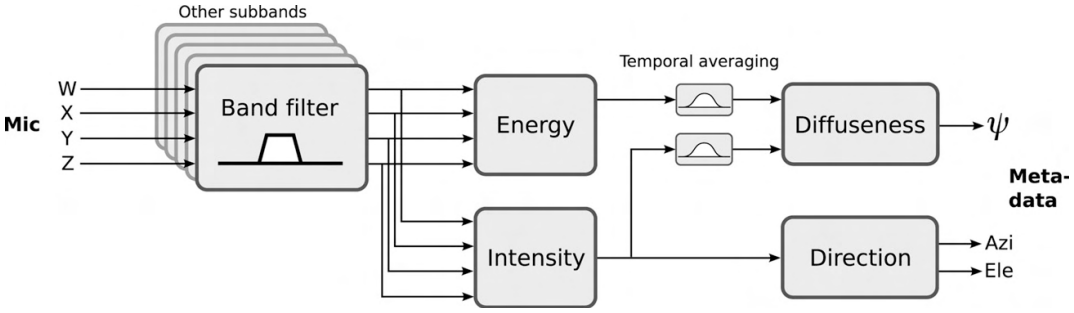

Figure 3.2: DirAC analysis stage. *From [91].*

Next, the sound field is analyzed in each frequency band individually. Energetic analysis is used to form estimates for diffuseness and direction from the energy and intensity of the sound field. Energy and intensity, in turn, are calculated from the sound pressure and the particle velocity, which can be approximated from the B-format microphone signals. Temporal averaging is applied in the analysis of diffuseness. The authors state that averaging is necessary due to the properties of human auditory system. Omitting the temporal averaging causes underestimated diffuseness and, consequently, loss of spaciousness. The analysis stage produces estimates for the diffuseness of the sound field (ψ), and azimuth and elevation angles for the direction of arriving sound in each frequency band.

In the synthesis stage, virtual directional microphone signals are created with alignment towards each loudspeaker in the reproduction setup. The virtual microphone signals are created from a linear combination of the original B-format microphone signals. Figure 3.3 displays a block diagram of the DirAC synthesis stage.

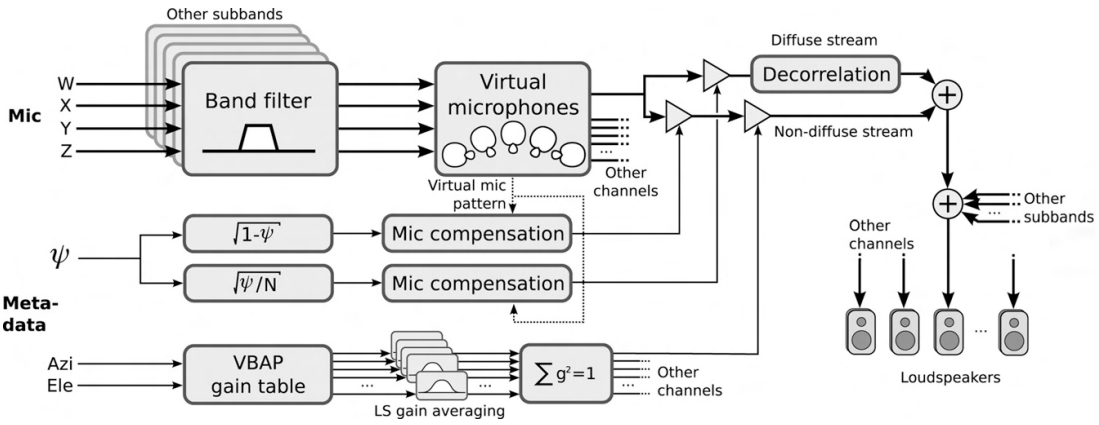

Figure 3.3: DirAC synthesis stage. *From [91].*

The original sound field is divided into diffuse and non-diffuse streams, that are treated differently. The non-diffuse synthesis aims to produce point-like virtual sources for the listener according to the analyzed direction of arrival. The stream consists of the virtual microphone signals multiplied by factor $\sqrt{1-\psi}$, which emphasizes the direct sound from the sources over the reverberant tail. After the multiplication microphone gain compensation is applied to overcome the loss of energy due to the directional virtual microphone patterns when compared to omnidirectional microphone synthesis.

The virtual sources are positioned with a modified vector-base amplitude panning (VBAP) [55] process. Originally, VBAP places the virtual sources between loudspeakers by adjusting the amplitudes of monophonic signals, but in this case the input signal is already a multichannel signal. Consequently, here VBAP is used to gate the virtual microphone signals, so that only a limited number of loudspeakers are used to reproduce a particular point-like virtual source. Temporal averaging is applied also in the non-diffuse stream by slowing down the loudspeaker gains. Fast changes in the analyzed direction cause fast changes in the loudspeaker gains, and concurrently produce a "bubbling" sound artifact.

The diffuse stream, in turn, contains mostly the reverberant part of the sound. The stream consists of the virtual microphone signals multiplied by $\sqrt{\psi/N}$, where N is the number of loudspeakers in the reproduction setup, to attenuate the level of non-diffuse sound. Decorrelation is necessary for the diffuse stream because the virtual microphone signals still have some interchannel coherence. The decorrelated diffuse part of the sound is distributed to all loudspeakers in the reproduction setup.

Other versions of DirAC processing are also presented. Pihlajamäki and Pulkki [52] present a low-delay version of DirAC intended for real time applications such as games and virtual worlds. Here, the processing uses different time-frequency resolution for the diffuse and non-diffuse streams. This way, the non-diffuse content is reproduced as soon as possible and the perceived delay in reduced. Latest proposition, by Politis and Pulkki [53], uses A-format microphone signals in the analysis and synthesis stages. A-format microphone setup is a tetrahedral array of four cardioid or subcardioid capsules. The A-format setup is commonly found in commercial microphones and, in effect, the B-format signals are often derived from A-format microphone signals. Avoiding the transformation into B-format enhances the estimation of DirAC parameters and improves the quality and efficiency of

the synthesis stage, according to the authors.

3.2.3 Recent systems

The classic CAVE Automatic Virtual Environment (CAVE) was first implemented in 1991 in University of Illinois at Chicago [12]. Since then, numerous installations have been made around the world improving on certain aspects of the original design as technology has matured. The CAVE is typically a cube-shaped virtual-reality space 3 m x 3 m x 3 m in size. The walls and sometimes also floor and ceiling are projection screens, or made of tiled display panels, with capability to show stereoscopic imagery.

The latest and most advanced of the CAVE-like systems is the Cornea built in King Abdullah University of Science and Technology (KAUST) [14]. The system is depicted in Figure 3.4. In Cornea all six faces of the cube are rear-projected projector screens effectively surrounding the user with visualizations. Each 3 m by 3 m screen is back-projected by four 4K projectors, each capable to produce 4096x2160 pixels and 10000 lumens. The whole system hence has 24 4K projectors. This results in approximately 4000 x 4000 pixel resolution for each screen, because the system uses passive stereoscopy and, consequently, two fully overlapping projectors are needed to produce imagery for half of each screen. The projection system is controlled by a cluster of 24 PCs including 96 graphics processing units in total.

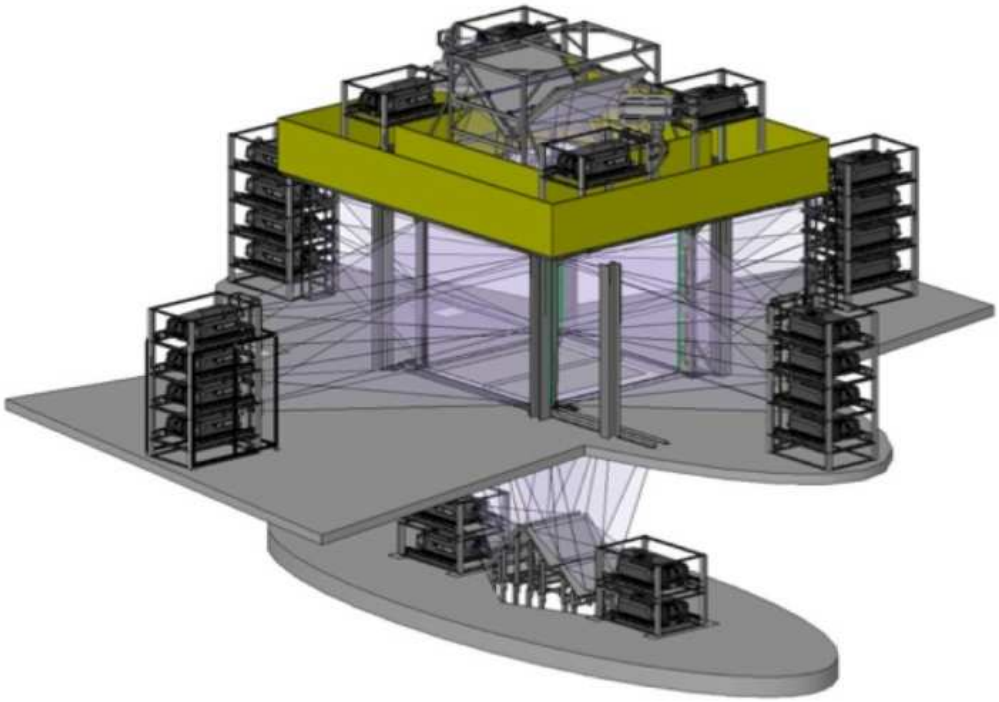

Figure 3.4: 3D model of the Cornea at KAUST. *From [14].*

The Cornea is built into acoustically treated room with low reverberation time. The

projection screens cause trouble considering the sound reproduction with loudspeakers installed outside the cube, because the screen material produces non-linear attenuation, reaching -30 dB above 1kHz, for the side walls. The floor and ceiling, on the other hand, are made of acoustically reflective plexiglass causing flutter echoes. The Cornea is equipped with active acoustics comprising of 16 microphones inside the cube and 20 loudspeakers and four subwoofers outside the projection screens. The active acoustics system enables the Cornea to be electronically coupled to a multi-channel reverberator, and the deficiencies of the acoustic conditions inside the cube can be overcome to some extent. Also, it is possible to simulate different reverberation characteristics for spaces of varying size. [17]

The Cornea is an example of a system where the video reproduction quality is taken to the maximum somewhat at the expense of audio. The Cornea is mainly intended for scientific visualization of complex multivariate datasets and thus the audio reproduction is not considered its major asset. Another project aiming for immersive high-definition video and audio reproduction is the AlloSphere at University of California, Santa Barbara [2]. The AlloSphere is designed and implemented to have symmetrical immersion capabilities in both audio and video modalities. It is not based on the classic CAVE design, but rather resembles IMAX theaters. The AlloSphere space is a spherical screen 10m in diameter. It can accommodate up to 30 people at once on a bridge spanning through the middle of the sphere.

The entire inner surface of the AlloSphere is covered by 14 active stereoscopic projectors each capable of producing 3000 lumens and resolution of 1400x1050. The video setup achieves reproduction of 19,2 Megapixels in total and it is considered sufficient regarding the acuity of human visual system. The sphere surface is made of 23% perforated aluminum sheets, measured to have only a little effect on the sound coming from loudspeakers outside the sphere. The sphere itself is housed in an anechoic chamber. The loudspeaker system has maximum of 512 loudspeakers placed in four circles running above and below the equator. Also, four subwoofers are installed below the sphere.

Spatial sound processing is achieved by three separate systems: vector-base amplitude panning (VBAP) and Ambisonics if virtual sources do not need to be localized inside the sphere, and wave field synthesis (WFS) if more detailed sound fields are necessary. Currently, the AlloSphere project is not yet fully completed, and the system houses only 6 projectors and 140 loudspeakers for prototyping purposes. Once completed, the AlloSphere will be used for scientific visualization for science and engineering, and also for the search of new forms of multimedia art.

3.3 Deployed audiovisual environment

The immersive audiovisual system used in the study at hand was developed and implemented by Gómez Bolaños [21] in 2011 as a part of his Master's Thesis. The environment is located in the Department of Signal Processing and Acoustics at the Aalto University School of Electrical Engineering. The main objectives in designing the environment were adaptability, meaning the system can be easily modified to different research purposes, and high fidelity reproduction in both auditory and visual modalities. Special attention

was paid to the sound reproduction system. The main features of the system are presented here, for further detail see [21].

3.3.1 Video setup

Three dimensional model of the setup is depicted in Figure 3.5 and a real life image of the environment is shown in Figure 3.6. The video setup consists of three HD video projectors with short throw lenses, installed to produce field of view of 226° at the viewing position. Luminosity of one projector is 1800 lumens. The images are front-projected to three acoustically transparent screens placed to follow the shape of the base of a pentagon. Physical size of one projected image is 2.5 x 1.88 meters. Aspect ratio of the screens is 4:3, which means that high definition (HD) resolution 1440 x 1080 pixels is used for each screen and, in consequence, the resolution of the full video system is 4320 x 1080 pixels. This results in pixel size of 1.74 x 1.74 mm which makes the pixels visible if viewed from a close distance. The viewing location is 1.78 meters from the centers of the screens, and in effect, the pixels are hardly visible at this distance.

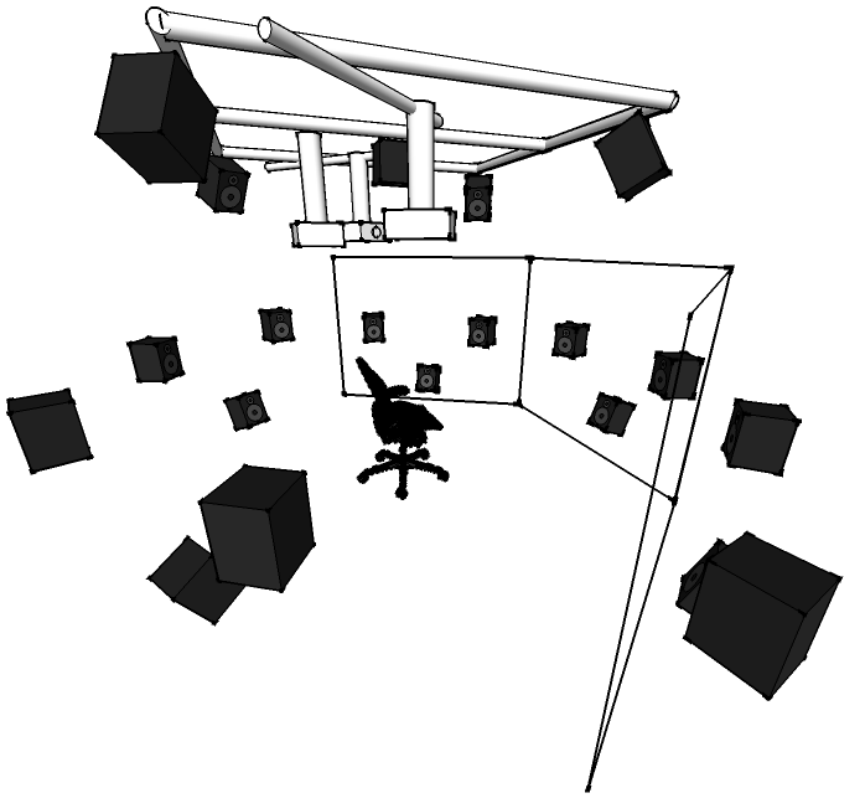

Figure 3.5: 3D representation of the deployed audiovisual environment. *From [21].*

3.3.2 Audio setup

The environment is implemented in a multipurpose room, which poses some difficulties considering the audio reproduction. Dimensions of the room are 8700 x 6150 x 3600

Figure 3.6: The deployed audiovisual environment. *From [21].*

mm with walls of painted concrete. The ceiling is covered with concave diffusers and the walls can be covered with heavy curtains in order to avoid flutter echoes and reduce reverberation time. Bass attenuators built of absorptive material are installed in three corners of the room to reduce the low frequency room modes.

The audio reproduction system consists of 20 loudspeakers in four different elevation levels with respect to the listening position, namely -35°, 0°, 40° and 90° as depicted in Figure 3.7. The sound reproduction is controlled by Directional Audio Coding (DirAC) that is a sound spatialization technique introduced in Section 3.2.2. The loudspeaker positions were chosen in order to provide sufficient number of triangles for the VBAP process inside DirAC. Especially, the horizontal plane was designed to have a good spatial definition.

The loudspeakers used in the setup are Genelec 1029A active loudspeakers with a frequency response of ±2.5 dB between 70 and 18000 Hz in the free field. Some of the loudspeakers are placed behind the video screens and their frequency responses have to be corrected for the slight attenuation caused by the screen canvas. The attenuation, measured 3 dB at 2 kHz and 5 dB at 10 kHz, is corrected with a high-pass shelving filter. The audio setup is calibrated to have an A-weighted equivalent level of 76 dB at the listening position. A-weighted equivalent noise level produced by the video projectors was measured to be 27 dB at the listening position. This was not considered too high because the reproduced sound will usually mask the noise.

In addition, headphone reproduction is possible in the environment. The space is equipped with infrared cameras to detect the headphone position and enable head-tracking. This facilitates the use of HRTFs for 3D audio reproduction with headphones if the room effect has to be avoided. The environment is controlled by software written in Max 5 with MSP and Jitter components [13]. The software runs on a Mac Pro computer with two graphics

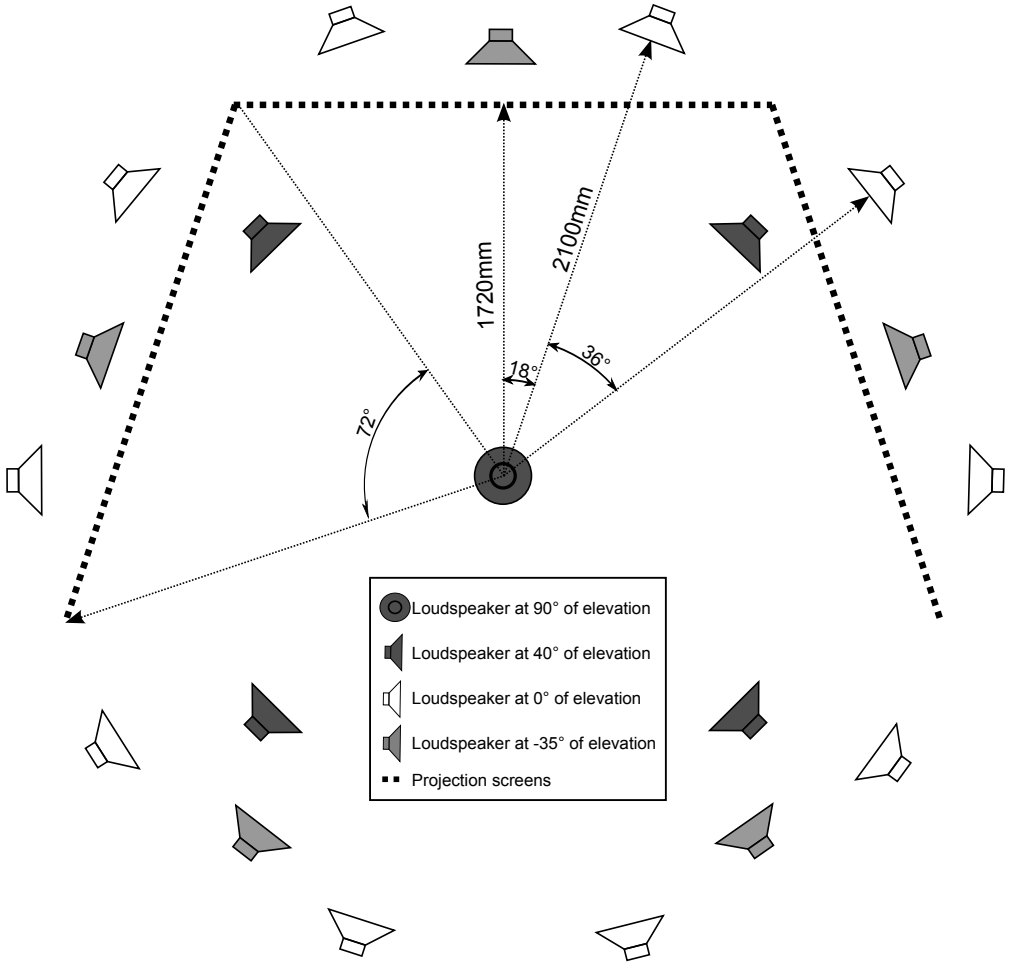

Figure 3.7: Loudspeaker layout of the deployed audiovisual environment. *From [21].*

cards. Max 5 allows controlling the audio and video streams in realtime with multiscreen video and complex loudspeaker setups. Also the user interface for the subjective evaluation task was programmed in Max 5.

3.3.3 Capturing system and content production

The audiovisual material for the environment is recorded from real life scenes with a Ladybug3 camera built by Point Grey Research, Inc. [74] and an A-format SPS200 microphone by Soundfield [85]. The complete system with the camera and microphone mounted on a tripod is shown in Figure 3.8. The Ladybug3 camera system has 5 cameras in the horizontal plane and one camera pointing upwards. This setup enables the recording of spherical videos covering approximately 80 % of full sphere. The camera produces 6 separate image files that are JPEG-compressed and streamed to a disk at 16 frames per second. The spherical video is stitched together from the separate image files in post processing and encoded with huffyuv lossless codec [77]. Final resolution of the full video is 12 Megapixels.

Figure 3.8: Ladybug 3 camera and Soundfield microphone.

Effectively, in this setup, data from approximately three horizontal cameras can be used in the reproduction due to the size of the video screen. This way the visual scene is consistent with the locations of the sound sources, and no distortions occur in the projected video. The original video has to be interpolated to match the native resolution of the video projectors. First, the video is interpolated in VirtualDubMod [38] and next cropped and sliced to three separate video streams in Max 5 using Jitter components.

Audio is recorded in A-format with a portable recording device. The microphone is mounted in front of and below the camera, so that it is not visible in the picture. In post-processing, the A-format signals are analyzed with A-format version of DirAC and synthesized into 20 loudspeaker signals in Matlab.

Chapter 4

Evaluating perceived audiovisual quality

> ❝ *Can telepresence be a true substitute for the real thing? Will we be able to couple our artificial devices naturally and comfortably to work together with the sensory mechanisms of human organisms?* ❞
>
> M. Minsky – *Telepresence*, 1980 [45]

In this Chapter a distinction is made between technically estimated quality and subjectively perceived quality. First, the concept of perceived quality is defined, and next, various objective and subjective metrics for perceived quality are presented through examples. Lastly in this Chapter, the problem of audiovisual content classification is touched, since content is the starting point in understanding perceived quality.

4.1 Perceived multimodal quality

Multimodal perception strives to integrate information from multiple sensorial channels into a unified experience, that contains more information than just the sum of the separate unimodal percepts, as discussed in Section 2.4. Consequently, also multimodal quality perception relies on integration of a set of quality perceptions constructed from sensorial input and higher cognitive processing. Quality perception is a dynamic process where, for example, content and task define the most important aspects affecting the perceived overall quality [34].

Designing an automatic system to estimate the perceived quality of a given artificial audiovisual stimulus would require knowing all the technical features that are relevant to a human observer in a given context and content. This is, however, impossible with current knowledge of how humans perceive complex multimodal stimuli [49]. Objective quality metrics concentrate often only on measuring the coding distortions impairing the transmitted video or audio and traditional Quality of Service (QoS) metrics, such as signal-to-noise

ratio (SNR) or total harmonic distortion (THD), do not relate reliably to the perceived quality.

Therefore, a change of paradigm is suggested from QoS towards Quality of Experience (QoE), when evaluating systems of greater sensorial complexity [70]. QoE is defined by the International Telecommunication Union (ITU) as: *"The overall acceptability of an application or service, as perceived subjectively by the end-user"* [64]. This means it comprises not only technical aspects but also human factors such as expectations, experience and cultural influences. This definition implies that measuring QoE requires subjective interaction tests with humans, whereas QoS can be evaluated objectively by a system developer not involved in the interaction process [22, 46]. The new challenge to researchers and engineers is to understand multimodal perception and the mechanism of quality impression much deeper than nowadays.

A theoretical model for multimodal quality perception has been suggested by Reiter [70]. His three level salience model tries to conceptualize the process of audiovisual quality impression formation in the human user. The model's general structure is shown in Figure 4.1.

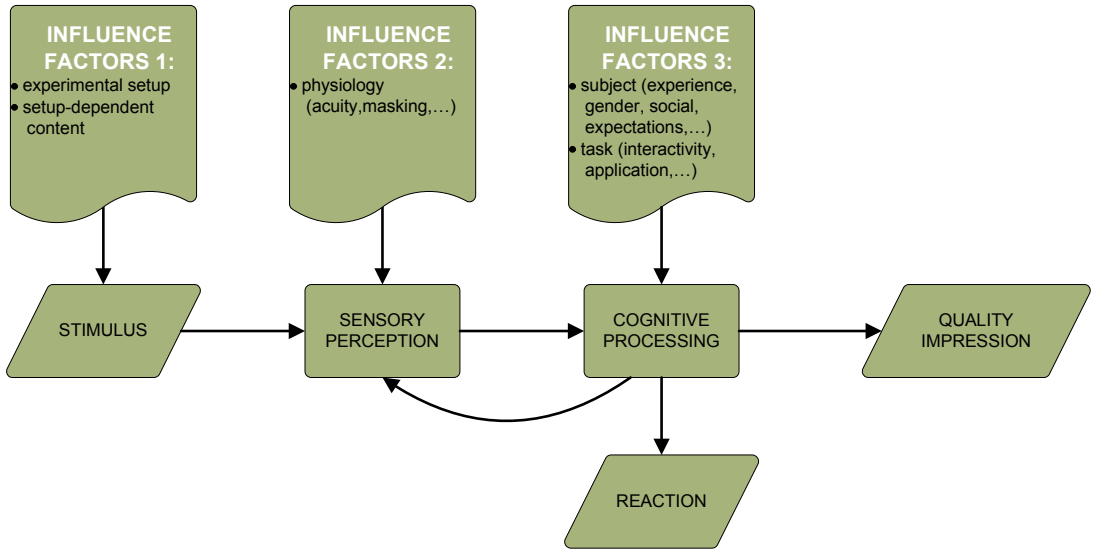

Figure 4.1: Reiter's general salience model for audiovisual application systems. *From [70].*

The model is divided into three levels according to different perceptual and cognitive functions and each level is related with a different set of influences. The first level describes the basis for human perception, namely, the stimuli entering our sensory system. In audiovisual content the stimuli can be influenced by factors related to their generation, *e.g.* audio-visual reproduction method.

Next two levels describe the core of human perception: sensory perception and cognitive processing. Second level is the sensory perception part where the physiology of the user, *e.g.* acuity of hearing and vision, is taken into account as a possible influence. Sensory perception feeds data to the third level, cognitive processing, which is responsible for producing a response by the user. The response can be an immediate action to an event

or it can be feedback to the sensory perception level in order to redistribute attention or shift focus to gain more insight of the stimuli. This feedback loop is similar to the construction of the Neisser's Perceptual Cycle described in Chapter 2.

Influence factors of the third level are the most difficult to quantify, for they are related to the individual's own interpretation of the stimuli. Possible influences rise from expectations, experience and task, for example. Eventually, the cognitive processing will produce an overall quality impression that is, according to the model, a function of all influence factors from each of the three levels.

Takatalo describes a more detailed theoretical model for the formation of subjective experiences in his Dissertation [89]. The model is called the *experiential cycle* and it is depicted in Figure 4.2. Takatalo states that the model can be applied to studying any human-environment interaction, but his dissertation is concentrated on studying experiences in entertainment virtual environments.

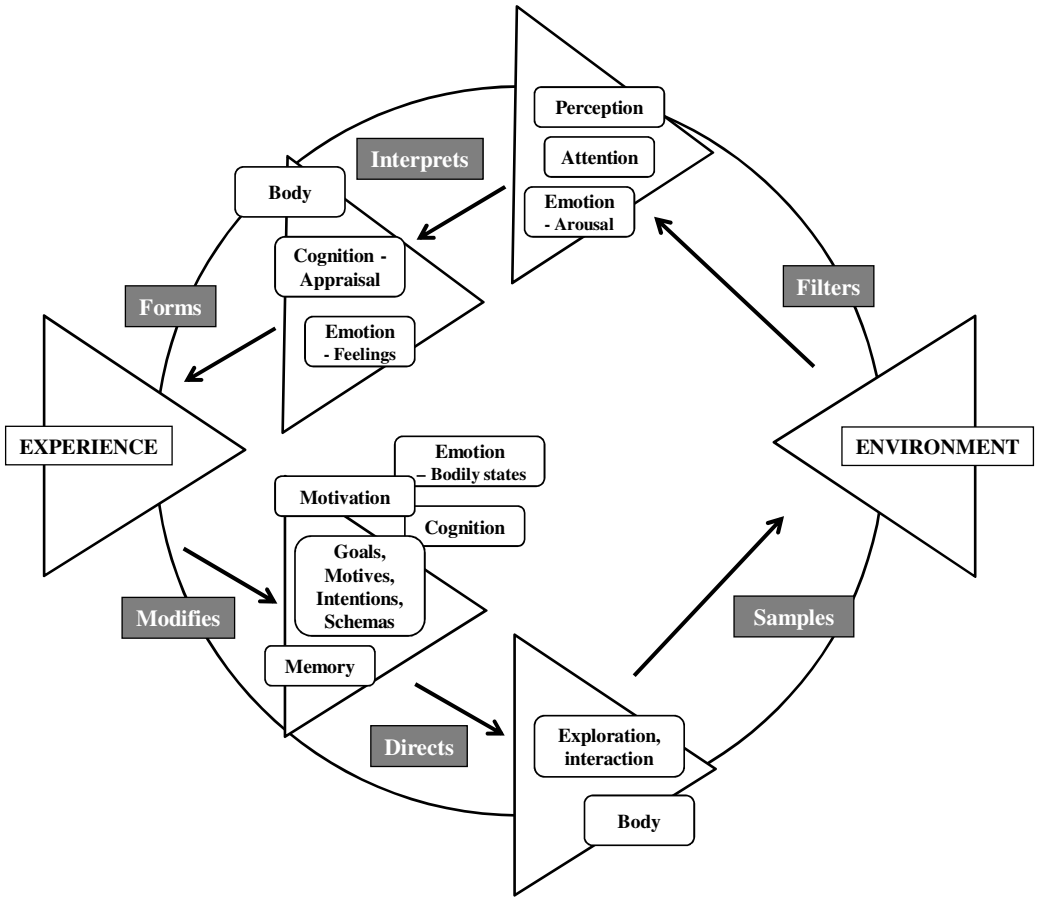

Figure 4.2: Takatalo's experiential cycle. *From [89].*

Takatalo's experiential cycle augments the Neisser's perceptual cycle presented in Section 2.3.1. Takatalo states that the perceptual cycle is good at describing the human-environment interaction in a generic and simple manner. Takatalo's experiential cycle, on the other hand, strives to add psychological multidimensionality to Neisser's basic model

that consisted only of thinking, planning and understanding. Takatalo argues that his model is well suited for studying conscious experience.

The experiential cycle begins with motivation, triggered by some external or internal event, to satisfy a specific need. Takatalo states, that multiple motives may affect an individual at any time, but they are organized into a hierarchy of goals according to their emotionally evaluated benefit or harm to the individual. The motives and goals ranked highest in the hierarchy direct our exploration of the environment. The higher the motive for a goal the more energy is invested in achieving it.

The degree of activation, or emotional arousal, is an indicator of alertness on a scale from deep sleep to high excitement. High level of arousal is also related to narrow attention span, and together the attention-arousal pair defined by motivation enable us to sample our environment at an optimal level in order to achieve our goals. The narrowed attention span also filters out irrelevant stimuli and helps us to focus on the stimuli that interest us.

The relevant stimuli are passed on into consciousness and they become cognitively interpreted. The interpretation aims to recognize the stimuli and relate them to each other and to schemas stored in our memory. The cognitive interpretation shapes our emotions, but also emotions shape our interpretation of the perceived stimuli as emotions can interrupt ongoing goals and substitute new ones. Finally, according to Takatalo, the cognitive processing produces an experience that occupies the conscious mind. The current experience also determines future experiences by modifying new motives.

4.2 Objective measures for perceived audiovisual quality

The quality metrics for unimodal representations have been studied for years and perception motivated objective methods for audio quality [60] and video quality [63] have already been standardized by ITU. Despite the advances in understanding unimodal quality, bimodal quality is still relatively unexplored issue and no comprehensive objective quality metric exist for the perceived quality in multimedia content [26, 33]. In this section the guidelines and attempts for developing an objective multimedia quality metric are presented.

4.2.1 Multimedia models

ITU has published requirements for an objective multimedia quality model in recommendation ITU-T J.148 [62] to guide the development of such a system. The fairly general model consists of individual quality measures for audio and video and their synchronization assessed from the input audio-visual stream. The quality attributes are combined by a multimedia integration function that accommodates human perceptual and cognitive processes known to be active in the formation of quality judgements. The function has a separate input for task-dependent influences that may affect the perceived quality.

The model has five separate outputs. The primary output is the overall perceived audiovisual quality. In addition, it produces measures for separate audio and video qualities, and also objective quality measure for audio, accounting for the influence of video quality

and vice versa.

However, the model is not ready yet. The recommendation points out different issues that need to be addressed in order to form the integration function and assess the perceived quality. Cross-modal influences such as differing quality levels in different modalities, cross-modal error frequency and error synchrony and also cross-modal masking effects need to be considered and additional data from subjective quality tests is needed. Also more information is needed to understand how other cognitive factors such as task, prior experience and knowledge affect the quality perception.

Given the well-functioning perceptual models for unimodal quality, a few studies have set out to form a linear combination of the separate objective unimodal quality attributes in an attempt to form an equation for audiovisual quality. A basic multimedia model by Hands [26] suggests a general formula for the overall perceived audiovisual quality. The predictive model is defined in Equation 4.1.

$$AVQ = a_0 + a_1 \cdot AQ + a_2 \cdot VQ + a_3 \cdot AQ \cdot VQ \tag{4.1}$$

In the equation AQ and VQ stand for the perceived audio and video quality and AVQ for their combined quality perception. The coefficients a_{1-3} are content dependent weights for the different modalities and a_0 is a scaling term that improves the fit between predicted and perceived quality. The predictive model is content dependent because it was noticed that AQ and VQ have different importances in slow-motion content, showing for example only a talking head, in contrast to high-motion content [7, 26]. For a full review of different proposed coefficients for the quality function, see [101].

The fusion model has received some comments wondering whether it is sufficient just to functionally combine the separate audio and video quality attributes. The main question is in what stage of the information processing chain of the human brain does multimodal integration take place. In addition, if audiovisual stimuli are partially processed in different brain regions than unimodal stimuli, there might be problems transferring the results from unimodal experiments to more realistic multimodal situations [73].

Garcia and Raake have proposed a different model for audiovisual quality in high-definition IPTV services based on impairment factors rather than quality factors [19]. It uses factors related to compression artifacts and packet loss calculated from the incoming bit-stream. The model assumes that certain impairment factors can be considered as additive on an appropriate rating scale and therefore the model is a linear combination of different impairment factors. The authors compared results from the predictive model with results from subjective evaluation of the same stimulus material and found very good correlation between the two.

4.2.2 Online prediction

Menkovski *et al.* [44] propose a novel solution combining subjective user feedback and objective quality metrics. They present a machine learning -based approach to online prediction of Quality of Experience. The QoE prediction model consists of technical parameters for network Quality of Service (QoS) and Mean Opinion Scores (MOS) for QoE

collected from users of a service. The idea is to gather a large number of MOS for QoE with different levels of estimated QoS. The machine learning algorithm uses supervised learning to eventually learn how the QoS parameters affect the perceived MOS. Continuous user feedback is necessary in order to get enough training points for the model to learn the underlying structure of the data. As more and more feedback is gathered, the model gets more accurate. The authors also state, that if changes in the measured environment happen, the model will adapt to the new situation as soon as there is enough feedback for the new conditions. The proposed method is best suited for measuring QoE on systems where user feedback is easily available, such as various multimedia content streaming applications. With this kind of approach the time consuming subjective testing can be avoided.

4.3 Subjective measures for audiovisual quality

Due to the lack of comprehensive objective quality measure for audiovisual applications, subjective evaluation of audiovisual quality is considered to be the most accurate method to estimate the human quality perception. In this Section, the standardized subjective evaluation procedures are first presented and a few usage cases reviewed. Next, new approaches for user-centered evaluation, that try to overcome the shortcomings of the standards, are presented.

4.3.1 International standards

ITU has recommended guidelines for conducting quantitative subjective quality assessment tests for audio-visual content in ITU-T Recommendation P.911 [66]. The recommendation defines methodology for evaluating subjective audio-visual quality of non-interactive systems with four different research methods and gives specifications for the listening and viewing conditions.

The most widespread and easiest to implement method in the recommendation is Absolute Category Rating (ACR) where short (<10 s) test stimuli are presented one by one and evaluated independently afterwards. Several studies concerning audiovisual quality have reported to use this method and it is best suited for evaluating systems with wide quality variations. In addition, this method is very similar to the natural way of using multimedia applications. A variation of the ACR is the Single Stimulus Continuous Quality Evaluation (SSCQE) method where the subjective quality is evaluated continuously during presentation of the stimulus. As opposed to ACR, the SSCQE method takes into account temporal variations of perceived quality and the results are presented by plotting the time during which the subjective score was higher than some threshold value.

Other methods in the recommendation have more discriminative power as they are pairwise comparison tests. They are good at finding small differences between the stimuli, because the test material is always evaluated in relation to other materials. The Pair Comparison Test (PC) presents the test sequences in pairs and the pairwise preference in the context of the test scenario is voted after each pair. In Degradation Category Rating (DCR) the test sequences are again presented in pairs, but the first stimulus is always the source

reference. The second stimulus is processed by, for example, the system under test and the subjects are asked to rate its impairment in relation to the reference.

Another recommendation by ITU related to subjective audio-visual quality assessment is ITU-T P.920, which defines interactive evaluation methods for audio-visual communication systems [67]. This recommendation is not, however, discussed here in more depth, because the scope of this thesis is limited to non-interactive assessment.

ITU's methodological recommendations are wide-spread and carefully implemented in multiple audiovisual quality evaluation problems. Using similar methodology makes results obtained in different laboratories comparable and experiments easy to reproduce. However, quantitative assessment of multimodal quality has been criticized for disregarding the understanding of participants' own interpretations and evaluation criteria for quality, and hiding the multivariate nature of multimodal quality under one Mean Opinion Score [33, 49, 88]. Therefore, a set of mixed methods research combining both qualitative and quantitative data have been developed in the recent years.

4.3.2 Mixed methods research

Mixed methods research is stated to complement the traditional quantitative and qualitative research. It is defined by Johnson *et al.* [31] as: *"The class of research where the researcher mixes or combines quantitative and qualitative research techniques, methods, approaches, concepts or language into a single study"*. It tries to combine the best features of both worlds by using words to add meaning to numbers and, on the other hand, numbers to add precision to words.

Considering Quality of Experience studies, a mixed method approach called Interpretation-based Quality (IBQ) has been proposed by Radun *et al.* [57] in the context of measuring perceived image quality of digital cameras. IBQ strives to capture the subjective differences perceived by the users with minimum guidance and intervention by the researchers. In IBQ, the participants are requested to provide a Mean Opinion Score (MOS) for, for example, perceived quality of a digital image. In addition, they are asked to freely describe a few characteristics of image quality they consider important. The assumption is that MOS tells only about a change in perceived quality, but the free descriptions should explain the reasons behind the changes.

Radun *et al.* found that naïve observers were capable of consistently estimating overall image quality without training. Also, the free descriptions were found to be able to separate different imaging devices used in the test through correspondence analysis. The authors conclude, that the IBQ is a fast method to study the most relevant aspects of perceived quality of a technology or a service in a similar setting the users would typically use the services. Later, the method has been used to study perceived quality of printed images [49], video quality [92], and quality of stereoscopic imagery [23]. In [92] the basic IBQ method was augmented to contain two session, where the first session was used to collect a large vocabulary of free descriptions for video quality. This vocabulary was used to form contextually valid assessment scales from the most frequent free descriptions. In the second session, video quality was estimated with the previously formed scales and a MOS scale.

Experienced Quality Factors procedure by Jumisko-Pyykkö *et al.* [33] is very similar to IBQ. The difference is that the qualitative descriptions are gathered in a semi-structured interview after the quantitative quality evaluation.

Open Profiling of Quality (OPQ) has been proposed by Strohmeier *et al.* [88]. OPQ uses quantitative psychoperceptual evaluation of excellence and qualitative sensory profiling in order to create two individual data sets describing the experienced quality. The psychoperceptual evaluation is based on standardized quantitative methodological recommendations (*e.g.* by ITU [66]). The sensory evaluation procedure is done later in a separate session where participants create and refine their own individual quality attributes related to each presented stimulus. The set of attributes is used to form relevant scales for an evaluation task where the stimuli are once again rated with the newly formed scales. These data sets are combined in the final stage of processing called external preference mapping, where a link is created between the quantitative and qualitative results. In other words, the method tries to understand the underlying cognitive processes of quantitative excellence by collecting a set of related perceptual quality attributes. The setting is similar to IBQ, but the attribute elicitation process is much lengthier and the whole quality measurement study can comprise up to three separate sessions.

The authors used the method in assessing the experienced audiovisual quality of a mobile 3D television in three separate experiments, where spatial sound parameters, visual representation mode (2D/3D) and video coding methods were varied. They found that the results complemented each other so that quantitative quality preferences were often explained by qualitative descriptions. For example, quantitative excellence was related to descriptions of depth and error-freeness when visual presentation mode and coding factors were varied. The authors conclude that without the qualitative data the reasons affecting quantitative preference would have been based on assumptions.

As a research method, OPQ has still some challenges to overcome, because the subjects' abilities to accurately describe the perceptual properties of a given stimulus have been reported to vary and, in consequence, the qualitative data can be inaccurate. The authors stress the importance of training with a simple description task prior to the evaluation in order to succeed with the sensory evaluation. Another issue is that the subjects need to participate in multiple sessions. This can cause problems with validity if a lot of participants drop-out between the sessions.

4.3.3 Questionnaires

Questionnaires are used especially in the field of presence research to retrospectively assess whether the participant experienced non-mediation, or "being there", during exposure to a virtual environment [29]. Dozens of questionnaires measuring different dimensions of presence have been proposed since the early 1990s. The large variety of questionnaires is mainly due to the lack of universal conceptualization of presence and, in consequence, the lack of standardized measurement criteria. Also, different applications may require different measurement tools [28].

Witmer and Singer [96] define involvement and immersion as conditions for presence, and created two questionnaires to evaluate factors that could have an influence on these two

preconditions. The first, Immersive Tendencies Questionnaire (ITQ) is a pre-test questionnaire estimating the subject's individual tendency to experience presence. It measures the participants tendencies in three subscales, namely focus, involvement and games, with 29 questions about common activities. It can be used in screening the participants prior to the experiment.

ITQ has been reported to have a good correlation with the second, main questionnaire called the Presence Questionnaire (PQ). Good correlation means that subjects with high ITQ score usually also score high in the PQ scale. PQ is mostly concentrating on the immersive and involvement attributes of the environment, estimating for example the consistency of multimodal information and scene realism originally with 32 questions. Slater [81] has criticized the Presence Questionnaire by stating that it confounds the actual reasons affecting the multidimensional presence experience. Slater argues that the PQ is only a measure for responses to various technical aspects of the system, rather than being a measure for the presence phenomenon itself.

Slater-Usoh-Steed Questionnaire (SUS) [84] has a rather different viewpoint on presence than the PQ. SUS is mainly interested in estimating how the participants remember the exposure to the virtual environment. The authors identify three presence indicators, namely the sense of being there, extent of the VE being more real than reality, and if the VE was thought of as a place visited. In the first version of the questionnaire these indicators were evaluated with three questions. Later the questionnaire has been augmented to include six questions with seven point rating scale.

The use of questionnaires to assess presence is wide-spread in the presence research field, but also criticized for being incapable to fully reflect the mental process the test participants undergo in the virtual environment. Slater [82] demonstrated the effect of a questionnaire attributing meaning to a made up concept when he tried to measure the perceived "colorfulness" of yesterday. Colorfulness was associated with good task completion during the day by some participants and for yesterday being pleasant rather than frustrating day by some. With this experiment he tried to show how an ambiguous concept can have different interpretations and that a questionnaire can elicit meaning to attributes that the participants didn't even know to exist prior to the experiment.

4.3.4 Physiological measurements

A few studies have reported to use physiological measurements to evaluate the level of immersion and feeling of presence the subject experiences in a virtual environment. The hypothesis with such studies normally is that the more real a virtual environment appears, the more similar to reality physiological responses it will evoke. In addition, using physiological measurements doesn't require the subject to understand the term 'presence'. Different interpretations of the concept can result in unstable results across subjects and environments [29]. Physiological measurements can be used as corroborative measures by correlating the measured physiological data with subjective evaluation. In this way, the results obtained from the subjective tests can be made objective.

Sanders and Scorgie used physiological data to study the effect of sound delivery methods on the user's sense of presence [79]. They measured heart rate, skin temperature and

electrodermal activity of the skin in order to estimate the emotional reactions to different stimuli. In addition to objective physiological measures they used subjective presence questionnaires and checked if correlation exists between the two methods.

The subjects played a video game simulating a combat situation while the physiological variables were measured. The questionnaire was filled after a 10 minutes gaming session. The questionnaire was a combination of Witmer and Singer's presence questionnaire and Slater's questionnaire. These questionnaires are presented in Section 4.3.3. The experiment was *between-subjects*[1] design where the sound reproduction method was varied between no sound, headphones, headphones with bass, and 5.1 speakers for different participants. This design was selected in order to minimize repeated exposure to the environment that could affect the sense of presence.

Questionnaire results indicate that there is no significant effect between the tested sound reproduction methods, but the absence of sound was observed to decrease the presence experience. This implies that virtual environments could be built with using headphones for sound delivery instead of large and expensive loudspeaker setups. Comparing the questionnaire data with the physiological measurements showed significant correlation for the electrodermal activity and heart rate with the presence questionnaire. Modulated frequency and larger amplitude in EDA is related to elevated arousal. Arousal, in turn, is considered to be one dimension of emotion so, in effect, adding sound to the virtual environment affected the level of presence by inducing arousal in the participants.

The temperature measurements did not correlate well with the questionnaire results but, in contrast, temperature was observed to drop when speakers were used for sound reproduction instead of the headphones. Also, adding subwoofer to the headphone setup resulted in a temperature drop. Skin temperature drop is considered to be related to fear emotion, as the body prepares to "fight or flight" by pumping blood from extremities to vital organs. The authors hypothesized the subwoofer added to the sensation of fear along with the better spatialization of sounds achieved by the 5.1 loudspeaker setup, although these effects weren't observable in the presence questionnaire results.

The authors addressed the problem that the questionnaires didn't directly ask emotion related questions while the physiological measurements indicated emotional changes. Development of emotion related questionnaires is encouraged. Another problem was selecting the audio-visual material. The video game the authors used in the study didn't contain much point-like sound sources but rather explosions that surrounded the player. In effect, the scenario didn't require the subject to be able to localize sounds in order to complete the mission. The authors point out that a more fitting scenario for evaluating sound reproduction methods would include the need for sound localization.

Another study by Meehan *et al.* [43] was reported to combine physiological measurements and a subjective questionnaire. They measured the same variables as Sanders and Scorgie, namely skin conductance, skin temperature and heart rate, but the difference was in the experimental setup. Meehan *et al.* used a head-mounted display (HMD) and a stressful virtual environment where the user had to perform tasks while standing on the ledge of a 6 m fall. The goal was to evoke as strong as possible physiological reactions in the user and see if the measured reactions correlate with the questionnaire results.

[1] Each subject is tested in one condition only *(vs. within-subjects)*

The results indicate that heart rate is the best physiological measure to differentiate between various levels of presence. It was noted to distinguish between different frame rates and whether haptic response was used or not. Skin conductance was the second best indicator for presence, but it didn't differentiate between different frame rates. The authors also tested if the physiological responses will diminish over multiple exposures to the environment and some attenuation was observed, although the responses never completely disappeared.

A serious problem with physiological measures is the need for scenarios that evoke a strong physiological response. Ordinary and dull situations like being in a virtual room, don't necessarily evoke measurable changes or the expected physiological response is unknown. Desired responses would be [82].

A different approach was presented by Kunka *et al.* [37] who combined gaze-tracking as a supporting tool when they used subjective testing to evaluate the quality of experience. Their method is based on the assumption that the heat map generated by the gaze-tracker should be focused around the image area related to the sound producing object when audio-visual correlation is high. Otherwise the subject's gaze would wander around the image and the sound would not be correlated with the image.

Their experiment consisted of subjective and objective parts. First they conducted subjective assessment of audio-visual interaction without using the gaze-tracker and then an objective assessment with different participants and the gaze-tracking system. The results showed good correlation between the two methods. Those contents rated highly in the subjective evaluation were also found to have a good interaction between seeing and hearing. Good interaction in this study was defined as a high ratio between focus points in the heat map versus the total length of the gaze plot, meaning that the subject's gaze was most of the time concentrated on a small area of the picture.

4.4 Audiovisual content classification

Reliable classification of audiovisual content is regarded essential in developing an objective multimedia quality metric. Perceived audiovisual quality is highly content dependent, as different types of content draw the user's attention towards different modalities with differing quality attributes [26]. Classifying audiovisual content is in its infancy and the first task would be to derive perceptually meaningful factors according which the classification could be performed [71].

Classifying audiovisual content relies heavily on classification of audio and video content separately and often considering only technical aspects of the content. Reiter [72] has proposed a set of dimensions according which the classification could be performed. His work relies on previously defined dimensions by Woszczyk *et al.* [98] who evaluated the quality of home theater systems. They proposed a set of four dimensions relevant for evaluating the perceived quality of home theater systems, presented in Table 4.1, each defined by another four attributes presented in Table 4.2.

Reiter reduced the dimensionality to three significant dimensions and two attributes in the context of content classification. He found that *Mood* dimension had similar importance

Table 4.1: Dimensions relevant for quality evaluation of home theaters according to Woszczyk *et al.* [98]

Dimension	**Definition**
Action	the sensation of dynamic intensity and power
Mood	the articulation and density of atmosphere
Motion	the illusion of physical flow and movement
Space	the illusion of being in a projected space

Table 4.2: Attributes related to dimensions of home theater quality evaluation according to Woszczyk *et al.* [98]

Attribute	**Definition**
Quality	distinctness, clarity, and detail of the impression
Magnitude	the strength of the impression
Involvement	the emotional effect on the viewer
Balance	relative contribution of auditory and visual sensations

in all different contents and consequently it could be suppressed from the analysis. The other three dimensions were found to have content depended importances. Considering the attributes, Reiter found that *Quality*, *Magnitude* and *Involvement* were highly correlated with each other and their influence can be approximated with just one question. The *Balance* attribute was significantly dependent on the content.

A different viewpoint is presented by Säämänen *et al.* [78] who set out to find what kind of content users produce with digital video cameras in order to be able to produce meaningful test sequences for benchmarking perceived quality of imaging devices. They came up with a 3-dimensional Videospace where the videos are represented according to subject-camera distance, scene lighting and object motion. First version of Videospace classification relied on expert ratings, but the authors suggest computer-based algorithms could perform the classification faster and with higher precision. Downside of Videospace is that it completely neglects the audio track in the videos.

Chapter 5

Experimental work

5.1 Objective

Aim of the study was to examine the impact of cross-modal effects on subjective evaluation of degradation in immersive multimedia content. The spatial width of audio and video reproduction were varied in an immersive audiovisual environment and subjective evaluations for magnitude of the degradation were collected. In addition, free descriptions of the most prominent aspects affecting the given degradation score were gathered. Finally, constrained correspondence analysis was applied in order to clarify the reasons for a particular degradation score in given conditions. The present study combines quality of experience methodology with features of presence research, and also tries to take the perceiving individual into consideration. The procedure used here is based on the Interpretation-based Quality (IBQ) methodology [57] and recommendations by ITU. Related research is presented in Section 2.4.3.

5.2 Research questions

1. How does the spatial width of audio and video reproduction affect the perceived degradation of immersive audiovisual stimuli.
2. Does content have an effect on the relative importance of spatial width of audio and video reproduction.
3. Can individual tendency to experience immersion affect the amount of perceived degradation.
4. What are the reasons behind a given degradation score.

5.3 Method

5.3.1 Participants

Participants were recruited with an open call posted to Aalto University's student forum and to those attending a basic course in sound technology. The participants were given a gift card worth of 15 euros. Overall, 23 people participated in the test. Six of them were female and the average age was 24,3 (SD=2,9). All reported to have normal or corrected to normal hearing and vision and none of them were professionally engaged with audio or video quality evaluation. In addition, five people participated in an informal pilot study to test the functionality of the procedure and suitability of the stimuli. The pilot testers did not participate in the actual test.

5.3.2 Apparatus

The tests were conducted using an immersive audiovisual environment developed and implemented by Gómez Bolaños in 2011 as a part of his Master's Thesis [21]. The environment is located in the Department of Signal Processing and Acoustics at Aalto University School of Electrical Engineering. The technical specifications of the environment are presented in Section 3.3.

The audio reproduction system consists of 20 loudspeakers in four different elevation levels, but in this study only 10 loudspeakers at 0° elevation with respect to the listening position were used. The audio files were recorded with 24 bits at 48 kHz with an A-format microphone. Audio spatialization from A-format microphone signals to 10 loudspeaker signals was achieved by Directional Audio Coding (DirAC) [56]. Sound pressure level of each content was adjusted to approximately match the one in the recording location.

The participants were seated in the center of the environment and the test was run by a program written in Max 5 automatically controlling the playback of the stimuli and collecting the answers. Signal processing for preparing the audio files was done using Matlab and video processing using VirtualDubMod and Max 5 with Jitter components.

5.3.3 Stimuli

Four different video segments were used in the test. The chosen scenes displayed team sports, choir music, conversation and traffic. Table 5.1 shows the classification of the video contents according to ITU's Recommendation P.910 [65] and also according to Videospace classification framework proposed by Säämänen *et al.* [78]. In addition, the sound scene is coarsely described for each content.

The actual stimuli were created by limiting field-of-view and spatial extent of audio reproduction of the original video. Table 5.2 summarizes different settings applied to the video segments in order to create the stimuli and Figure 5.1 visualizes the settings. Video field-of-view was limited using an overlapping matrix of dark-grey pixels. The matrix was applied symmetrically to the left and right edges of the screen.

Table 5.1: Video content classification

	Floorball
Description	Game of floorball with movement across the field.
Video details	High
Video motion	High
Subject distance	12 m
Lighting (1-10)	5
Audio	Point-like moving sources and transients. Non-diffuse stream dominant.
	Choir
Description	20 people choir singing Rahmaninov in a church.
Video details	Moderate
Video motion	Low
Subject distance	6 m
Lighting (1-10)	3
Audio	Wide and still source. Diffuse stream dominant.
	Conversation
Description	Two men discussing over table in Greek.
Video details	Low
Video motion	Low
Subject distance	2 m
Lighting (1-10)	5
Audio	Two still point-like sources. Non-diffuse stream dominant.
	Traffic
Description	Trucks and cars accelerating from traffic lights. Also two bicyclists passing the camera close by.
Video details	Moderate
Video motion	High
Subject distance	20 m
Lighting (1-10)	7
Audio	Point-like moving sources and noise. Diffuse stream dominant.

Spatial extent of audio reproduction was controlled by manipulating the DirAC synthesis stage. The 360° condition was synthesized with the real loudspeaker directions. In 180° condition only 6 frontal loudspeakers were used and the DirAC synthesis was told that those loudspeakers actually situated at ±90° of the listening position would have been situated at ±179°. The remaining four loudspeakers were linearly distributed between these extremes. In *stereo* condition only two loudspeakers situated at ±18° were used. This time, DirAC was told that there were four loudspeakers, two at ±45° and two at

Table 5.2: Settings applied to the original video content

Video	**Audio**
Full (226°, 4320 px)	Full 2D (360°)
Limited to 2/3 (150,7°, 2880 px)	Scaled to frontal 180°
Limited to 1/3 (75,3°, 1440 px)	Scaled to 36° (stereo)
	Mono from back

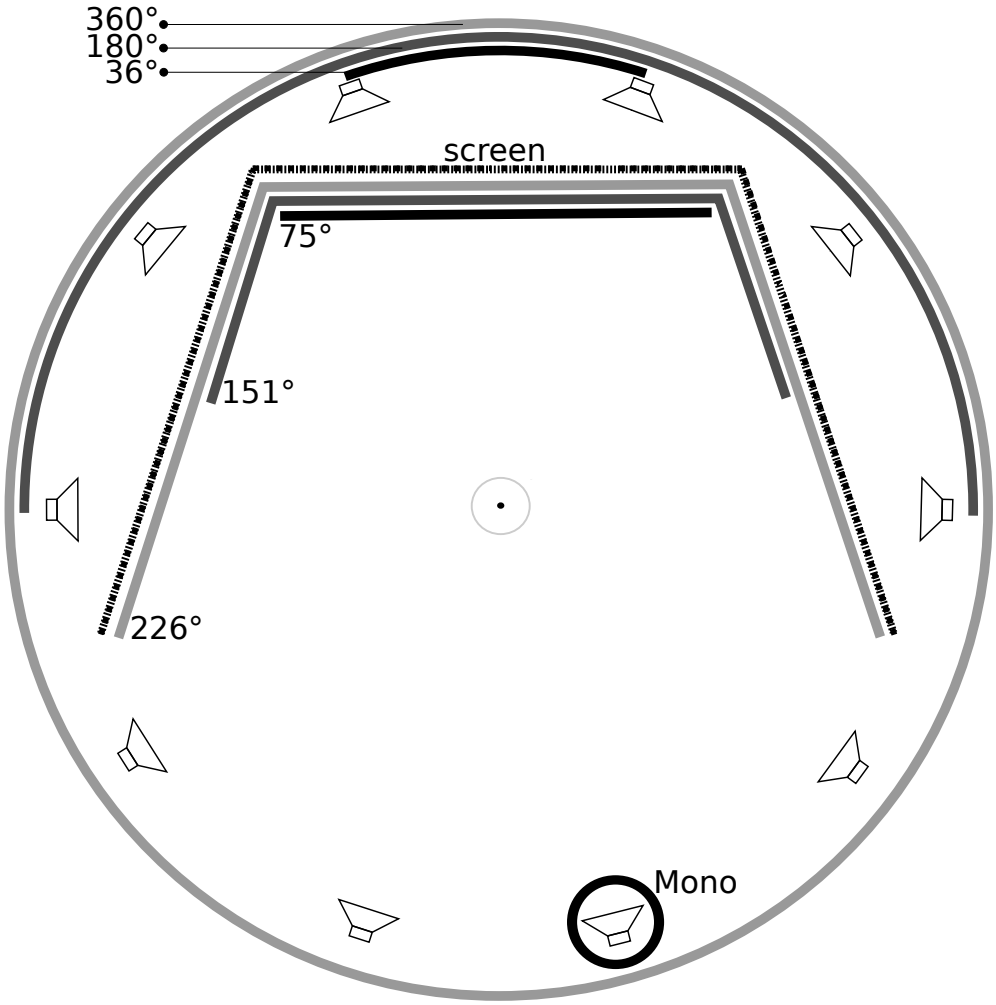

Figure 5.1: Settings for audio and video reproduction.

±135°. The resulting signals were summed in both sides to the loudspeakers at ±18°. The *mono* condition was created by summing all the original 360° condition loudspeaker signals to one loudspeaker located at 162° behind the listening position.

This processing resulted in asymmetric degradation of audio and video. In video some information was entirely lost due to the occluding matrix, while in audio only direction

information was impaired due to spatially narrower reproduction. In addition, the processing introduced audiovisual conflicts between audio and video event locations. The conflicts were more pronounced the narrower the sound reproduction was.

5.3.4 Procedure

The test was divided into two parts. It began with Immersive Tendencies Questionnaire (ITQ) [96] and continued to Degradation Category Rating (DCR) [66] of the stimuli videos. Aim of the ITQ was to get information about the individual tendencies of the participants to be involved or immersed. The questionnaire contains 23 questions and thus produces a 23 dimensional feature vector describing the individual's tendency to experience immersion. In this study the questions were answered on a continuous scale with end-points denoted with suitable minimum and maximum attributes. These feature vectors were used to classify the participants into two groups corresponding to low and high relative tendency for immersion.

Next, in the Degradation Category Rating, the participants were shown pairs of videos. First video of the pair was always the reference, and the second video was the degraded stimulus. The reference stimulus had full 226° field-of-view and 360° audio reproduction, while the degraded stimulus was processed with some combination of the conditions presented in Table 5.2. After each pair the participants were shown an answering sheet where they had to mark how much the second video differed in comparison to the reference in their opinion. The answering sheet is depicted in Figure 5.2.

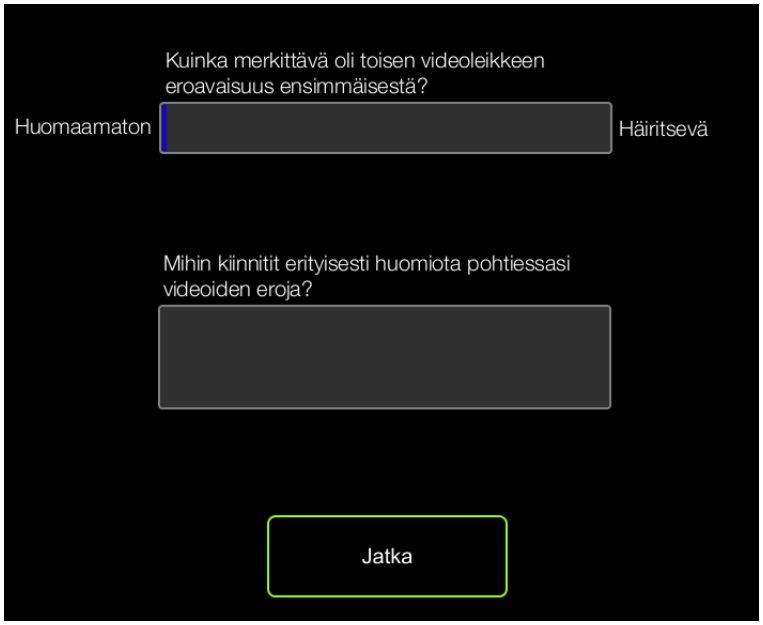

Figure 5.2: Answering sheet in the DCR test.

The degradation evaluation was done on a visual analogue scale with end-points denoted as *Imperceptible (Huomaamaton)-Annoying (Häiritsevä)*. The slider produced degradation scores on a scale 1...100, where 1 corresponds to no perceived degradation and 100 to

annoying degradation. In addition to the slider, the answering sheet contained an open question box where the participants could freely describe a few things they paid attention to when comparing the two videos.

The audio and video conditions produced 12 different stimulus videos for each content. The stimuli were shown in random order within one content. Also the contents were randomized in a way that any one of the four contents was shown first before moving to next one. One reliability check stimulus was added to each content and, consequently, the whole test comprised of $4 \cdot 13 = 52$ pairs of videos.

Training session of four pairs was completed before the actual test began. During training it was made sure the participants had understood the task, but no additional information about the test conditions was given. The participants were instructed to pay attention to both audio and video when making the voting after each pair. Also, they were encouraged to consider what information about the surrounding world was missing from the second stimulus that was present in the reference, and how disturbing that loss of information was.

The DCR test was divided into four sections, 13 stimulus pairs each. A short break was held after each section. The whole test from the ITQ questionnaire to finish took 63 min on average (SD=11,8).

5.4 Results

5.4.1 How does the spatial extent of audio and video reproduction affect the perceived degradation of immersive audiovisual stimuli

Figure 5.3 displays the mean degradation scores with 95% confidence intervals for each combination of audio and video conditions. Scores from the four different contents are pooled together in respective conditions for this analysis. Dunnett's modified Tukey-Kramer pair-wise multiple comparison test was performed to study the differences in means. All the video conditions were found to have significantly different means at $p = 0.05$ in every audio condition, the difference diminishing as the audio reproduction gets narrower.

With full video, the spatial width of audio has clear effect on the perceived degradation. 360° and *mono* conditions differed significantly from the others, while 180° and 36° conditions formed a group with no difference in means. The effect of audio width diminishes as the video width is decreased. With 2/3 video condition only the mono audio is significantly different from the other audio settings and for the 1/3 video condition the audio condition has no effect on perceived overall degradation.

5.4.2 Does content have an effect on the relative importance of spatial width of audio and video reproduction

Next, the data was divided according to the four different contents. Figures 5.4, 5.5 and 5.6 show the mean degradation scores with 95% confidence intervals for each audio condition

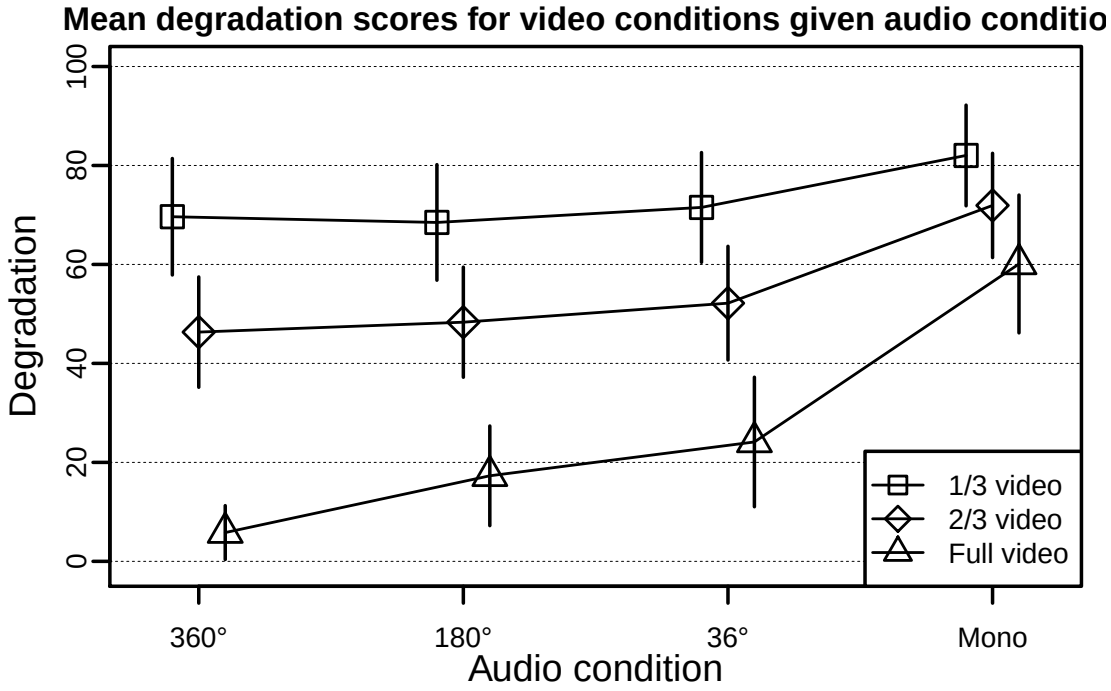

Figure 5.3: Mean score for each audio and video condition. Whiskers denote the 95% confidence interval.

and content in full, 2/3 and 1/3 video condition, respectively. In full and 2/3 video conditions, the only significant content effect at $p = 0.05$ was observed between *conversation* and *traffic* in 36° audio condition. In 1/3 video condition the mean degradation for *floorball* content was significantly larger than the mean degradation for *choir* content in 360° audio condition. Similarly, in 180° audio condition *floorball* and *conversation* contents had significantly higher mean degradations than *choir*.

5.4.3 Can individual tendency to experience immersion affect the amount of perceived degradation

Internal consistency of the Immersive Tendencies Questionnaire data was estimated and the reliablility of test scores was found acceptable (Cronbach's $\alpha = 0.71$). Originally, Witmer and Singer designed the questionnaire to measure immersive tendency in three subscales: involvement, focus and games. Later, Weibel *et al.* [94] performed factor analysis on the questionnaire scores to examine the dimensionality of immersive tendency, and found the original division to three subscales unsupported. They discovered two subscales, emotional involvement and absorption.

In this study, principal component analysis was performed on the questionnaire data and item correlations with the two primary dimensions was observed. First dimension explained 18,6% of the total variance and was significantly correlated with 5 questionnaire items. The items refer to emotional reactions during media usage and gaming similarly

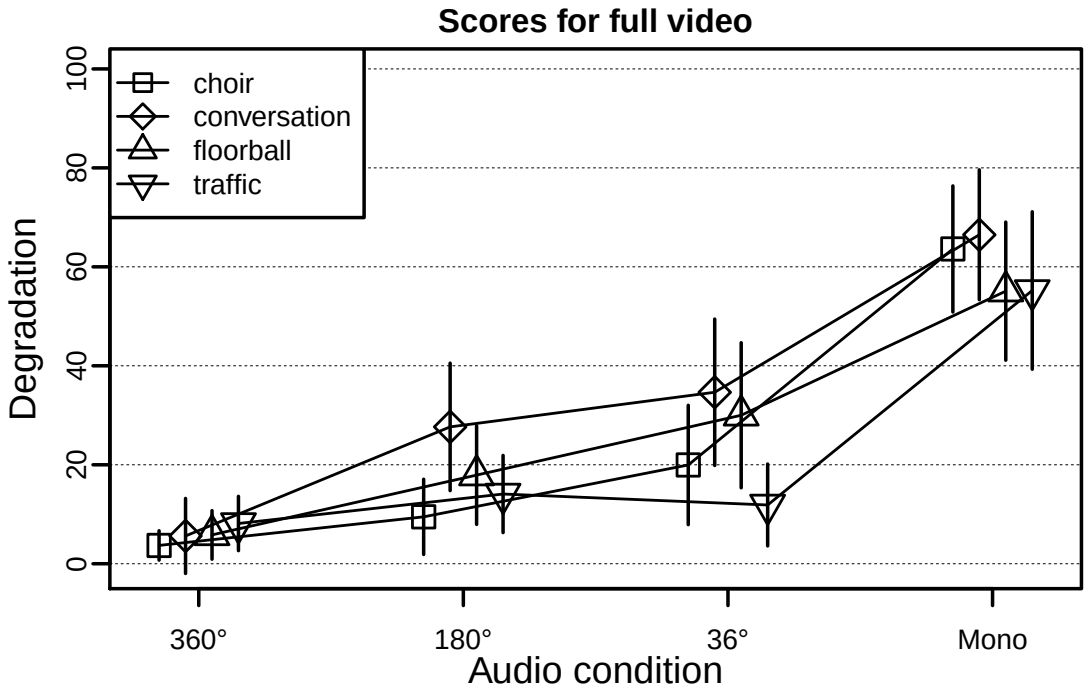

Figure 5.4: Mean score for each audio condition and content with full video. Whiskers denote the 95% confidence interval.

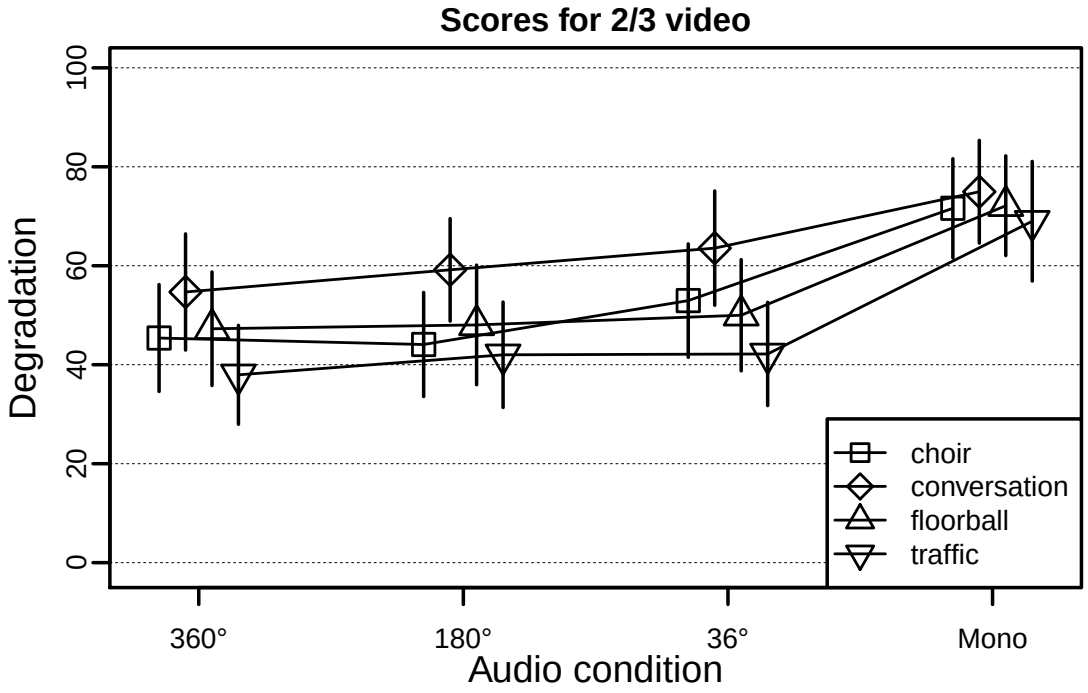

Figure 5.5: Mean score for each audio condition and content with 2/3 video. Whiskers denote the 95% confidence interval.

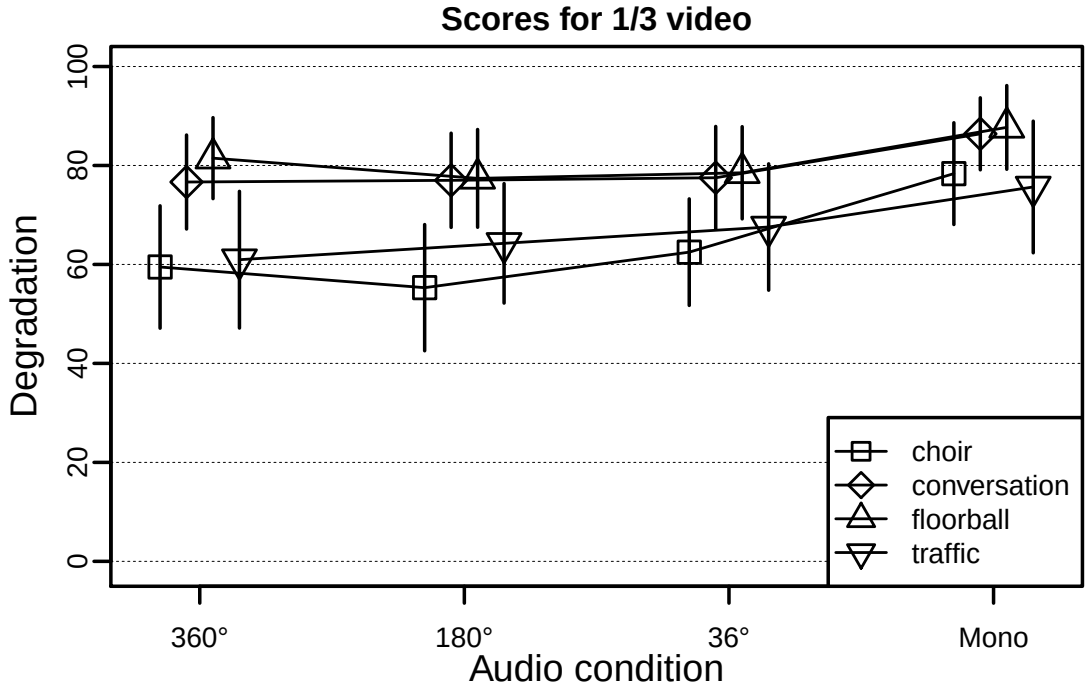

Figure 5.6: Mean score for each audio condition and content with 1/3 video. Whiskers denote the 95% confidence interval.

as in Weibel *et al.*'s study, and consequently, this dimension was dubbed *emotional involvement*. The second dimension explained 13,6% of the variance and was significantly correlated with 4 items related to ability to block distractions and tendency to forget one's surroundings. These items were again in accordance with Weibel *et al.*'s findings and the second dimension was dubbed *absorption*.

Based on the scores in the first two dimensions, the participants were divided into two groups: low tendency for immersion ($N_{low} = 11$) and high tendency for immersion ($N_{high} = 12$). Figure 5.7 displays the mean overall degradation scores separated by groups for low and high tendency to experience immersion. The low tendency for immersion group gives constantly higher degradation scores but the difference is not statistically significant in any of the test cases at $p = 0.05$.

5.4.4 What are the reasons behind a given degradation score

The free description data were analyzed by first going through all the answers and searching for expressions that could be classified, or coded, into more general categories. The expressions were first transformed to their basic form. Next, similar expressions were grouped together under broader categories using Atlas.ti software [3]. The qualitative answers were originally given in Finnish and the categorization was also performed in Finnish. Inter-rater agreement was found acceptable, when comparing the codings made by the author and an independent coder using the same codes (Cohen's $\kappa = 0.66$). Fi-

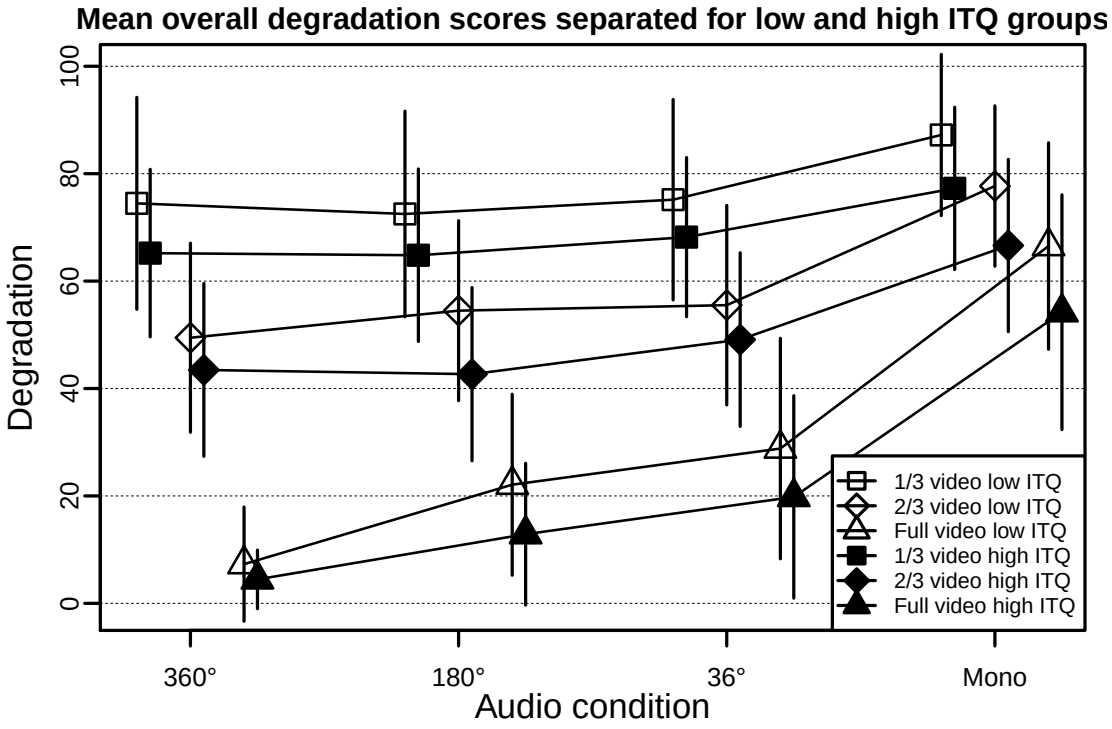

Figure 5.7: Mean score for each audio and video condition, grouped by immersive tendencies. Whiskers denote the 95% confidence interval.

nally, the expressions and codes were translated into English by the author, and subtle nuance differences may have been introduced in this stage. The translation was made keeping in mind the actual meaning of the expressions rather than translating each word as accurately as possible. Ten most frequently used codes with related typical expressions are listed in Table 5.3.

Each of the coded expressions was also related to a specific setting for audio and video reproduction, and content. A contingency table was formed, where the rows corresponded to all the different combinations for audio, video and content and columns were the ten most frequent codes. The table showed how many times each code was mentioned with each of the reproduction conditions.

Constrained (or canonical) Correspondence Analysis (CCA) was used to analyze the structure of the contingency table. The analysis was made with *vegan* -package [50] in R language. CCA is often applied in ecology studies, where environmental variables are used to constrain ordination of different sites and their vegetation. The hypothesis is that vegetation is controlled by environment, and consequently, CCA is used to ordinate the contingency table constructed from sites and their vegetation constrained to be related to a second matrix containing environmental variables, *e.g.* soil type or elevation. Constrained ordination aims to display only the variation explained with constraining variables, while rest of the variation is thought to be caused by some unknown or unmeasured variables in the environment.

Table 5.3: Frequency table for the codes extracted from the qualitative data.

Typical expression	**Code**	**Frequency**
"narrow picture" *"image missing from the sides"*	Video width	380
"sound from the behind" *"sound from one point"*	Audio direction	274
"sound from narrower area"	Audio width	116
"can't see the talkers/game" *"important section of the image missing"*	Essential content missing	108
"sound doesn't localize according to the video" *"miss-synchronized audio and video"*	Audiovisual mismatch	86
"small difference in audio" *"maybe something in audio"*	Something in audio	46
"more noise in audio" *"sound quality was poor"*	Audio quality	42
"grainy video" *"brighter/darker video"*	Video quality	30
"like they are singing in a small room" *"scary, because you don't know anything about the surroundings"*	Poor sense of space	30
"the audio was more quiet" *"audio wasn't as loud"*	Quiet audio	25
	N=	**1137**

In this study, the constraining environmental variables were thought to be the spatial width of audio and video reproduction, and content. Sites were taken to be the different combinations of the three. Vegetation related to a site was, in this study, the ten most frequent codes obtained from the free descriptions. Effectively, the codes are species that live in a specific site described by environmental variables.

The obtained model with three constraints was assessed for significance by using permutation tests. Marginal effects test, when each term is eliminated from the model containing all other terms, indicated all the constraints are significant at $p = 0.05$. Similarly, significances of the obtained CCA axis were tested with permutation test and the five first axis were found to be significant at $p = 0.05$. The constrained model explains 63% of the total inertia, and the three first CCA axis explain 49%, 30% and 10% of the constrained model's inertia.

Figure 5.8 displays the two first axis obtained from the CCA. Sites (reproduction method-/content combinations) are denoted with black circles and free description codes with blue

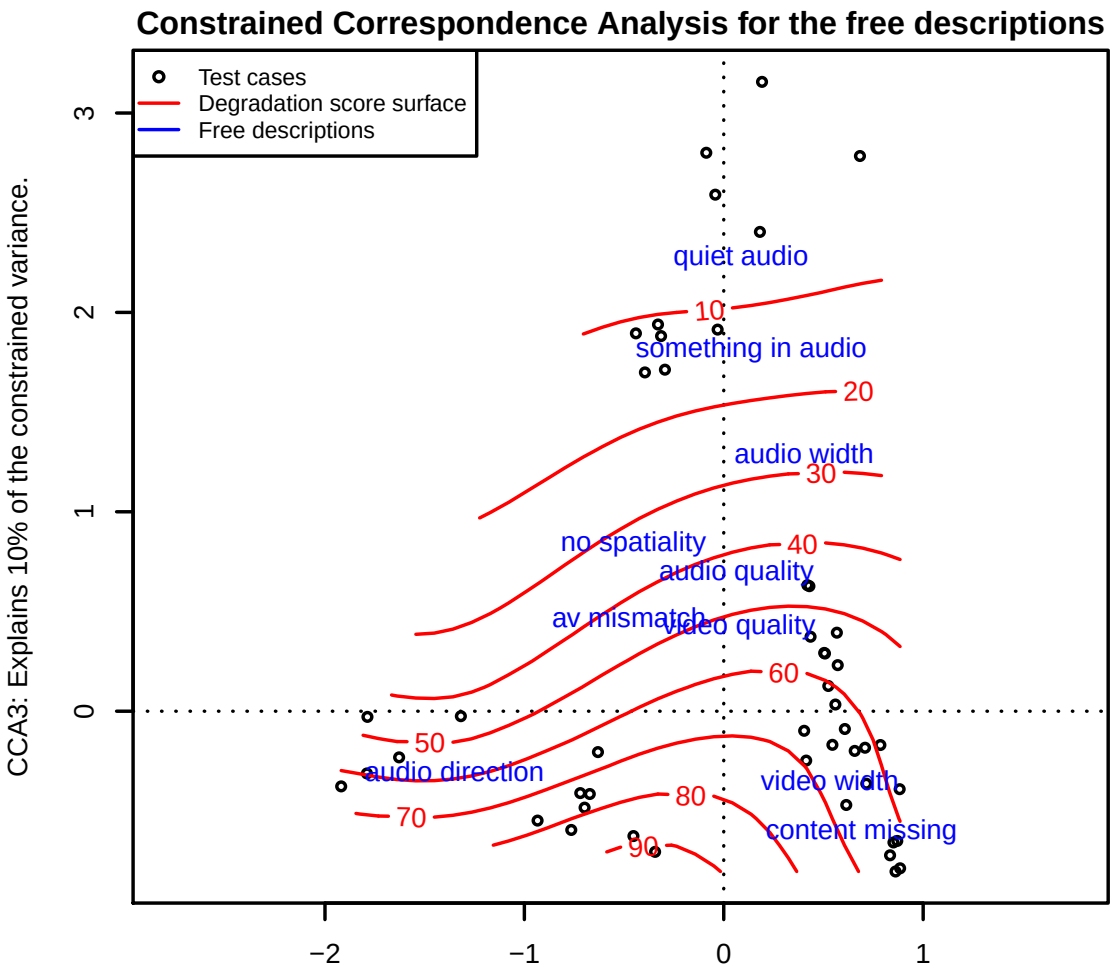

Figure 5.8: Ordination of test conditions and free descriptions constrained by audio and video reproduction type and stimulus content.

text. Both of them are scaled proportional to axis eigenvalues. In addition, the Figure shows fitted surface obtained from the median degradation scores related to each site. The degradation score was not used in the model building, but rather added afterwards to reveal more information of the relationship of perceived degradation to environmental variables.

Degradation scores are negative scores in a way that the larger the score the bigger the perceived degradation. Figure 5.8 shows that worst scores reside close to descriptive codes about audio direction, video width and missing content. Similarly, codes related to something undefined about audio, audio quietness and audio width are ordinated to sites with low degradation score, *i.e.* small perceived degradation.

Figures A-1, A-2 and A-3 show the environmental variable categories in the ordinated space for video width, audio width and content, respectively. The full video condition is clearly separated from the two other video widths, but the category is really widespread. By looking at Figure 5.8 none of the video related descriptions fall under the full

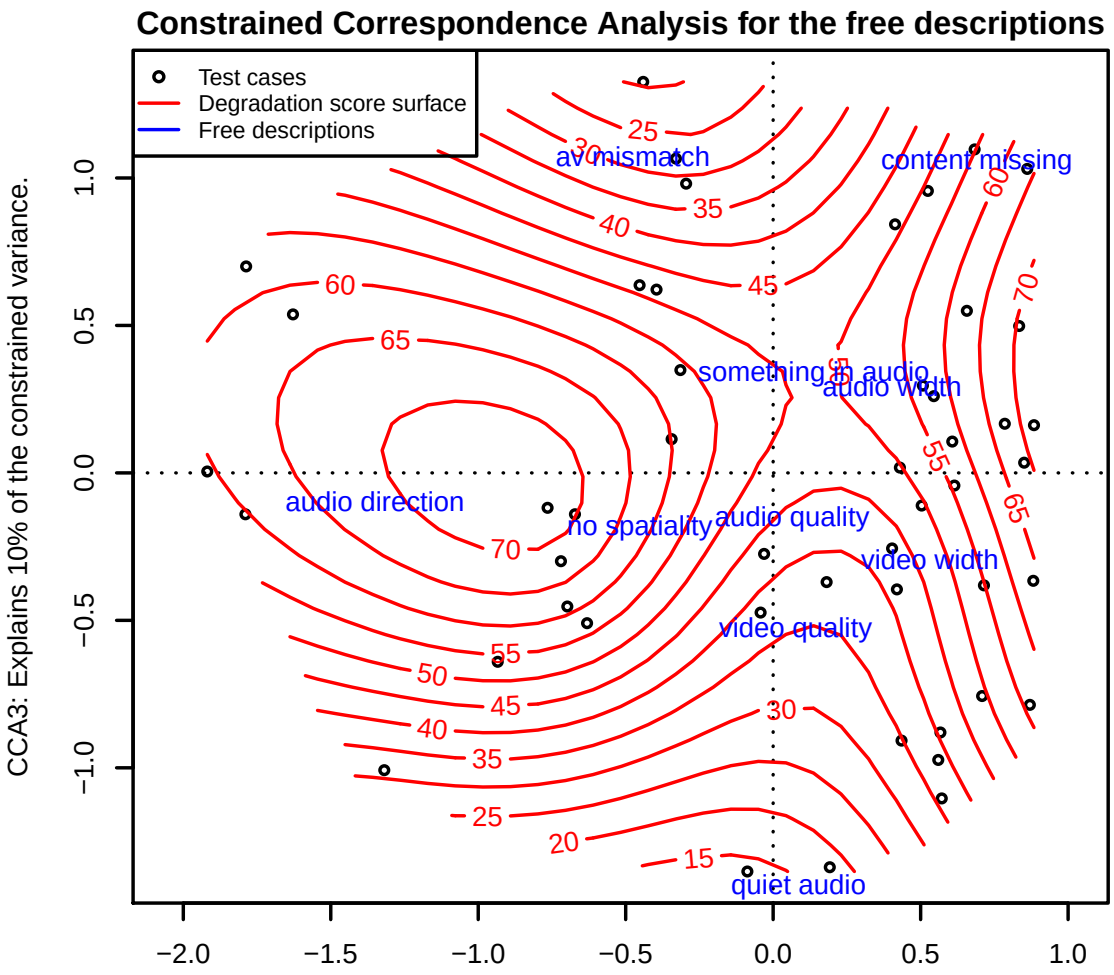

Figure 5.9: Ordination of test conditions and free descriptions constrained by audio and video reproduction type and stimulus content.

video category, whereas the narrower video categories get mentions about video width and missing content, and also receive the highest degradation scores.

Regarding audio categories, the *mono* condition is clearly separated from the three others. Also 360° condition is almost separated from the 180° and 36° categories. Content categories are fully overlapping each other and no distinction can be made between them in these dimensions.

Figure 5.9 displays the CCA ordination in dimensions 1 and 3 and Figures A-4, A-5 and A-6 the environmental variables in the respective dimensions. Here, a distinction between contents can be observed. Conversation and floorball contents are separated from the traffic and choir contents. Looking at Figure 5.9 reveals that conversation and floorball are situated close to descriptions of audiovisual mismatch, audio width and missing content, whereas traffic and choir are close to mentions of audio and video quality and video width.

Chapter 6

Conclusions

In this work, immersive audiovisual displays were presented from multiple points of view, and a subjective user experiment was conducted in order to add new knowledge about multimodal perception based on previous studies. First, the human audiovisual perception was reviewed beginning from physiological structures of the sensory organs and progressing to theories of perception and consciousness. Next, enabling technologies for immersive audiovisual systems were presented along with a few recently built audiovisual environments. Last, objective and subjective methodology for studying the quality experience in an audiovisual application was reviewed and the user experiment conducted using an immersive audiovisual display.

In the conducted experiment, spatial width of both audio and video reproduction were varied and the impact on overall perceived quality degradation observed. Video width was found to be the dominant element in defining the degradation of an audiovisual stimulus. With full video width the effect of spatial width of audio was also significant, but as the video width was reduced, the effect of audio width almost disappeared. These findings are in accordance with previous studies concerning home theater systems by Bech *et al.* [6] and stereophonic audio setups [5].

Content did not have a noticeable effect on the perceived degradation in this study, in contrast to previous studies, where the effect of content on perceived audiovisual quality has been strong, *e.g.* [36]. *Floorball* and *conversation* contents were slightly separated from *choir* and *traffic* contents. Tendency to experience immersion had some repeating impact on the given degradation scores, but the effect was not significant. Larger sample could have produced also statistically significant results.

Correspondence analysis was applied to explain the results for quantitative analysis of subjective degradation evaluation. Degradation scores alone do not tell the actual reasons for the perceived degradation and qualitative analysis is needed to aid the interpretation. In the present study, the constrained model explained 63% of the total inertia of the data. The unexplained variation could be further reduced by incorporating new constraining variables, for example, the free descriptions could have been divided according to low and high tendency for immersion participants and the groups added as a new constraint.

The correspondence analysis results suggests that wrong audio direction, narrower video

width and missing content are the causes for highest degradation scores, whereas audio width, some undefined change in audio, or quietness of audio are not degrading the overall quality experience. These findings support the results obtained by quantitative analysis.

Reasons for the separation of contents were found from the correspondence analysis map, where missing content and audiovisual mismatch were identified as reasons causing degradation in *floorball* and *conversation* contents. These contents were also the only ones, where the video width reduction actually cut away essential information about the environment, *e.g.* the persons having the conversation. On the other hand, these contents had also point-like sound sources that were largely missing from the two other contents. This potentially caused the clearly perceived audiovisual mismatch. Nevertheless, the dimension separating the contents explained only 10% of the constrained variance, so the effect is not very strong

To sum up, this experiment confirmed the results obtained from previous studies and extended them to apply also to immersive audiovisual displays. Also, this experiment showed, that DirAC processing can be reliably used as a sound spatialization technique in perception studies.

In future, new methods to describe multimedia content are needed. Current descriptions rely heavily only on describing video content while neglecting the audio. Modern audio technologies, such as DirAC, enable interpreting the sound scene in greater detail than before and this should be used in advantage. Choosing the contents is considered an essential part of designing a successful experiment and accurate descriptions of both audio and video scenes would certainly help in producing reproducible research. Moreover, future objective multimedia quality metric will be dependent on accurate knowledge about cross-modal effects present in the audiovisual content it is supposed to measure.

Appendix A

Constrained Correspondence Analysis Figures

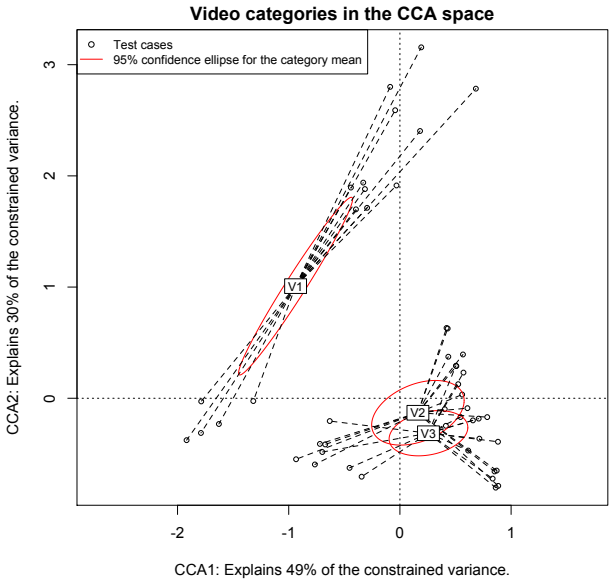

Figure A-1: 95% confidence ellipses for the video category means in the ordination space dimensions 1 and 2. V1=Full, V2=2/3, V3=1/3.

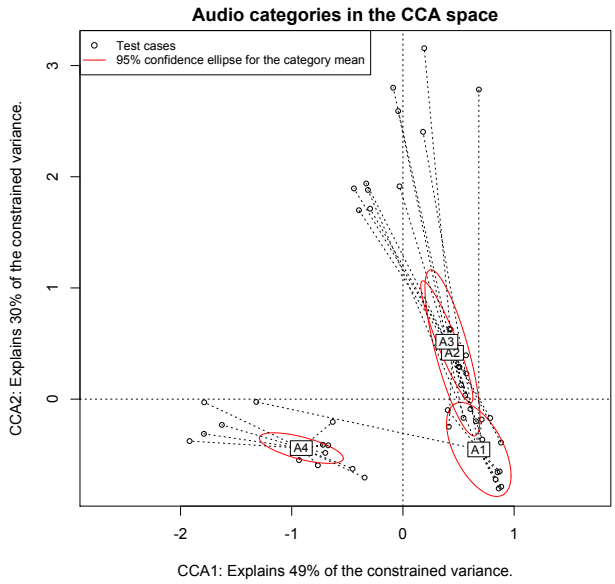

Figure A-2: 95% confidence ellipses for the audio category means in the ordination space dimensions 1 and 2. A1=360°, A2=180°, A3=36°, A4=Mono.

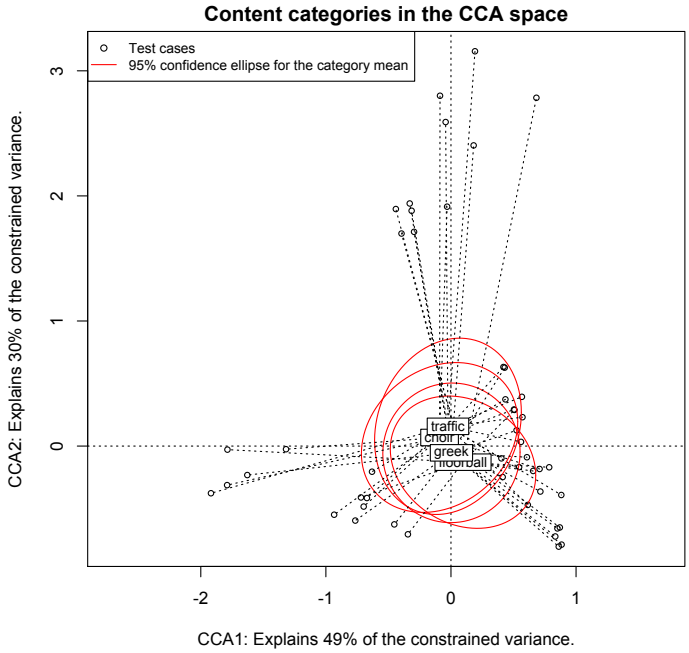

Figure A-3: 95% confidence ellipses for the content means in the ordination space dimensions 1 and 2.

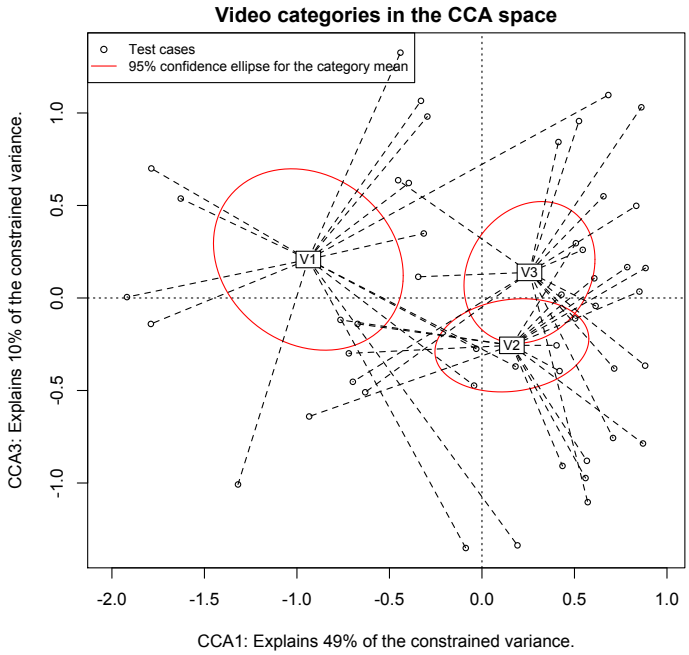

Figure A-4: 95% confidence ellipses for the video category means in the ordination space dimensions 1 and 3. V1=Full, V2=2/3, V3=1/3.

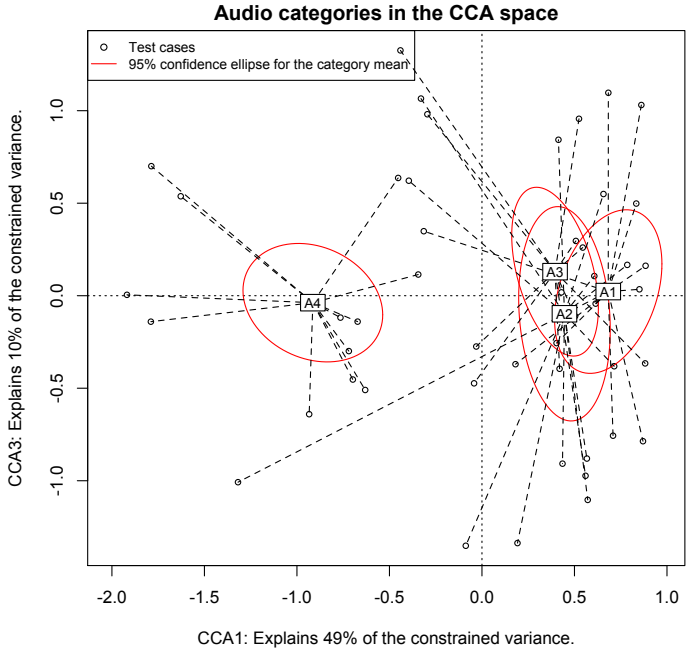

Figure A-5: 95% confidence ellipses for the audio category means in the ordination space dimensions 1 and 3. A1=360°, A2=180°, A3=36°, A4=Mono.

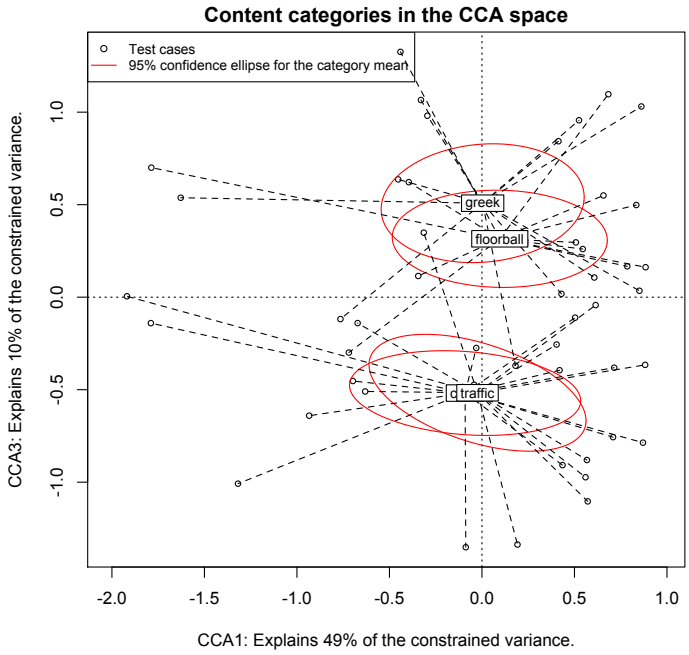

Figure A-6: 95% confidence ellipses for the content means in the ordination space dimensions 1 and 3.

Bibliography

[1] D. Alais and D. Burr. The ventriloquist effect results from near-optimal bimodal integration. *Current biology*, 14:257–262, February 2004.

[2] X. Amatriain, J.A. Kuchera-Morin, T. Hollerer, and S.T. Pope. The allosphere: Immersive multimedia for scientific discovery and artistic exploration. *IEEE Multimedia*, 16(2):64–75, 2009.

[3] ATLAS.ti Scientific Software Development GmbH. Atlas.ti v.6, 2011. Available from: http://www.atlasti.com/.

[4] M. F. Bear, B. W. Connors, and M. A. Paradiso. *Neuroscience: Exploring the Brain*. Lippincott Williams & Wilkins, Philadelphia (PA), 3rd edition, 2007.

[5] S. Bech. The Influence of Stereophonic Width on the Perceived Quality of an Audio-Visual Presentation Using a Multichannel Sound System. In *AES 102nd Convention*, Munich, Germany, 1997.

[6] S. Bech, V. Hansen, and W. Woszczyk. Interaction between audio-visual factors in a home theater system: Experimental results. In *AES 99th Convention*, New York, (NY), 1995.

[7] J.G. Beerends and F.E. De Caluwe. The influence of video quality on perceived audio quality and vice versa. *Journal-Audio Engineering Society*, 47:355–362, 1999.

[8] D. R. Begault. Auditory and non-auditory factors that potentially influence virtual acoustic imagery. In *AES 16th International Conference on Spatial Sound Reproduction*, Rovaniemi, Finland, 1999.

[9] A. J. Berkhout, D. De Vries, and P Vogel. Acoustic control by wave field synthesis. *J. Acoust. Soc. Am*, 93(5):2764–2778, 1993.

[10] J. Blauert. *Spatial Hearing - The Psychophysics of Human Sound Localization*. The MIT Press, Cambridge (MA), 1997.

[11] A. W. Bronkhorst. The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions. *Acta Acustica united with Acustica*, 86(1):117–128, 2000.

[12] C. Cruz-neira, D. J. Sandin, and T. A. DeFanti. Surround-Screen Projection-Based Virtual Reality : The Design and Implementation of the CAVE. In *SIGGRAPH '93 Proceedings of the 20th annual conference on Computer graphics and interactive techniques*, pages 135–142, New York, (NY), 1993.

[13] Cycling 74. Max 5 / MSP / Jitter. Available from: http://cycling74.com/.

[14] T. A. DeFanti, D. Acevedo, R. A. Ainsworth, M. D. Brown, S. Cutchin, G. Dawe, K-U Doerr, A. Johnson, C. Knox, R. Kooima, F. Kuester, J. Leigh, L. Long, P. Otto, V. Petrovic, K. Ponto, A. Prudhomme, R. Rao, L. Renambot, D. J. Sandin, J. P. Schulze, L. Smarr, M. Srinivasan, P. Weber, and G. Wickham. The future of the CAVE. *Central European Journal of Engineering*, 1(1):16–37, November 2010.

[15] J. A. Deutsch and D. Deutsch. Attention: Some theoretical considerations. *Psychological Review*, 70(1):80–90, 1963.

[16] J. Driver and C. Spence. Attention and the crossmodal construction of space. *Trends in Cognitive Sciences*, 2(7):254–262, 1998.

[17] S. Ellison and P. Otto. Acoustics for reproducing sound at the visualization labs at the King Abdullah University of Science and Technology: A case study. In *159th Meeting on Acoustics, Acoustical Society of America*, volume 9, Baltimore, (MD), 2010.

[18] J.S. Farris. *The Human-Web Interaction Cycle: A Proposed And Tested Framework Of Perception, Cognition, And Action On The Web*. PhD thesis, Kansas State University, 2003.

[19] M. N. Garcia and A. Raake. Impairment-factor-based audio-visual quality model for IPTV. In *Quality of Multimedia Experience (QoMEX)*, San Diego (CA), 2009.

[20] M. A. Gerzon. Ambisonics in Multichannel Broadcasting and Video. *Journal of Audio Engineering Society*, 33(11):859–871, 1985.

[21] J. Gómez Bolaños. *Design and Implementation of an Immersive Audiovisual Environment*. Master's thesis, Aalto University School of Electrical Engineering, 2011.

[22] A. Gotchev, A. Smolic, S. Jumisko-Pyykkö, D. Strohmeier, G. B. Akar, P. Merkle, and N. Daskalov. Mobile 3D television: development of core technological elements and user-centered evaluation methods toward an optimized system. In *Proc. SPIE Multimedia on Mobile Devices*, San Jose (CA), 2009.

[23] J. Häkkinen, T. Kawai, J. Takatalo, T. Leisti, J. Radun, A. Hirsaho, and G. Nyman. Measuring stereoscopic image quality experience with interpretation based quality methodology. In *IS&T / SPIE International Symposium on Electronic Imaging*, volume 6808, 2008.

[24] K. Hamasaki, T. Nishiguchi, R. Okumura, Y. Nakayama, and A. Ando. 22.2 Multichannel Sound System for Ultra High-Definition TV. In *Society of Motion Picture and Television Engineers Technical Conference*, 2007.

[25] S. Handel. *Perceptual Coherence: Hearing and Seeing*. Oxford University Press, New York, (NY), 2006.

[26] D. S. Hands. A Basic Multimedia Quality Model. *IEEE Transactions on multimedia*, 6(6):806–816, 2004.

[27] D. H. Hubel. *Eye, Brain and Vision*. W. H. Freeman, Online book, 1995. Available from: http://hubel.med.harvard.edu/book/bcontex.htm.

[28] W. IJsselsteijn and J. Van Baren. Measuring Presence : A Guide to Current Measurement Approaches. *OmniPres Project IST-2001-39237, Deliverable 5*, 2004.

[29] W. IJsselstein, H. de Ridder, J. Freeman, and S. E. Avons. Presence: concept, determinants, and measurement. In *Proc. SPIE Human Vision and Electronic Imaging V*, volume 3959, pages 520–529, 2000.

[30] K. Johnsen and B. Lok. An evaluation of immersive displays for virtual human experiences. In *IEEE Virtual Reality Conference, 2008.*, pages 133–136, Reno (NV), 2008.

[31] R. B. Johnson and A. J. Onwuegbuzie. Mixed Methods Research: A Research Paradigm Whose Time Has Come. *Educational Researcher*, 33(7):14–26, October 2004.

[32] A. Joly, N. Montard, and M. Buttin. Audio-visual quality and interactions between television audio and video. In *International Symposium on Signal Processing and its Applications (ISSPA)*, Kuala Lumpur, Malaysia, 2001.

[33] S. Jumisko-Pyykkö, J. Häkkinen, and G. Nyman. Experienced Quality Factors - Qualitative Evaluation Approach to Audiovisual Quality. In *IS&T / SPIE conference Electronic Imaging, Multimedia on Mobile Devices*, 2007.

[34] S. Jumisko-Pyykkö, U. Reiter, and C. Weigel. Produced quality is not perceived quality - A quantitative approach to overall audiovisual quality. In *3DTV Conference*, pages 1–4, 2007.

[35] M. Karjalainen. *Kommunikaatioakustiikka*. Helsinki University of Technology, Department of Signal Processing and Acoustics, Espoo, Finland, 2009.

[36] J. Korhonen, U. Reiter, and E. Myakotnykh. On the relative importance of audio and video in the presence of packet losses. In *Second International Workshop on Quality of Multimedia Experience (QoMEX)*, pages 64–69, Trondheim, Norway, June 2010.

[37] B. Kunka, B. Kostek, M. Kulesza, P. Szczuko, and A. Czyzewski. Gaze-tracking-based audio-visual correlation analysis employing quality of experience methodology. *Intelligent Decision Technologies*, 4:217–227, 2010.

[38] A. Lee. VirtualDubMod. Available from: http://virtualdubmod.sourceforge.net.

[39] J-S. Lee, F. De Simone, and T. Ebrahimi. Video coding based on audio-visual attention. In *IEEE International Conference on Multimedia and Expo*, pages 57–60, New York, (NY), June 2009.

[40] D. J. Levitin, K. MacLean, M. Mathews, and L. Chu. The perception of cross-modal simultaneity. In *International Journal of Computing and Anticipatory systems*, volume 5, pages 323–329, 2000.

[41] D. W. Massaro and D. S. Warner. Dividing attention between auditory and visual perception. *Perception and Psychophysics*, 21(6), 1977.

[42] H. McGurk and J. MacDonald. Hearing lips and seeing voices. *Nature*, 264:746–748, 1976.

[43] M. Meehan, B. Insko, M. Whitton, and F. P. Brooks, Jr. Physiological measures of presence in stressful virtual envinronments. In *29th Annual Conference on Computer Graphics and Interactive Techniques*, pages 645–652, San Antonio (TX), 2002.

[44] V. Menkovski, G. Exarchakos, and A. Liotta. Online QoE prediction. In *Quality of Multimedia Experience (QoMEX)*, pages 118–123, Trondheim, Norway, 2010. IEEE.

[45] M. Minsky. Telepresence. *Omni*, 1980.

[46] S. Möller, K-P. Engelbrecht, C. Kühnel, I. Wechsung, and B. Weiss. A taxonomy of quality of service and quality of experience of multimodal human-machine interaction. In *Quality of Multimedia Experience (QoMEX)*, pages 7–12, San Diego (CA), 2009.

[47] J. J. Nassi and E. M. Callaway. Parallel processing strategies of the primate visual system. *Nature reviews. Neuroscience*, 10(5):360–372, May 2009.

[48] U. Neisser. *Cognition and reality: Principles and implications of cognitive psychology*. W H Freeman, New York, (NY), 1976.

[49] G. Nyman, J. Radun, T. Leisti, J. Oja, H. Ojanen, J-L. Olives, T. Vuori, and J. Häkkinen. What do users really perceive - Probing the subjective image quality. In *IS&T / SPIE Electronic Imaging*, volume 6059, 2006.

[50] J. Oksanen, F. G. Blanchet, R. Kindt, P. Legendre, P. R. Minchin, R. B. O'Hara, G. L. Simpson, P. Solymos, M. H. H. Stevens, and H. Wagner. vegan: Community Ecology Package, 2011. Available from: http://cran.r-project.org/package=vegan.

[51] R. Patterson, M. D. Winterbottom, and B. J. Pierce. Perceptual Issues in the Use of Head-Mounted Visual Displays. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 48(3):555–573, 2006.

[52] T. Pihlajamäki and V. Pulkki. Low-delay Directional Audio Coding for Real-Time Human-Computer Interaction. In *AES 130th Convention*, pages 1–8, London, UK, 2011.

[53] A. Politis and V. Pulkki. Broadband analysis and synthesis for DirAC using A-format. In *AES 131st Convention*, New York, (NY), 2011.

[54] J.D. Prothero and H.G. Hoffman. Widening the Field-of-View Increases the Sense of Presence in Immersive Virtual Environments. Technical report, Human Interface Technology Laboratory, 1995.

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

[56] V. Pulkki. Spatial sound reproduction with directional audio coding. *Journal of Audio Engineering Society*, 55(6):503–516, 2007.

[57] J. Radun, T. Virtanen, G. Nyman, and J-L. Olives. Explaining multivariate image quality-Interpretation-based quality approach. In *International Congress of Imaging Science*, pages 119–121, Rochester, (NY), 2006.

[58] J. P. Rauschecker. Cortical processing of complex sounds. *Current Opinion in Neurobiology*, 8(4):516–521, 1998.

[59] J. P. Rauschecker and S. K. Scott. Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. *Nature neuroscience*, 12(6):718–724, June 2009.

[60] ITU-R BS.1387 Recommendation. Method for objective measurements of perceived audio quality. *ITU Radiocommunication Sector*, 2001.

[61] ITU-T J.100 Recommendation. Tolerances for transmission time differences between the vision and sound components of a television signal. In *ITU Telecommunication Standardization Sector*, 1990.

[62] ITU-T J.148 Recommendation. Requirements for an objective perceptual multimedia quality model. *ITU Telecommunication Standardization Sector*, 2003.

[63] ITU-T J.247 Recommendation. Objective perceptual multimedia video quality measurement in the presence of a full reference. *ITU Telecommunication Standardization Sector*, 2008.

[64] ITU-T P.10 Recommendation and Amendment 1. Vocabulary for performance and quality of service. *ITU Telecommunication Standardization Sector*, 2007.

[65] ITU-T P.910 Recommendation. Subjective video quality assessment methods for multimedia applications. In *ITU Telecommunication Standardization Sector*, 1999.

[66] ITU-T P.911 Recommendation. Subjective audiovisual quality assessment methods for multimedia applications. *ITU Telecommunication Standardization Sector*, 1998.

[67] ITU-T P.920 Recommendation. Interactive test methods for audiovisual communications. *ITU Telecommunication Standardization Sector*, 1996.

[68] U. Reiter. Subjective Assessment of the Optimum Number of Loudspeaker Channels in Audio-Visual Applications Using Large Screens. In *AES 28th International Conference*, pages 102–109, Piteå, Sweden, 2006.

[69] U. Reiter. *Bimodal audiovisual perception in interactive application systems of moderate complexity.* PhD thesis, Technische Universität Ilmenau, Germany, 2009.

[70] U. Reiter. Perceived Quality in Consumer Electronics - from Quality of Servive to Quality of Experience. In *13th IEEE International Symposium on Consumer Electronics (ISCE)*, Kyoto, Japan, 2009.

[71] U. Reiter. Towards a Classification of Audiovisual Media Content. In *129th AES Convention*, San Francisco, (CA), 2010.

[72] U. Reiter. Reducing the Cost of Audiovisual Content Classification by Experiment. In *AES 131st Convention*, New York, (NY), 2011.

[73] U. Reiter and J. You. Estimating perceived audiovisual and multimedia quality - A survey. In *14th IEEE International Symposium on Consumer Electronics (ISCE)*, Braunschweig, Germany, 2010.

[74] Point Grey Research. Ladybug3. Available from: http://www.ptgrey.com/.

[75] A. Rimell and M. Hollier. The significance of cross-modal interaction in audio-visual quality perception. In *IEEE 3rd Workshop on Multimedia Signal Processing*, pages 509–514, 1999.

[76] A. Rimell and A. Owen. The effect of focused attention on audio-visual quality perception with applications in multi-modal codec design. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 6, pages 2377–2380, 2000.

[77] B. Rudiak-Gould. Huffyuv v2.1.1. Available from: http://neuron2.net/www.math.berkeley.edu/benrg/huffyuv.html.

[78] T. Säämänen, T. Virtanen, and G. Nyman. Videospace: classification of video through shooting context information. *Proc. SPIE*, 7529(752906), 2010.

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

[80] S. H. Schwartz. *Visual Perception: A Clinical Orientation*. McGraw-Hill Professional Publishing, New York, (NY), 4th edition, 2009.

[81] M. Slater. Measuring presence: A response to the Witmer and Singer presence questionnaire. *Presence: Teleoperators & Virtual Environments*, 8(5):560–565, 1999.

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

[83] M. Slater. Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 364(1535):3549–57, December 2009.

[84] M. Slater, M. Usoh, and A. Steed. Depth of presence in virtual environments. *Presence: Teleoperators and Virtual Environments*, 3(2):130–144, 1994.

[85] Soundfield. SPS200. Available from: http://www.soundfield.com/.

[86] C. Spence. Audiovisual multisensory integration. *Acoustical Science and Technology*, 28(2):61–70, 2007.

[87] D. Strohmeier and S. Jumisko-Pyykkö. How does my 3D video sound like?-Impact of loudspeaker set-ups on audiovisual quality on mid-sized autostereoscopic display. In *3DTV Conference: The True Vision-Capture, Transmission and Display of 3D Video*, pages 73–76. IEEE, 2008.

[88] D. Strohmeier, S. Jumisko-Pyykkö, and K. Kunze. Open Profiling of Quality: A Mixed Method Approach to Understanding Multimodal Quality Perception. *Advances in Multimedia*, vol. 2010:1–28, 2010.

[89] J. Takatalo. *Content-Oriented Experience in Entertainment Virtual Environments.* PhD thesis, Institute of Behavioural Sciences, University of Helsinki, 2011.

[90] D. L. Valente and J. Braasch. Subjective scaling of spatial room acoustic parameters influenced by visual environmental cues. *The Journal of the Acoustical Society of America*, 128(4):1952–1964, October 2010.

[91] J. Vilkamo, T. Lokki, and V. Pulkki. Directional Audio Coding: Virtual Microphone-Based Synthesis and Subjective Evaluation. *Journal of Audio Engineering Society*, 57(9):709–724, 2009.

[92] T. Virtanen, J. Radun, P. Lindroos, and S. Suomi. Forming valid scales for subjective video quality measurement based on a hybrid qualitative/quantitative methodology. *SPIE-IS&T Electronic Imaging*, 6808:1–11, 2008.

[93] S. Waxman. *Clinical Neuroanatomy.* McGraw-Hill Professional Publishing, New York, (NY), 26th edition, 2009.

[94] D. Weibel, B. Wissmath, and F. W. Mast. Immersion in mediated environments: the role of personality traits. *Cyberpsychology, behavior and social networking*, 13(3):251–6, June 2010.

[95] W. Wirth, T. Hartmann, S. Böcking, P. Vorderer, C. Klimmt, H. Schramm, T. Saari, J. Laarni, N. Ravaja, F.R. Gouveia, F. Biocca, A. Sacau, L. Jäncke, T. Baumgartner, and P. Jäncke. A process model of the formation of spatial presence experiences. *Media Psychology*, 9(3):493–525, 2007.

[96] B. G. Witmer and M. J. Singer. Measuring Presence in Virtual Environments: A Presence Questionnaire. *Presence*, 7(3):225–240, 1998.

[97] N. Wood and N. Cowan. The cocktail party phenomenon revisited: How frequent are attention shifts to one's name in an irrelevant auditory channel? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21(1):225–260, 1995.

[98] W. Woszczyk, S. Bech, and V. Hansen. Interaction between audio-visual factors in a home theater system: Definition of subjective attributes. In *AES 99th Convention*, New York, NY, 1995.

[99] S. Yantis and J. Jonides. Abrupt visual onsets and selective attention: Voluntary versus automatic allocation. *Journal of Experimental Psychology: Human Peception and Performance*, 16(1):121–134, 1990.

[100] W. A. Yost. *Fundamentals of Hearing: An Introduction.* Academic Press, Inc., San Diego (CA), 3rd edition, 1994.

[101] J. You, U. Reiter, M. M. Hannuksela, M. Gabbouj, and A. Perkins. Perceptual-based quality assessment for audio-visual services: A survey. *Signal Processing: Image Communication*, 25:482–501, 2010.

[102] S. K. Zielinski, F. Rumsey, S. Bech, B. de Bruyn, and R. Kassier. Computer games and multichannel audio quality - The effect of division of attention between auditory and visual modalities. In *AES 24th International Conference on Multichannel Audio*, 2003.