

Game Sound Technology and Player Interaction: Concepts and Developments

Mark Grimshaw
University of Bolton, UK

Information Science
REFERENCE

INFORMATION SCIENCE REFERENCE
Hershey • New York

Director of Editorial Content: Kristin Klinger
Director of Book Publications: Julia Mosemann
Acquisitions Editor: Lindsay Johnston
Development Editor: Joel Gamon
Publishing Assistant: Milan Vracarich Jr.
Typesetter: Natalie Pronio
Production Editor: Jamie Snavelly
Cover Design: Lisa Tosheff

Published in the United States of America by
Information Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue
Hershey PA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-global.com
Web site: <http://www.igi-global.com>

Copyright © 2011 by IGI Global. All rights reserved. No part of this publication may be reproduced, stored or distributed in any form or by any means, electronic or mechanical, including photocopying, without written permission from the publisher. Product or company names used in this set are for identification purposes only. Inclusion of the names of the products or companies does not indicate a claim of ownership by IGI Global of the trademark or registered trademark.

Library of Congress Cataloging-in-Publication Data

Game sound technology and player interaction : concepts and development / Mark Grimshaw, editor. p. cm.

Summary: "This book researches both how game sound affects a player psychologically, emotionally, and physiologically, and how this relationship itself impacts the design of computer game sound and the development of technology"-- Provided by publisher. Includes bibliographical references and index. ISBN 978-1-61692-828-5 (hardcover) -- ISBN 978-1-61692-830-8 (ebook) 1. Computer games--Design. 2. Sound--Psychological aspects. 3. Sound--Physiological effect. 4. Human-computer interaction. I. Grimshaw, Mark, 1963-

QA76.76.C672G366 2011
794.8'1536--dc22

2010035721

British Cataloguing in Publication Data

A Cataloguing in Publication record for this book is available from the British Library.

All work contributed to this book is new, previously-unpublished material. The views expressed in this book are those of the authors, but not necessarily of the publisher.

Chapter 8

Perceived Quality in Game Audio

Ulrich Reiter

Norwegian University of Science and Technology, Norway

ABSTRACT

This chapter reviews game audio from a Quality of Experience point of view. It describes cross-modal interaction of auditory and visual stimuli, re-introduces the concept of plausibility, and discusses issues of interactivity and attention as the basis for the qualitative, high-level salience model being suggested here. The model is substantiated by experimental results indicating that interaction or task located in the audio domain clearly influences the perceived audio quality. Cross-modal influence, with interaction or task located in a different (for example, visual) domain, is possible, but is significantly harder to predict and evaluate.

INTRODUCTION

Perceived quality in game audio is not a question of audio quality alone. As audio is usually only a part in an overall game concept consisting of graphics, physics, artificial intelligence, user input, feedback and so forth, audio has been considered to play a relatively minor role in the overall experience that a game provides. Consequently, a lot of effort has been put into providing near photo-realistic representations of (virtual) game scenarios to the player, but only little into audio. Interestingly,

this assessment has had to be revised over the last years. Learning from other artistic fields like cinema, in which storytelling is a central means of providing “user experience”, game developers have come to know that audio can trigger emotions and provide additional information otherwise hard to convey. Today, although budgets are still limited compared to other aspects of game engineering, audio in games is given more attention by the game developers than ever before.

But there is more to audio in games than just an emotional support for a story. Most games are user-centered and non-linear, as opposed to the linear story telling of traditional, non-interactive

DOI: 10.4018/978-1-61692-828-5.ch008

content presentation. Therefore, the audio has to be manipulated in real-time depending on the player's actions. Real-time processing of audio can become computationally very demanding and is a problem for complex game scenarios. This has introduced the concept of plausibility: the main goal in game audio is not to have an audio simulation as exact and close to reality as possible, but to render audio that is plausible in the game scenario, and that provides an overall quality impression that matches the other aspects of the game.

One fact well known from home cinema applications is that an improved quality in video can also increase the subjectively perceived audio quality, and that the reverse effect also exists (Beerends & De Caluwe, 1999). It is therefore a most interesting question to see whether these effects can be exploited to increase the subjectively perceived overall quality of a game without actually increasing the computational load. Instead of just rendering more details (equivalent to a higher simulation depth), focusing on those details that are actually relevant in a certain context could provide a much higher Quality of Experience (QoE) (see Farnell, 2011 for a discussion of relevancy and redundancy in procedural audio design).

The central question is, therefore, which stimuli in a game scenario are of most importance? Can information that is difficult and cost-intensive to convey in one modality be presented in another modality with less effort but similar perceptual impact? What role does interactivity play in the perception of quality? What are the technical parameters that can influence the perceived quality of a game, and which other factors exist that potentially dominate the perceptual process?

This chapter aims at identifying and discussing general quality criteria in multimedia application systems with a focus on games. These criteria contain technical as well as human factors. In order to understand these factors, the first section touches upon the mechanisms of human percep-

tion: well-known facts about visual and auditory perception are summarized briefly.

The second section presents a discussion of cross-modal influences, that is, interaction between auditory and visual stimuli in the perceptual apparatus, and cross-modality in general. A survey detailing the most accepted theories of how audio-visual (bimodal) perception is achieved in the human brain is given. This is far more complex than just adding the results of auditory and visual processing and is therefore worth an extended discussion. This is followed by examples of effects in bimodal perception (based on research in the fields of psychology and cognitive sciences) that can be relevant in the context of game audio.

The third section discusses the concept of auditory and audio-visual plausibility. It briefly compares the requirements for exact (room) acoustic simulations versus real-time rendering and details the constraints resulting for computer games.

The next section gives an overview on issues related to interactivity, such as latency, user input, and perceptual feedback. Interactivity is closely related to the generation of presence, defined as the "perceptual illusion of non-mediation", or simply the feeling of "being there". The concept of presence is discussed as an indirect measure for perceived quality.

The fifth section elaborates on the concept of attention. The perception of multiple streams is discussed and an introduction to the general model of the Perceptual Cycle according to Neisser is given. From this, the concepts of selective attention and divided attention are discussed and capacity limits of the human perceptual system are explained.

Finally, in the sixth section, the resulting factors (technical as well as human) are arranged to form a qualitative model describing human audio-visual perception based on saliency of stimuli. Such a model can serve as a basis for determining the QoE in games in general and specifically for game audio. Experimental results documenting inner-

modal versus cross-modal effects on perceived audio quality are summarized.

Finally, a summary is given that reviews the most important concepts leading to the salience model presented in the preceding section. Further research potential is defined.

MECHANISMS OF HUMAN PERCEPTION

Vision (sight) and audition (hearing) are the most important human senses for playing games. In the real world, these senses provide us with information about the more remote surroundings, as opposed to taste (gustation), smell (olfaction), and touch (taction or pressure) which provide information about our immediate vicinity. Because vision and audition communicate spatial and temporal relations of objects, and because the necessary technology to stimulate the two is readily available on computer systems used in the home, most games only stimulate the two.

Visual Perception

Vision mainly serves to indicate spatial correlation of objects, as the human visual system seldom responds to direct light stimulation. Rather, light is reflected by objects and thus transmits information about certain characteristics of the object. The direction of a visually perceived object corresponds directly to the position of its image on the retina, the place where the light receptors are located in the eye. At the same time, a visual stimulus occupies a position in perceptual space that is defined relative to a distance axis, as well as to the vertical and horizontal axes.

In the determination of an object's distance to the eye, there are a number of potential cues of depth. These include monocular mechanisms like interposition, size, and linear perspective as well as binocular cues like convergence and disparity. All of these are usually evaluated jointly, allow-

ing us to solve even ambiguous situations with contradicting sensory information.

All these depth cues can be exploited even when the environment is at rest. As soon as motion (of objects or of the head) is present, motion parallax takes on an important role in depth perception. Motion parallax describes the fact that the image of an object far away from the viewer moves more slowly across the retina than the image of an object at a close distance. Motion parallax also provides cues in the monocular case.

Auditory Perception

Auditory stimuli are perceived to be localized in space. The sound is not heard within the ear, but it is phenomenally positioned at the source of the sound. In order to localize a sound, the auditory system relies on binaural and monaural acoustic cues. Directional hearing in the horizontal plane (azimuth) is dominated by two mechanisms which exploit binaural time differences and binaural intensity differences. For sinusoidal signals, interaural time differences (ITDs, the same stimulus arriving at different times at the left and the right ear) can be interpreted by the human hearing system as directional cues from around 80Hz up to a maximum frequency of around 1500Hz. This maximum frequency corresponds to a wavelength of roughly the distance between the two ears. For higher frequencies, more than one wavelength fits between the two ears, making the comparison of phase information between left and right ear equivocal (Braasch, 2005). For signals with frequencies above 1500Hz, interaural level differences (ILDs) between the two ears are the primary cues (Blauert, 2001). Regardless of the source position, ILDs are small at low frequencies. This is because the dimensions of head and pinnae (the outer ear visible on the side of the head) are small in comparison to the wavelengths at frequencies below about 1500Hz. Therefore they do not represent any noteworthy obstacle for the propagation of sound.

Directional hearing in the vertical plane (elevation) is dominated by monaural cues. These stem from direction-dependent spectral filtering caused by reflection and diffraction at the torso, head, and pinnae. Each direction of incidence (for instance, defined in terms of azimuth and elevation) is related to a unique spectral filtering for each individual. This spectral filtering can be described by head-related transfer functions (HRTFs). In addition to providing localization of sounds in the vertical plane, these spectral cues are also essential for resolving front-back confusions (Blauert, 2001). Pulkki (2001) reports that, for elevation perception, frequencies around 6kHz are especially important.

In everyday situations, localization of sound sources seldom relies on auditory cues alone. Knowledge of the potential source of a sound (for example, airplane noises from above, or crunching shoes from below) aids in the localization process. Visual cues heavily influence the localization of sound sources.

CROSS-MODAL INTERACTION BETWEEN AUDIO AND VIDEO

Human perception in real world situations is a multi-modal, recursive process. Stimuli from different modalities usually complement each other and make the perceptual process more unequivocal. Only those stimuli that can actually be perceived by the primary receptors of sound, light, pressure and so on contribute to an overall impression (which is the result of any perceptual process). The human perceptual process, because of its complexity, cannot easily be explained in a simple block diagram without neglecting important features. A number of descriptive models exist, but these only cover certain aspects of the process, depending on the level of abstraction at which the respective model is located.

Relatively little is known about the mechanisms of multi-modal processing in the human brain.

The main questions with respect to audio-visual perception are: At what level of perceptual processing do cross-modal interactions occur? And what mechanism underlies them?

Joint Processing of Audio-Visual Stimuli

As early as 1909, Brodmann suggested a division of the cerebral cortex into 52 distinct regions, based on their histological characteristics (Brodmann, 1909). These areas, today called *Brodmann areas*, have later been associated to nervous functions. The most important areas in the audio-visual game context are Primary Visual Cortex (V1), Visual Association Cortex (V2 and V3), as well as Primary Auditory Cortex and anterior and posterior transverse temporal areas (H). This division suggests that the different modalities are related to separate regions of the brain, and that processing of stimuli is performed separately for each modality.

Taking a closer look at the brain reveals that the neurons of the neocortex are arranged in six horizontal layers, parallel to the surface. The functional units of cortical activity are organized in groups of neurons. These are connected by four types of fibers, of which the association fibers are especially interesting when looking at information exchange between cortical areas. Short association fibers (called loops) connect adjacent gyri, whereas long association fibers form bundles to connect more distant gyri in the same hemisphere. These association bundles give fibers to and receive fibers from the overlying gyri along their routes. They occupy most of the space underneath the cortex.

There are many such connections between different functional areas of the neocortex such that information can be exchanged between them and true multi-modal processing can be achieved. Goldstein (2002) gives an example of a red, rolling ball entering our field of view. Locally distinct neurons are then activated by either motion, shape, or color. Subsequently, dorsal and

ventral streams are also activated. Although the involved neurons are locally distinct, we perceive one singular object, not separate *rolling*, *red color*, or *round shape*.

Until now, it is unclear how the processing of multiple characteristics of a single object is organized. A number of theories have been suggested to explain this binding problem, and the exploration of binding in the visual system has become a heavily discussed topic. According to Goldstein (2002), the most prominent theory, suggested by Singer, Engel, Kreiter, Munk, Neuenschwander, and Roelfsema (1997), assumes that visual objects are represented by groups of neurons. These so-called cell-assemblies are activated jointly, producing an oscillatory response. This way, neurons belonging to the same cell-assembly can synchronize. Whenever the reaction to stimuli is synchronized, this means that the respective cortical areas are processing data coming from one single object or context.

Yet, this binding by synchrony theory has left doubts with respect to the interpretation and processing of the synchrony code. For example, Klein, König, and Körding (2003) postulate that “many properties of the mammalian visual system can be explained as leading to optimally sparse neural responses in response to pictures of natural scenes” (p. 659). According to Goldstein (2002), many others argue that binding can be explained by (selective) attention. Attention is discussed below.

Dominance of Single Modalities

Very often the dominance of visual stimuli over other modalities is accepted naturally as a given. In fact, looking at our everyday experiences we might be inclined to accept this posit without further discussion: because “seeing is believing”, we often think that we tend to trust our eyes more than the other senses. Yet, this appraisal is often due to the fact that in the real world we seldom have to face contradictions in the multi-modal

stimuli perceived by our senses. There is actually no need to consciously further evaluate the different percepts in terms of relevance, because they usually complement (and not contradict) each other.

In order to actually verify any naturally given order of significance of the perceived stimuli, it is necessary to present the human perceptual system with contradictory sensory information and see what the generally dominating modality is—if there is any. There have been a number of scientific efforts to explain in a *perceptual relevance model* how the human perceptual system weighs the different contradicting percepts.

Two such models have been proposed to describe how perceptual judgments are made when signals from different modalities are conflicting. One of these models suggests that the signal that is typically most reliable dominates the competition completely in a winner-take-it-all fashion: the judgment is based exclusively on the dominant signal. In the context of spatial localization based on visual and auditory cues, this model is called *visual capture* because localization judgments are made based on visual information only. The other model suggests that perceptual judgments are based on a mixture of information originating from multiple modalities. This can be described as an optimal model of sensory integration which has been derived based on the maximum-likelihood estimation (MLE) theory. This model assumes that the percepts in the different modalities are statistically independent and that the estimate of a property under examination by a human observer has a normal distribution. In engineering literature, the MLE model is also known as the Kalman Filter (Kalman & Bucy, 1961).

Battaglia, Jacobs, and Aslin (2003) report that several investigators have examined whether human adults actually combine information from multiple sensory sources in a statistically optimal manner (that is, according to the MLE model). They explain:

According to this model, a sensory source is reliable if the distribution of inferences based on that source has a relatively small variance; otherwise the source is regarded as unreliable. More-reliable sources are assigned a larger weight in a linear-cue-combination rule, and less reliable sources are assigned a smaller weight. (Battaglia et al., 2003, p. 1391)

Looking at it this way, visual capture is just a special case of the MLE model: the highly reliable percept (the visual cue) is assigned a weight of one, whereas the less reliable percept (the auditory cue) is assigned a weight of zero.

Battaglia et al. (2003) describe an experiment designed to answer the question whether human observers localize events presented simultaneously in the auditory and visual domain in a way that is best predicted by the visual capture model or by the MLE model. Their report suggests that both models are partially correct and that a hybrid model may provide the best account of their subjects' performances. As greater amounts of noise were added to the visual signal, subjects used more and more information perceived via the auditory channel, as suggested by the MLE model. Yet most notably, according to their analysis, test subjects seemed to be biased towards using visual information to a greater extent than originally predicted by the MLE model. This means that the model used in the experiments committed a systematic error by constantly underestimating the test subjects' use of visual information (thus overestimating the use of auditory information).

Shams, Kamitani, and Shimojo (2000, 2002) describe experiments in which visual illusion was induced by sound, resulting in the auditory cue outweighing the visual cue. They presented test subjects with flashes of light and beeps of sound: whenever a single flash of light was accompanied by multiple auditory beeps, the single flash was perceived as multiple flashes. They conclude that this alteration of the visual percept is caused by cross-modal perceptual interactions, rather than

having cognitive, attentional, or other origins. This is especially interesting as there was no degradation in the quality of the visual percept offered, which otherwise inevitably provokes the human perceptual system to rely on other modalities.

To sum up, the combined results of these experiments suggest that there is no clear, generalized bias of humans toward any of the available modalities in terms of dominance. Apparently, there is no such thing as a general dominance of visual percepts over other stimuli. Instead, whenever such a bias toward any of the available modalities exists, this seems to be highly dependent on the context. Whereas Battaglia et al. (2003) tested subjects for contradicting localization cues and were presented with a bias toward the visual percept, Shams et al. (2000) tested subjects for temporal variations of cues and were presented with a bias toward the auditory percept. This actually indicates that the human perceptual system tends to prefer those senses (give a higher weight to those percepts) that promise a higher degree of reliability or resolution for the presented perceptual problem: Whereas the horizontal resolution of the human auditory system is roughly 2 to 3 degrees for sinusoidal signals coming from a forward direction (Zwicker & Fastl, 1999), the resolution of the visual system is at least 100 times as high, about 1 min. of arc (Howard, 1982). On the other hand, the time resolution of the auditory system allows to resolve the temporal structure of sounds as close as 2ms (Zwicker & Fastl, 1999), whereas the human visual system can be tricked into believing in a continuously moving object when presented with only 24 sampled pictures of the continuous movement per second.

AUDITORY AND AUDIO-VISUAL PLAUSIBILITY

In classic room acoustic simulation, the time necessary to render the room audible (in other words, to perform the room acoustic simulation

itself), is often considered second-rank. Instead, the (acoustic) similarity between the simulation and the real situation is considered most important. In games, this situation is reversed: the available computational power is critical, and rendering has to be performed in real-time. Therefore, the concept of plausibility is applied: as long as there is no obvious contradiction between the visual and the acoustic representation of a virtual scene, the human senses merge auditory and visual impressions. Hence, it is usually possible to replace a cost-intensive geometry-based room acoustic simulation with a generic reverberation algorithm, for example, with combinations of all pass filters and delays according to Schroeder (1962, 1970), with nested all pass filters according to Gardner (1992), or with feedback delay line structures according to Jot and Chaigne (1991). This way, the auditory part of the presentation provides a rough sketch of the room's characteristics, whereas the visual part complements the overall impression with an increased level of detail. As long as the information provided in the two modalities is not contradictory, there is a high chance that the player's perceptual apparatus merges the stimuli and blends them to form a single, multi-modal representation of the scene.

In general, it might be arguable whether a "perfect" reproduction of the properties of a real life experience will ever be possible in a computer game at all (with the assumption that a simulation is good enough as long as there is no perceptual difference to reality detectable by the human senses in the given situation). A lesser interpretation of this applies to scenes which have no counterpart in reality: their appearance needs to be plausible in every aspect and also in a sense of perfect agreement between the cues offered by the system in the different perceptual domains.

In the context of games, this requirement can be further reduced. Because the visual representation of the scene is limited to a region in the frontal area and is not supposed to fill the field of view entirely, it suffices to require that the one part of

the virtual scene that is displayed (audio-visually) is perceived as plausible. It is thus accepted that stimuli coming from the surrounding real world (which cannot be entirely excluded in a typical computer game playing environment) might interfere with those from the virtual scene.

Furthermore, the time and investment necessary to develop completely accurate auditory and visual models is as much of a limiting factor for how much detail will be rendered, as is the computational power alone. It is therefore reasonable to focus only on the most important stimuli and leave out those that would go unnoticed in a real world situation. In order to do so, it is necessary to predict what the most important stimuli or objects in the overall audio-visual percept are.

INTERACTIVITY ISSUES AND PRESENCE

The concept of interactivity has been defined by Lee, Jin, Park, and Kang (2005) and Lee, Jeong, Park, and Ryu (2007) based on three major viewpoints: technology-oriented, communication-setting oriented, and individual-oriented views. Here, the technology-oriented view of interactivity is adopted, which "defines interactivity as a characteristic of new technologies that makes an individual's participation in a communication setting possible and efficient" (Lee et al., 2007).

Steuer (1992) holds that interactivity is a stimulus-driven variable which is determined by the technological structure of the medium. According to Steuer, interactivity is "the extent to which users can participate in modifying the form and content of a mediated environment in real time" (p. 14)—in other words, the degree to which users can influence a target environment. He identifies three factors that contribute to interactivity:

- *speed* (the rate at which input can be assimilated into the mediated environment)

- *range* (the number of possibilities for action at any given time)
- *mapping* (the ability of a system to map its controls to changes in the mediated environment in a natural and predictable manner).

These factors are related to technological constraints that come into play when an application is supposed to provide interactivity to the user, as is the case for computer games. These technological constraints are briefly discussed in the following subsections.

Latency

Latency is one of the main concerns in computer games. Latency in the context of interactivity can be defined as the time that elapses between a user input and the apparent reaction of the system to that input. It is closely related to Steuer's *speed* factor.

Latencies are introduced by individual components of the system. These components may include input devices, signal processing algorithms, device drivers, communication lines and so on. Although these components may interact in more than one way on a game platform, a system's end-to-end latency should not vary over time to make it predictable.

Meehan, Razaque, Whitton, and Brooks (2003) report a study in which they tested the perceived sense of presence (see below) for two different end-to-end latencies in a Virtual Environment (VE). The low latency was 50ms, the high latency was 90ms. Test subjects were presented with a relaxing environment that was switched to a threatening one and their response was observed. Meehan et al. report that subjects in the low-latency group had a higher self-reported sense of presence and a statistically higher change in heart rate between presentations of the two situations.

MacKenzie and Ware (1993) conducted the first quantitative experiments with respect to effects of visual latency. Participants completed a

Fitts' Law target acquisition task in which they had to move the mouse from a starting point to a target, with a latency of between 25ms and 225ms from moving the mouse to actually seeing the cursor move on the screen. The authors report that the threshold at which latency started to affect the performance was approximately 75ms. This effect was also dependent on the difficulty of the task: the harder the task, the greater was the adverse effect caused by increased latency.

Wenzel (1998, 1999, 2001) has published a number of reports about the impact of system latency on dynamic performance in virtual acoustic environments with a focus on localization of sound sources. The bottom line is that depending on the source velocity of the audio signal itself, localization of sound sources might be impaired when total system latency (end-to-end latency) is higher than around 60ms for audio-only presentations (Wenzel, 1998). On the other hand, error rates in an active localization task, tested on an HRTF-based reproduction system, showed comparable error rates for both low and very high latencies suggesting that subjects were largely able to ignore latency altogether (Wenzel, 2001).

Nordahl (2005) examined the impact of self-induced footstep sounds on the perception of presence and latency. Interestingly, for audio-visual feedback in a VE, the maximum sound delay that was possible without latency being perceived as such was around 50% higher than for the audio-only feedback case (mean values of 60.9ms against 41.7ms). Nordahl explains this as attention being focused mainly on the visual, rather than the auditory feedback in the audio-visual case.

Looking at these experimental results, it is difficult to draw a general conclusion on the maximum allowed latency for computer games. Apparently, the perception of latency as such depends on the system setup itself (screen, loudspeakers/headphones, for example), on the task, and on the content that is displayed. At the same time, measuring total system latency correctly is not a trivial task. Therefore, a general recommendation

would be to keep latency as low as possible within any such system, that is, preferably below 50ms.

Input and Perceptual Feedback

Perceptual feedback is the response that a system provides to the player's input. In games, perceptual feedback is usually provided in the auditory and visual domains. Input provided by the player can, in the general case, consist of any kind of signal accepted by the system for controlling it: speech, gesture, haptic control, eye tracking and so forth.

Input and perceptual feedback are related to Steuer's (1992) *mapping* factor and his *range* factor is related to the kind of interaction that is offered by the game. This depends strongly on the goal of the application or game itself. In a first-person shooter, players might expect a different range of interaction than in a business simulation game.

Hence, both input and perceptual feedback define the degree of interactivity a game player can experience.

Presence

Closely related to interactivity is presence. Larsson, Västfjäll, and Kleiner (2003) define presence in interactive audio-visual application systems or VEs "as the feeling of 'being there'" (p. 98), and as the element that generates involvement of the user. Lombard and Ditton (1997) define presence in a broader sense as the "perceptual illusion of nonmediation" (p. 24).

According to Steuer (1992), the level of interactivity (degree to which users can influence the target environment) has been found to be one of the key factors for the degree of involvement of a user. Steuer has found vividness (ability to technologically display sensory rich environments) to be the second fundamental component of presence. Along the same lines, Sheridan (1994) assumes the quality and extent of sensory information that is fed back to the user, as well as exploration and manipulation capabilities, to be crucial for the

subjective feeling of presence. Other factors have been found to be determinants for presence but these depend on the theoretical concept applied by the researcher.

Ellis (1996) points out that presence may not necessarily be the ultimate goal of every interactive audio-visual application system. He holds that successful task accomplishment can be far more important than presence, especially in situations "where the medium itself is not the message" (p. 253). This is easily accepted for player-game interaction, but is also applicable to communication between players in a multi-player game environment, when players have to team up to achieve a certain goal.

ATTENTION

When being confronted with an increased number of stimuli, the human perceptual apparatus will try to keep up with the processing required for the input on offer. Generally, this can be achieved using different strategies. According to Pashler (1999), all of them are usually referred to as *attention*.

Many human activities require that information from a multitude of sources is taken in. When we attempt to monitor one stream of information, we pay attention to the source. Usually, natural scenes are multi-modal, thus providing information in more than one modality. Also, natural scenes usually provide more than one informational stream. The question is then, how is attention distributed if a multitude of information is presented in more than one stream? What role does multi-modality of the information play in computer games?

Perception of Multiple Streams

Eijkman and Vendrik (1965) conducted one of the earliest studies on the perception of bimodal stimuli. They asked test subjects to detect increments in the intensity of light and tones. The stimuli lasted one second and were presented

either separately or simultaneously. Subjects detected the increments in one modality without interference from simultaneously monitoring the other modality, and performance of detection was comparable to that of only monitoring one modality. Other studies, for example, Shiffrin and Grantham (1974) and by Gescheider, Sager, and Ruffolo (1975), also support these results for presentations of short bimodal stimuli.

As the stimuli presented in the auditory and the visual modalities were not contextually related in the study of Eijkman and Vendrik (1965), they constituted what could be called separate perceptual streams. Yet, detection of increments in the duration of the same stimuli was showing marked interference. This suggests that temporal judgments might be processed by the same processing system (the same cortical areas), a theory that is further supported by the findings of Shams et al. (2000, 2002) already discussed in the subsection on visual dominance.

Interestingly, other studies combining auditory and visual discrimination tasks showed modest but considerable decrements in terms of performance. This was observed when test subjects were confronted with bimodal stimuli in comparison to unimodal ones. To give an example, Tulving

and Lindsay (1967) presented test subjects with tones and patches of light. Subjects were asked to judge the intensity of either tone or light, and results were compared to the bimodal judgment of intensity of both stimuli.

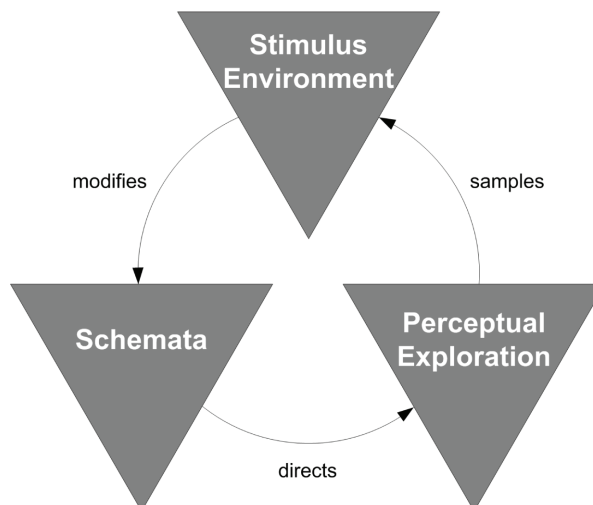
All of these studies characteristically involve magnitude judgments rather than categorical judgments. Therefore, the performance of test subjects in the bimodal case might have been limited by the difficulty of maintaining a standard in memory against which to judge the inputs, rather than by the influence of a second modality itself.

The Perceptual Cycle

Neisser's model of the Perceptual Cycle describes perception as a setup of schemata, perceptual exploration and stimulus environment (Farris, 2003). These elements influence each other in a continuously updated circular process, see Figure 1. Thus, Neisser's model describes at a very abstract level how the perception of the environment is influenced by background knowledge, which in turn is updated by the perceived stimuli.

In Neisser's model, schemata represent an individual's knowledge about the environment. Schemata are based on previous experiences and

Figure 1. The Perceptual Cycle after Neisser. (Adapted from Farris, 2003)



are located in the long term memory. Neisser attributes to them the generation of certain expectations and emotions that steer our attention in the further exploration of our environment. The exploratory process consists, according to Neisser, in the transfer of sensory information (the stimulus) into the short-term memory. In the exploratory process, the entirety of stimuli (the stimulus environment) is compared to the schemata already known. Recognized stimuli are given a meaning, whereas unrecognized stimuli will modify the schemata, which will then in turn direct the exploratory process further (Goldstein, 2002, Farris, 2003).

Returning to the area of games, the differences in schemata between human individuals cause the same stimulus to provoke different reactions in different game players. Following Neisser's model, new experiences (those that cause a modification of existing schemata) are especially likely to generate a higher load in terms of processing requirements. Schemata therefore also control the attention that we pay toward stimuli. The exploratory process is directed in the same way for multi-modal stimuli as for unimodal stimuli.

Selective Attention

An unmanageable number of studies have tried to identify and describe the strategies that are actually used in the human perceptual process. Pashler (1999) gives an overview and identifies two main concepts of attention: attention as based on exclusion (gating) or based on capacity (resource) allocation. The first concept defines the *mechanism that reduces processing of irrelevant stimuli* to be attention. It can be regarded as a filtering device that keeps out stimuli from the perceptual machinery that performs the recognition. Attention is therefore identified with a purely exclusionary mechanism.

The second concept construes *the limited processing resource* (rather than the filtering device) as attention. It suggests that when attention is given

to an object, it is perceptually analyzed. When attention is allocated to several objects, they are processed in parallel until the capacity limits are exceeded. In that case, processing becomes less efficient or eventually impossible.

Neither of the two concepts can be ruled out by the many investigations performed in the scientific community up to now. Instead, assuming either the gating or the resource interpretation, all empirical results can be explained in some way or other. As a result it must be concluded that both capacity limits and perceptual gating characterize human perceptual processing. This combined concept is termed controlled parallel processing (CPP). It claims that parallel processing of different objects is achievable but optional. At the same time, selective processing of a single object is possible, largely preventing other stimuli from undergoing full perceptual analysis.

In fact, further conceptualizing attention might not even be possible unless we understood the neural circuitry and operations that underlie these processes in detail. Rather, in the context of bimodal perception it is interesting whether there are separate perceptual attention systems associated with different sensory modalities or whether a unified multi-modal attention system exists. Are visual and auditory attention the same thing? According to Pashler (1999), investigations have shown that humans are capable of selecting visual stimuli in one location in space and auditory stimuli in another.

Spence, Nicholls, and Driver (2001) have examined the effect of expecting a stimulus in a certain modality upon human performance. They measured the reaction time to a stimulus located in the auditory, visual, or tactile modality between different frequencies of occurrence (equal number of targets in all modalities against a 75% majority of targets located in one modality). Spence et al. report that reaction times for targets in the unexpected modalities were slower than for the expected modality or no expectancy at all. They further state that shifting attention away from the

tactile modality was taking longer than shifting from the auditory or visual modality. These results show that performance not only depends on what actually happens, but also on what is anticipated by a game player. Yet, it must also be noted that in this study a faster response time for the most likely modality was always related to priming from an event in the same modality on the previous trial, and not to the expectancy as such.

Alais and Blake (1999) have found evidence that attention focused on a visual object markedly amplifies neural activity produced by features of the attended object. They applied single-cell and neuroimaging studies and reinforce that visual attention modulates neural activity in several areas within the visual cortex. They state that “attentional modulation seems to involve a boost in the gain of responses of cells to their preferred stimuli, not a sharpening of their stimulus selectivity” (p. 1015).

These findings clearly indicate that the perceptual process is actually controlled by attention. They can not fully answer the question whether there is one multi-modal attention or whether attentions are associated with modalities. However, there are indicators that favor the latter.

Divided Attention and Perceptual Capacity Limits

One of these indicators is that capacity limits appear to be more severe when multiple stimuli are presented in the same modality compared with multiple modalities (Pashler, 1999; Reiter, Weitzel, and Cao, 2007; Reiter & Weitzel, 2007; Reiter, 2009). This means that capacity limits may occur earlier and more frequently if the main task and the so-called *distractors* (stimuli that are not directly related to the task/the direct focus of attention) are located in the same modality.

In an overview article, Lavie (2001) examines the capacity limits in selective attention. Lavie reasserts and concludes what evidence from several studies suggests: that selective attention as discussed in the previous section can either

result in selective perception (concept of gating or *early selection*) or in selective behavior (resource allocation or *late selection*). Most importantly, she argues that the choice of mechanism actually applied depends on the perceptual load. At low perceptual load, irrelevant information continues to be processed—early selection fails and late selection becomes necessary. When the perceptual load is high, irrelevant information is not processed and resource allocation is no longer needed. She cites a number of experimental studies that support these conclusions: processing of distractors ceases when the perceptual capacity is exhausted.

Interestingly, Lavie claims that distractor processing depends on perceptual capacity limits, rather than on limited information contained in the relevant stimuli. This makes the MLE model second-rank in importance: In the MLE model, limited information contained in the relevant stimuli should entail the processing of additional cues among the distractors to check for reliability of that limited information and the correctness of its interpretation. Following Lavie, this is either not possible when the perceptual load is high, or attention needs to be shifted to formerly irrelevant information.

PERCEPTUAL SALIENCE AND SALIENCE MODEL

Landragin, Bellalem, and Romary (2001) suggest that in the absence of information about the history of an interactive process, a (visual) object can be considered salient when it attracts the user’s visual attention more than the other objects. This definition of salience originally valid for the visual domain can easily be extended to what might be called *multi-modal salience*, meaning that:

- certain properties of an object attract the user’s general attention more than the other properties of that object

- certain objects attract the user's attention more than other objects in that scene.

A salience model in the game context requires a user model of perception, as well as a task model. The user model describes familiarity of the game player with the objects' properties, as attention on the properties of an object may vary with background and experience of the player. Whereas an avatar of a human being or a human speech utterance can be considered more or less equally salient to all players (because its significance to humans is embedded genetically), an acoustically trained person might focus more on the reverberation in a virtual room than a visually oriented person. The task model describes the fact that salience depends on intentionality: depending on the task a player is given, his focus will shift accordingly.

Salience also depends on the physical characteristics of the objects themselves. In the auditory domain it is known that certain noises with increased measures of properties like *sharpness* or *roughness* call the attention more than others (Zwicker & Fastl, 1999). Adding to this, salience can be due to spatial or temporal disposition of the objects in a scene.

One of the most interesting aspects of a salience model in the context of computer games is its dependency on the degree of interactivity that the game offers to the player. If the player is allowed to interact freely with the objects in a virtual scene, then it is quite easy to determine the player's focus. Obviously, the player's focus will be on the object he is currently manipulating, so there is a clear indication of where to create a higher agreement of modalities. Consequently, games with fewer interaction possibilities are less likely to provide a sense of *being there* to the player. Thus, interactivity is important for the perceived realism of games in two different ways: first, it allows the player to do something in the virtual world, and second, it allows the application to determine the player's momentary focus. This information can then be used to enhance the

audio-visual appearance of the object in focus, for instance, by making the sound (effects) related to that object more realistic in terms of acoustic details, frequency range, localization and so on.

Salience Model

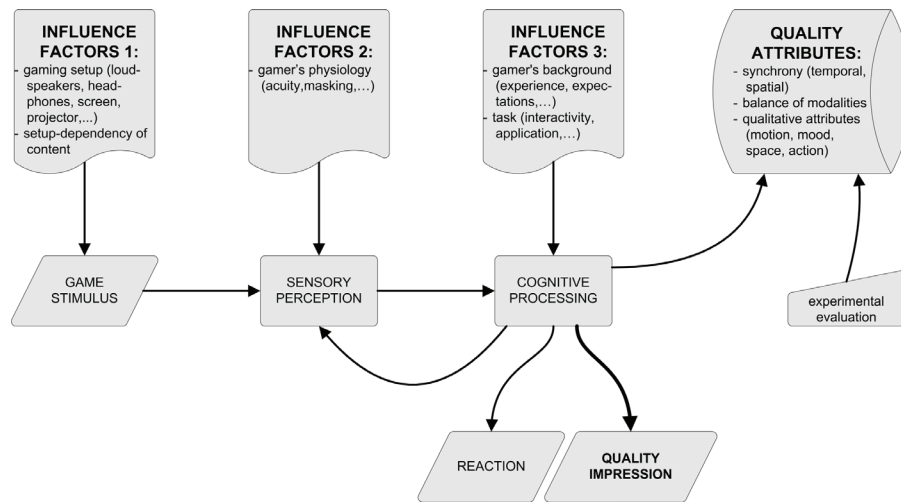
Obviously, there are situations in which the game engine has no or only limited information about the player's current focus. In these cases, it appears to be useful to have a salience model classifying the objects contained in the game scene. No such generalized multi-modal salience model exists, yet. For the rather limited scope of a gaming situation, a qualitative salience model is suggested here.

The salience model comprehends the influence factors that control the level of saliency of the objects in a game scene. Figure 2 shows how such a salience model may be structured: it is reasonable to start from the basis of human perception, the stimuli. In games, stimuli are generated by the game system itself, so they depend on a number of factors—the influence factors of level 1. These comprise the audio and visual reproduction setups, as well as input devices used for player feedback to (and control of) the system, like keyboard, joystick, mouse, or any other dedicated input device. Influence factors of level 1 are those related to the generation and control of stimuli.

The core elements of human perception are sensory perception on the one hand and cognitive processing on the other. Sensory perception can be affected by a number of influence factors of level 2. These involve the physiology of the user (acuity of vision and hearing, masking effects caused by limited resolution of the human sensors and so on), as well as other factors directly related to the physical perception of stimuli.

Cognitive processing produces a response by the player. This response can be obvious, like an immediate reaction to a stimulus, or it can be an internal response like re-distributing attention/shifting focus or just entering another turn of

Figure 2. A salience model for perceived quality in audio-visual games



the Perceptual Cycle. Obviously, the response is governed by another set of influence factors of level 3. These span the widest range of factors and are also the most difficult to quantify: experience, expectations, and socio-cultural background of the player; difficulty of task in a specific game situation; degree of interactivity; and so forth. Influence factors of level 3 are related to the processing and interpretation of the perceived stimuli.

Cognitive processing will eventually lead to a certain quality impression that is a function of all influence factors of types 1–3. This quality impression cannot be directly quantified. It needs additional processing to be uttered in the form of ratings on a quality (or quality impairment) scale, as semantic descriptors and so on. The overall quality impression is, in turn, the result of evaluating single or combined quality attributes. For example, Woszczyk, Bech, and Hansen (1995) have developed a number of attributes that are believed to be relevant for an overall audio-visual quality impression: they organize these attributes (*quality, magnitude, involvement, balance*) into 4 dimensions of perception (*motion, mood, space, action*), resulting in a 4 by 4 matrix of quality criteria. Yet, a quantification of their impact is hardly possible as of now. This is because the

individual attribute's weight not only depends on the audio-visual game scene under assessment (the stimuli), but also on the experimental methodology itself. An attribute that is explicitly asked for might be assumed to be of higher importance by a test player (we know from our experience that only important things are asked for in any kind of test). The player's attention will be directed toward the attribute under assessment, which distorts unbiased perception of the audio-visual scene as a whole. Therefore, the player's reaction in terms of quality rating can be assumed to be influenced as well.

Experimental Results

A number of experiments have shown that player interaction with an audio-visual game might have an effect on the perceived overall quality (Jumisko-Pyykkö, Reiter, and Weigel, 2007; Reiter et al., 2007; Reiter & Weitzel, 2007; Reiter & Jumisko-Pyykkö, 2007; Reiter, 2009). In these experiments, the general assumption was that by offering an attractive interactive content, or by assigning the user a challenging task, this user would become more involved and thus experience a subjectively higher overall quality. Along the

same lines, it was hypothesized that the subject's ability to differentiate between different levels of quality would decrease with an increase in difficulty of task/degree of interaction. The results show that this is not generally the case. However, when both task and main varying quality attribute were located in the same modality, such an effect could be observed.

More specifically, in the first experiment (Jumisko-Pyykkö et al., 2007; Reiter and Jumisko-Pyykkö, 2007) subjects were presented with a scenario located in a virtual sports gym. In the center of the gym, a loudspeaker was positioned that played back music/speech signals with varying amounts of reverberation (time and strength). Subjects were asked to rate the quality of reverberation under three different degrees of interaction:

1. No interaction (watch task): subjects were automatically moved on a pre-defined motion path through the virtual scenario
2. Limited interaction (watch and press button task): subjects were moved on a pre-defined motion path through the virtual scenario, but were asked to press a button whenever a certain object appeared within their field of view
3. Full interaction (navigate and collect task): subjects were asked to move freely through the scenario by using the computer mouse and to collect as many objects as possible by approaching them.

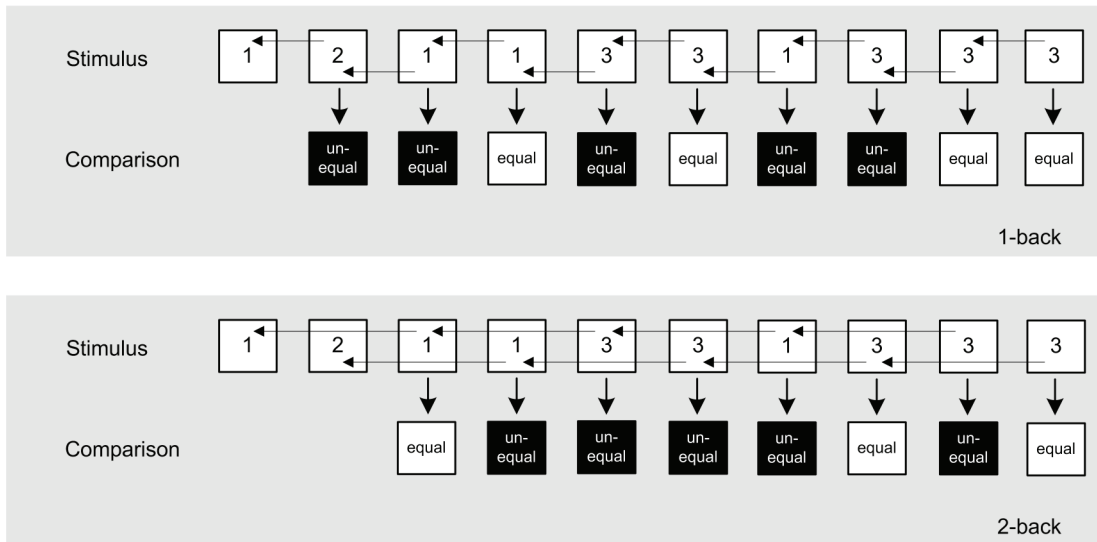
Interestingly, the ability of subjects to rate the quality of reverberation correctly did not vary with the degree of interaction/difficulty of the task (Friedman $X^2=3.3$, $df=2$, $p>0.05$, ns). Although subjects claimed to have experienced more difficulties in the interactive tasks, this did not show in the statistical analysis of the collected data. Three possible explanations were looked at. The first was that the quality differences were too obvious, that is, the steps between the different amounts of reverberation were too big. This is

possible but was not regarded as probable, given the results of informal experiments with a similar variation in reverberation. The second, was that the tasks (pressing a button, and navigating/collecting objects) were not demanding enough and that it was too easy for subjects to dedicate part of their attention towards the quality-rating task. This was contradicted by the claims of the subjects themselves: a large majority claimed to have been distracted by the navigation task. The third possible explanation was subsequently looked at in further experiments: The additional cognitive load (pressing a button, navigating while collecting objects) was located in the visual and haptic domains, whereas the quality differences to be rated were located in the auditory domain.

In a second round of experiments (compare Reiter et al., 2007; Reiter, 2009), both the additional cognitive load and the quality variations were located in the auditory domain. A virtual room (replica of the entrance hall of a large university building) was equipped with a virtual loudspeaker in the center, and subjects were asked to navigate freely through the room using a computer mouse. The loudspeaker played back a randomized sequence of numbers from 1 to 4 read out loud. The reverberation time of the room acoustic simulation could be adjusted between 1.0s and 3.0s in 0.5s steps, with 2.0s considered the "reference" reverberation time. In the experiment, the reverberation time was changed from reference to another value at a single random point in time during a transition time frame beginning 5 seconds after the start and ending 5 seconds before the end of each 30 second trial. A modified Degradation Category Rating scale according to Recommendation ITU-T P.911 (1998) was used, consisting of 5 levels (much shorter, shorter, equal, longer, much longer), to have subjects compare the test reverberation time with the reference reverberation time.

The additional cognitive load consisted of a so-called *n*-back working memory task, similar to what has been introduced by Kirchner (1958). Here, subjects were asked to semantically compare

Figure 3. Presented stimuli and correct answers (“Comparison”) for 1-back and 2-back continuous-matching-tasks



the current stimulus (the current number) with the one presented n steps back, see Figure 3. In the experiment, n was varied between 0 (no additional load) and 2 (high additional load).

The hypothesis was, again, that with increasing difficulty of the task, subjects would commit more errors in correctly rating the reverberation time as a measure of perceived quality. Here, for the statistical analysis, the rating errors were re-structured according to flaw size, such that each 0.5s deviation would result in one error point. The subsequent analysis was performed on error points. A complete description of the experiment can be found in (Reiter, 2009, pp. 203-212).

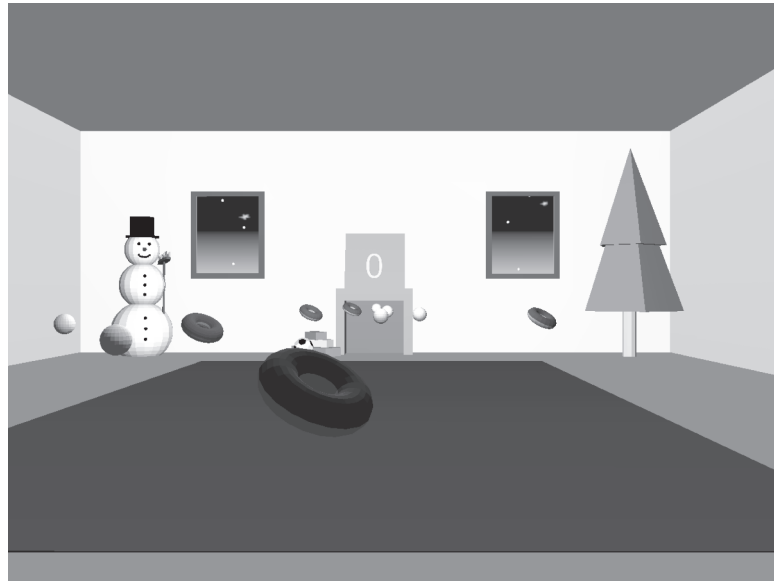
A comparison of the error rates for “navigation only” with “navigation with 2-back task” resulted in a highly significant difference ($T=20, p \leq 0.01$). Comparing these results to the first experiment described above, it becomes apparent that inner-modal influence of task is significantly greater than cross-modal influence. This might indicate that humans perform a pre-processing of stimuli that—depending on modality—takes place in separate areas of the brain. Thus, in situations where stimuli that belong to different modali-

ties have to be processed at the same time, we are better able to parallelize and distribute the processing accordingly. This is also suggested by the common theories of capacity limits in human attention, see above.

Game Example

In a third experiment (Reiter & Weitzel, 2007), inspired by Zielinski, Rumsey, Bech, de Bruyn, and Kassier (2003) and Kassier, Zielinski, and Rumsey (2003), it has been shown that cross-modal influence of interaction is very well possible when stimuli and interaction/task are carefully balanced. For this, a simplified Space Invaders-like arcade game has been created, in which two different types of objects (donuts and snowballs) moved through a virtual room. Motion of objects was straight towards the baseline, on which the player could move left and right. Players were instructed to collect as many donuts as possible and to avoid collisions with snowballs. Each collected donut resulted in an increase of the player’s score whereas a collision with a snowball decreased the score. The current game score was displayed

Figure 4. Grey-scale screenshot of the game scenario



on the screen near the chimney, which served as the source of the flying objects. Figure 4 shows a screenshot of the game scenario.

A typical background music track for an arcade game was chosen for the game. For the experiment, each subject carried out a passive and an active session. The active session involved playing the computer game and evaluating the sound quality of the game music. This session was designed to cause a division of attention between evaluating the audio quality and reaching a high score. In the passive session, subjects were asked to evaluate the audio quality while a game demo was presented. Here, the attention of the subjects was assumed to be directed to the audio quality exclusively.

In both sessions, active or passive, either the original (20kHz) game music, or a low-pass filtered version with cut-off frequencies at $f_c = 11\text{kHz}$, 12kHz or 13kHz was played. This was complemented by an anchor with $f_c = 4\text{kHz}$. Thus a total of 5, 3-test items, 1 anchor item, and 1 reference item (corresponding to the original full-range signal) were presented to the players in the experiment. After each round of the game,

players were asked to rate the perceived tonal quality degradation using the standardized ITU-T P.911 (Recommendation ITU-T P.911, 1999), 5-level impairment scale.

A total of 32 subjects participated in the experiment. Seven players were female and 25 were male (age $M = 25.7$, $SD = 5.36$). Regarding their listening experience, 20 subjects were considered initiated assessors and 12 classified as naive assessors. The group of initiated assessors had already gained abilities and knowledge in rating audio quality in preceding unimodal and bimodal subjective assessments. All participants reported normal hearing and normal or corrected to normal visual acuity.

A Wilcoxon T test showed that the quality ratings of the active session varied significantly from the ratings of the passive session for cut-off frequencies up to 12kHz . A significant decrease in rating correctness was shown for the active session in comparison to the passive session for the anchor item ($T = 37$, $p \leq 0.01$), the cut-off frequency $f_c = 11\text{kHz}$ ($T = 452.50$, $p \leq 0.01$), and the cut-off frequency $f_c = 12\text{kHz}$ ($T = 812$, $p \leq 0.01$). For the cut-off frequency of 13kHz and the

reference item, no significant differences could be found ($T = 630.50$ and $T = 75$, resp., $p > 0.05$, ns).

The data analysis showed that the ratings of the tonal quality degradations in the active session differed significantly from those in the passive session. The low-pass filtering in the active session was rated as being less perceptible compared to the passive session, for which active players turned into passive viewers. More generally speaking, the experiment shows that an influence of interaction performed in one modality (visual-haptic) upon the perception of quality in another modality (in this case, auditive) is possible. Thus, cross-modal influences are possible.

In order for a cross-modal influence to exist, the characteristics of stimuli and interaction/task must be carefully balanced. At this time, it is not possible to determine or quantify that balance a priori. However, some of the influence factors that contribute to this balance have been identified in the salience model in Figure 2 above. These influence factors need to be quantified and this is a task for the future.

SUMMARY AND CONCLUSION

This chapter has reviewed some of the most important issues of perceived quality of audio in computer games. The main conclusion is that audio quality in games, as perceived by a game player, is not independent of other factors (apart from sound quality itself). Because games usually provide information and feedback to the player in more than the auditory modality, it is necessary also to take into account other modalities when judging the impact and quality of audio. A rating of audio quality alone, without the gameplay context, is not meaningful.

The physical mechanisms of human auditory and visual perception are well understood. Cross-modal interaction between the two domains, that is, perceptual processing in the human brain, needs further research, before it is possible to model such

processes. Still, whether it is possible to come up with a generalized model of cross-modal perceptual processing at all is highly questionable. It is assumed by many that its complexity exceeds by far the possibilities for designing a suitable model. Yet, it seems feasible to aim at perceptual models that are valid for certain perceptual scenarios only. A specific game-playing scenario can be one of these, as factors like setup (computer screen, loudspeakers/headphones, input devices) and task are of rather small variance across users, given a certain use case. This has been demonstrated in the game example above. A salience model as described in this contribution could therefore serve as a starting point for the exploitation of salience effects.

Saliency is closely related to distribution of attention and perceptual capacity limits. The experimental results summarized in this chapter indicate that effects of capacity limits are more dominant inner-modally than cross-modally. At the same time, capacity limits seem to be more predictable inner-modally than cross-modally.

Unless we have better models of the perceptual processing underlying the generation of a subjective quality impression, it will be difficult to predict the perceived quality of audio in a multi-modal context in general, or in a game context as discussed here. Nevertheless, both the experiments described, and the literature and effects reviewed here, suggest that there is potential for exploitation of such perceptual constraints.

Future research should therefore concentrate on methodologies for the subjective evaluation of audio-visual quality, or multi-modal quality in general. Only a few recommendations exist for performing audio-visual experiments and the impact of interactivity—as naturally given in any gameplay—on the perceived quality is, until now, simply not considered at all. Once proper recommendations exist, it will be much easier to compare and validate experimental results, thus paving the way for a quantification of the salience model described in this chapter.

REFERENCES

- Alais, D., & Blake, R. (1999). Neural strength of visual attention gauged by motion adaptation. *Nature Neuroscience*, 2(11), 1015–1018. doi:10.1038/14814
- Battaglia, P. W., Jacobs, R. A., & Aslin, R. N. (2003). Bayesian integration of visual and auditory signals for spatial localization. *Journal of the Optical Society of America*, 20(7), 1391–1397. doi:10.1364/JOSAA.20.001391
- Beerends, J. G., & De Caluwe, F. E. (1999). The influence of video quality on perceived audio quality and vice versa. *Journal of the Audio Engineering Society. Audio Engineering Society*, 47(5), 355–362.
- Blauert, J. (2001). *Spatial hearing: The psychophysics of human sound localization* (3rd ed.). Cambridge, MA: MIT Press.
- Braasch, J. (2005). Modelling of binaural hearing. In Blauert, J. (Ed.), *Communication acoustics* (pp. 75–108). Berlin: Springer Verlag. doi:10.1007/3-540-27437-5_4
- Brodmann, K. (1909). *Vergleichende Lokalisationslehre der Grosshirnrinde in ihren Prinzipien dargestellt auf Grund des Zellenbaues*. Leipzig, Germany: Johann Ambrosius Barth Verlag.
- Eijkman, E., & Vendrik, J. H. (1965). Can a sensory system be specified by its internal noise? *The Journal of the Acoustical Society of America*, 37, 1102–1109. doi:10.1121/1.1909530
- Ellis, S. R. (1996). Presence of mind... A reaction to Thomas Sheridan's "Musing on telepresence." *Presence (Cambridge, Mass.)*, 5, 247–259.
- Farnell, A. (2011). Behaviour, structure and causality in procedural audio. In Grimshaw, M. (Ed.), *Game sound technology and player interaction: Concepts and developments*. Hershey, PA: IGI Global.
- Farris, J. S. (2003). *The human interaction cycle: A proposed and tested framework of perception, cognition, and action on the web*. Unpublished doctoral dissertation. Kansas State University, USA.
- Gardner, W. G. (1992, November). A realtime multichannel room simulator. Paper presented at the 124th meeting of the Acoustical Society of America.
- Gescheider, G. A., Sager, L. C., & Ruffolo, L. J. (1975). Simultaneous auditory and tactile information processing. *Perception & Psychophysics*, 18, 209–216.
- Goldstein, E. B. (2002). *Wahrnehmungspsychologie* (2nd ed.). Berlin: Spektrum Akadem. Verlag.
- Howard, I. P. (1982). *Human visual orientation*. New York: Wiley.
- Jot, J. M., & Chaigne, A. (1991). Digital delay networks for designing artificial reverberators. Paper presented at the AES 90th Convention. Preprint 3030.
- Jumisko-Pyykkö, S., Reiter, U., & Weigel, C. (2007). Produced quality is not perceived quality—A qualitative approach to overall audiovisual quality. In *Proceedings of the 3DTV Conference*.
- Kalman, R. E., & Bucy, R. S. (1961). New results in linear filtering and prediction problems. *Journal of Basic Engineering*, 83, 95–108.
- Kassier, R., Zielinski, S., & Rumsey, F. (2003). Computer games and multichannel audio quality part 2—Evaluation of time-variant audio degradation under divided and undivided attention. *AES 115th Convention*. Preprint 5856.
- Kirchner, W. K. (1958). Age differences in short-term retention of rapidly changing information. *Journal of Experimental Psychology*, 55(4), 352–358. doi:10.1037/h0043688

- Klein, D. J., König, P., & Körding, K. P. (2003). Sparse spectrotemporal coding of sounds. *EURASIP Journal on Applied Signal Processing*, 7, 659–667. doi:10.1155/S1110865703303051
- Landragin, F., Bellalem, N., & Romary, L. (2001). Visual salience and perceptual grouping in multimodal interactivity. In *Proceedings of International Workshop on Information Presentation and Natural Multimodal Dialogue IPNMD*.
- Larsson, P., Västfjäll, D., & Kleiner, M. (2003). On the quality of experience: A multi-modal approach to perceptual ego-motion and sensed presence in virtual environments. In *Proceedings of First ISCA ITRW on Auditory Quality of Systems AQS-2003*, 97-100.
- Lavie, N. (2001). Capacity limits in selective attention: Behavioral evidence and implications for neural activity. In Braun, J., & Koch, C. (Eds.), *Visual attention and cortical circuits* (pp. 49–60). Cambridge, MA: MIT Press.
- Lee, K. M., Jeong, E. J., Park, N., & Ryu, S. (2007). Effects of networked interactivity in educational games: Mediating effects of social presence. In *Proceedings of PRESENCE2007, 10th Annual International Workshop on Presence*, 179-186.
- Lee, K. M., Jin, S. A., Park, N., & Kang, S. (2005). Effects of narrative on feelings of presence in computer/video games. In *Proceedings of the Annual Conference of the International Communication Association (ICA)*.
- Lombard, M., & Ditton, Th. (1997). At the heart of it all: The concept of presence. *Journal of Computer-Mediated Communication*, 3.
- MacKenzie, I. S., & Ware, C. (1993). Lag as a determinant of human performance in interactive systems. In *Proceedings of the ACM Conference on Human Factors in Computing Systems – INTERCHI'93*, 488-493.
- Meehan, M., Razzaque, S., Whitton, M. C., & Brooks, F. P., Jr. (2003). Effect of latency on presence in stressful virtual environments. In *Proceedings of IEEE Virtual Reality*, 141-148.
- Nordahl, R. (2005). Self-induced footsteps sounds in virtual reality: Latency, recognition, quality and presence. In *Proceedings of PRESENCE 2005, 8th Annual International Workshop on Presence*, 353-354.
- Pashler, H. E. (1999). *The psychology of attention*. Cambridge, MA: MIT Press.
- Pulkki, V. (2001). *Spatial sound generation and perception by amplitude panning techniques*. Unpublished doctoral dissertation. Helsinki University of Technology, Finland.
- Recommendation ITU-TP.911. (1998/1999). *Subjective audiovisual quality assessment methods for multimedia applications*. Geneva: International Telecommunication Union.
- Reiter, U. (2009). *Bimodal audiovisual perception in interactive application systems of moderate complexity*. Unpublished doctoral dissertation. TU Ilmenau, Germany.
- Reiter, U., & Jumisko-Pyykkö, S. (2007). Watch, press and catch—Impact of divided attention on requirements of audiovisual quality. In Jacko, J. (Ed.), *Human-Computer Interaction, Part III, HCI2007* (pp. 943–952). Berlin: Springer Verlag.
- Reiter, U., & Weitzel, M. (2007). Influence of interaction on perceived quality in audiovisual applications: Evaluation of cross-modal influence. In *Proceedings of 13th International Conference on Auditory Displays (ICAD)*, 380-385.
- Reiter, U., Weitzel, M., & Cao, S. (2007). Influence of interaction on perceived quality in audiovisual applications: Subjective assessment with n-back working memory task. In *Proceedings of AES 30th International Conference*.

Schroeder, M. R. (1962). Natural sounding artificial reverberation. *Journal of the Audio Engineering Society, Audio Engineering Society, 10*(3), 219–223.

Schroeder, M. R. (1970). Digital simulation of sound transmission in reverberant spaces (part 1). *The Journal of the Acoustical Society of America, 47*(2), 424–431. doi:10.1121/1.1911541

Shams, L., Kamitani, Y., & Shimojo, S. (2000). What you see is what you hear. *Nature, 408*, 788. doi:10.1038/35048669

Shams, L., Kamitani, Y., & Shimojo, S. (2002). Visual illusion induced by sound. *Brain Research. Cognitive Brain Research, 14*, 147–152. doi:10.1016/S0926-6410(02)00069-1

Sheridan, T. B. (1994). Further Musings on the Psychophysics of Presence. *Presence (Cambridge, Mass.), 5*, 241–246.

Shiffrin, R. M., & Grantham, D. W. (1974). Can attention be allocated to sensory modalities? *Perception & Psychophysics, 15*, 460–474.

Singer, W., Engel, A. K., Kreiter, A. K., Munk, M. H. J., Neuenschwander, S., & Roelfsema, P. R. (1997). Neuronal assemblies: necessity, signature and detectability. *Trends in Cognitive Sciences, 1*(7), 252–261. doi:10.1016/S1364-6613(97)01079-6

Spence, C., Nicholls, M. E. R., & Driver, J. (2001). The cost of expecting events in the wrong sensory modality. *Perception & Psychophysics, 63*(2), 330–336.

Steuer, J. (1992). Defining virtual reality: Dimensions determining telepresence. *The Journal of Communication, 42*(4), 73–93. doi:10.1111/j.1460-2466.1992.tb00812.x

Tulving, E., & Lindsay, P. H. (1967). Identification of simultaneously presented simple visual and auditory stimuli. *Acta Psychologica, 27*, 101–109. doi:10.1016/0001-6918(67)90050-9

Wenzel, E. M. (1998). The impact of system latency on dynamic performance in virtual acoustic environments. In *Proceedings of the 15th International Congress on Acoustics and 135th Meeting of the Acoustical Society of America*, 2405-2406.

Wenzel, E. M. (1999). Effect of increasing system latency on localization of virtual sounds. In *Proceedings of the AES 16th International Conference on Spatial Sound Reproduction*, 42-50.

Wenzel, E. M. (2001). Effect of increasing system latency on localization of virtual sounds with short and long duration. In *Proceedings of 7th International Conference on Auditory Displays (ICAD)*. 185-190.

Woszczyk, W., Bech, S., & Hansen, V. (1995). Interactions between audio-visual factors in a home theater system: Definition of subjective attributes. *AES 99th Convention*. Preprint 4133.

Zielinski, S., Rumsey, F., Bech, S., de Bruyn, B., & Kassier, R. (2003). Computer games and multichannel audio quality—The effect of division of attention between auditory and visual modalities. In *Proceedings of the AES 24th International Conference on Multichannel Audio*, 85-93.

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

KEY TERMS AND DEFINITIONS

Binaural: Literally means “having or relating to two ears”. Binaural hearing, along with frequency cues, lets humans determine the direction of incidence of sounds.

Brodman Areas: 52 different regions of the cortex, defined on the basis of the organization of cells. Named after Korbinian Brodmann’s maps of cortical areas in humans, published 1909.

CognitivE Load: A term describing the load on working memory during instruction (problem solving, thinking, reasoning).

Dorsal Stream: Also known as the parietal stream, the “where” stream, or the “how” stream, proposed to be involved in the guidance of actions and recognizing where objects are in space.

Fitts’ Law: A model of the human movement in human-computer interaction and ergonomics which predicts that the time required to rapidly move to a target area is a function of the distance to and the size of the target.

Localization: The ability to detect the direction of incidence of a sound.

Monaural: Literally means “having or relating to one ear”.

Multi-Modal: More than one perceptual modality involved, usually the auditory and the visual domain, sometimes also including haptics.

Perceptual Cycle: A model describing human perception as a cyclic setup of schemata, perceptual exploration, and stimulus environment which influence each other in a continuously updated process, first introduced by US psychologist Ulric Neisser.

Presence: The feeling of being present in an artificial environment, for example a game scenario in a jungle.

Quality of Experience: The overall acceptability of an application or service, as perceived subjectively by the end-user. Quality of Experience includes the complete end-to-end system effects (client, terminal, network, services infrastructure and so on). Overall acceptability may be influenced by user expectations and context.

Salience: The state or quality of an item that stands out relative to neighboring items.

Schema: Previous knowledge, something we already understand or are familiar with.

Single-Cell Recording: A technique used in brain research to observe changes in voltage or current in a neuron, thus measuring a neuron’s activity.

Space Invaders: An arcade video game designed by Tomohiro Nishikado, released in 1978, with the aim of defeating waves of aliens with a cannon, earning as many points as possible.

Ventral Stream: Also known as the “what” stream, associated with object recognition and form representation.