Bimodal Audiovisual Perception in Interactive Application Systems of Moderate Complexity

Dissertation

zur Erlangung des akademischen Grades
Doktoringenieur (Dr.-Ing.)

vorgelegt der Fakultät für Elektrotechnik und Informationstechnik
der Technischen Universität Ilmenau

von Dipl.-Ing. Ulrich Reiter
geboren am 1. Juli 1970 in Gelsenkirchen

Gutachter:
1. Univ.-Prof. Dr.-Ing. Karlheinz Brandenburg
   Technische Universität Ilmenau
2. Dr. Durand R. Begault Ph.D.
   NASA Ames Research Center, Mountain View, CA, USA
3. Dr.-Ing. Thomas Sporer
   Fraunhofer Institut für Digitale Medientechnik, Ilmenau

Tag der Einreichung: 17. Dezember 2007
Tag der wissenschaftlichen Aussprache: 21. August 2009

urn:nbn:de:gbv:ilm1-2009000273

Zusammenfassung

Die vorliegende Dissertation beschäftigt sich mit Aspekten der Qualitätswahrnehmung von interaktiven audiovisuellen Anwendungssystemen moderater Komplexität, wie sie z.B. durch den MPEG-4 Standard definiert sind. Die Frage, welche Faktoren Einfluss auf die wahrgenommene Qualität von audiovisuellen Anwendungssystemen haben ist entscheidend dafür, wie die nur begrenzt zur Verfügung stehende Rechenleistung für die Echtzeit-Simulation von 3D Szenen und deren Darbietung sinnvoll verteilt werden soll. Während Qualitätsfaktoren für unimodale auditive als auch visuelle Stimuli seit langem bekannt sind und entsprechende Modelle existieren, müssen diese für die bimodale audiovisuelle Wahrnehmung noch hergeleitet werden. Dabei ist bekannt, dass eine Wechselwirkung zwischen auditiver und visueller Qualität besteht, nicht jedoch, wie die Mechanismen menschlicher audiovisueller Wahrnehmung genau arbeiten. Es wird auch angenommen, dass der Faktor Interaktion einen wesentlichen Einfluss auf wahrgenommene Qualität hat.

Das Ziel dieser Arbeit war, ein System für die zeitsparende und weitgehend automatisierte Durchführung von subjektiven audiovisuellen Wahrnehmungstests im gegebenen Kontext zu erstellen und es für einige exemplarische Experimente einzusetzen, welche erste Aussagen über audiovisuelle Wechselwirkungen und den Einfluss von Interaktion auf die Wahrnehmung erlauben sollten. Demzufolge gliederte sich die Arbeit in drei Aufgabenbereiche: die Erstellung eines geeigneten Testsystems auf der Grundlage eines vorhandenen, jedoch in seiner Audiofunktionalität noch eingeschränkten MPEG-4 Players, das Sicherstellen von Vergleichbarkeit und Wiederholbarkeit von audiovisuellen Wahrnehmungstests durch definierte Testmethoden und -bedingungen, und die eigentliche Durchführung der aufeinander abgestimmten Experimente mit anschließender Auswertung und Interpretation der gewonnenen Daten. Dazu wurde eine objektbasierte, modulare Audio-Engine mitentworfen und -implementiert, welche basierend auf den Möglichkeiten der MPEG-4 Szenenbeschreibung alle Fähigkeiten zur Echtzeitberechnung von Raumakustik bietet. Innerhalb des entwickelten Testsystems kommuniziert der MPEG-4 Player mit einem hardwaregestützten Benutzerinterface zur Eingabe der Qualitätsbewertungen durch die Testpersonen. Sämtliche relevanten Ereignisse, die während einer Testsession auftreten, können mit Hilfe eines Logging-Tools aufgezeichnet und für die weitere Datenanalyse mit Statistikprogrammen exportiert werden.

Eine Analyse der existierenden Testmethoden und -empfehlungen für unimodale Wahrnehmungstests sollte zeigen, ob deren Übertragung auf den audiovisuellen Fall möglich ist. Dabei wurde deutlich, dass bedingt durch die fehlende Kenntnis der zugrundeliegenden Wahrnehmungsprozesse zunächst eine Unterteilung nach den Zielen der durchgeführten Experimente sinnvoll erscheint. Weiterhin konnten Einflussfaktoren identifiziert werden, die die bimodale Wahrnehmung im gegebenen Kontext steuern.

Bei der Durchführung der Wahrnehmungsexperimente wurde die Funktionsfähigkeit des erstellten Testsystems verifiziert. Darüber hinaus ergaben sich erste Anhaltspunkte für den Einfluss von Interaktion auf die wahrgenommene Gesamtqualität: Interaktion in der auditiven Modalität verringert die Fähigkeit, Audioqualität korrekt beurteilen zu können, während visuell gestützte Interaktion (cross-modal) diesen Effekt nicht zwingend generiert.

Abstract

The dissertation at hand deals with aspects of quality perception of interactive audiovisual application systems of moderate complexity as e.g. defined in the MPEG-4 standard. Because in these systems the available computing power is limited, it is decisive to know which factors influence the perceived quality. Only then can the available computing power be distributed in the most effective and efficient way for the simulation and display of audiovisual 3D scenes. Whereas quality factors for the unimodal auditory and visual stimuli are well known and respective models of perception have been successfully devised based on this knowledge, this is not true for bimodal audiovisual perception. For the latter, it is only known that some kind of interdependency between auditory and visual perception does exist. The exact mechanisms of human audiovisual perception have not been described. It is assumed that interaction with an application or scene has a major influence upon the perceived overall quality.

The goal of this work was to devise a system capable of performing subjective audiovisual assessments in the given context in a largely automated way. By applying the system, first evidence regarding audiovisual interdependency and influence of interaction upon perception should be collected. Therefore this work was composed of three fields of activities: the creation of a test bench based on the available but (regarding the audio functionality) somewhat restricted MPEG-4 player, the preoccupation with methods and framework requirements that ensure comparability and reproducibility of audiovisual assessments and results, and the performance of a series of coordinated experiments including the analysis and interpretation of the collected data. An object-based modular audio rendering engine was co-designed and -implemented which allows to perform simple room-acoustic simulations based on the MPEG-4 scene description paradigm in real-time. Apart from the MPEG-4 player, the test bench consists of a haptic Input Device used by test subjects to enter their quality ratings and a logging tool that allows to journalize all relevant events during an assessment session. The collected data can be exported comfortably for further analysis using appropriate statistic tools.

A thorough analysis of the well established test methods and recommendations for unimodal subjective assessments was performed to find out whether a transfer to the audiovisual bimodal case is easily possible. It became evident that - due to the limited knowledge about the underlying perceptual processes - a novel categorization of experiments according to their goals could be helpful to organize the research in the field. Furthermore, a number of influencing factors could be identified that exercise control over bimodal perception in the given context.

By performing the perceptual experiments using the devised system, its functionality and ease of use was verified. Apart from that, some first indications for the role of interaction in perceived overall quality have been collected: interaction in the auditory modality reduces a human's ability of correctly rating the audio quality, whereas visually based (cross-modal) interaction does not necessarily generate this effect.

Danksagung

Die vorliegende Arbeit entstand im Rahmen meiner Tätigkeit als wissenschaftlicher Mitarbeiter am Institut für Medientechnik der Technischen Universität Ilmenau. Mein Dank gilt den Mitarbeitern und Studierenden des Fachgebietes Elektronische Medientechnik, welche mir die Erstellung der vorliegenden Dissertation ermöglichten. Mein besonderer Dank gilt meinem Doktorvater Herrn Univ.-Prof. Dr.-Ing. Karlheinz Brandenburg und den beiden Korreferenten, Herrn Dr. Durand R. Begault Ph.D. und Herrn Dr.-Ing. Thomas Sporer. Weiterhin danke ich meinen Kollegen Herrn Dipl.-Inf. Mathias Schwark und Frau MSc. eng. Satu Jumisko-Pyykkö für die angeregten und kritischen Diskussionen bei der Erstellung dieser Arbeit.

Meiner Frau Katja danke ich für ihre Unterstützung und Geduld während der Erstellung dieser Dissertation. Ohne sie wäre diese Arbeit nicht möglich gewesen.

Ulrich Reiter

Contents

Part I

Introduction

1 Introduction

1.1 Background

After years of continuously growing simulation depth of computer based interactive application systems, there is finally a tendency visible which promises a shift of paradigm. Apparently, the next improvements in these kinds of systems can only be achieved by following a new approach in simulation techniques. Computing power available on a standard personal computer almost always lags behind the necessities and expectations of a user. Formulated in Moore's Law[1], we have seen a continuous growth in computing power capacities from generation to generation. Still we are developing applications which only run on the fastest computers available. Therefore, computing power still seems to be the limiting factor for the further advance in the area of interactive application systems, even if these are only of moderate complexity.

Most of today's interactive application systems aim at simulating an accurate representation of the real world by focusing on the (arguably) most important human sense, vision. Auditory stimuli are used in these systems to enhance the overall impression of realism. Still, the stimuli of the two modalities are rendered and presented mostly independently from the other modality. The level of detail in the respective (visual or auditory) simulation is kept as high as possible with regard to computing power available, independently from the level of detail in the other modality and independently from the user's current focus.

This approach apparently contradicts our real world experiences. In the real world, we experience a simultaneous stimulation of all our senses, providing us with a redundancy of information about objects surrounding us. We deliberately or not choose the object or event which is of most interest to us, and perceive its characteristics multi-modally. Yet, not all stimuli that are perceived by our senses are equally important in the generation of an overall impression. Depending on a number of factors (experience, context, mood, to name a few) our perceptual processor subconsciously selects those stimuli that are important - and downgrades or completely discards others of less importance.

Unfortunately, the mechanisms underlying the selection and weighing process of multi-modal percepts are not fully understood. In fact, very little is known about them that could be used universally, in spite of severe efforts of scientists working in the fields of cognition, neurophysiology, psychology, communication sciences, and so on. Yet, if we confine ourselves to certain boundary conditions and drop the pretension of universally explaining the perceptual processes involved, we can find that in certain situations certain patterns of cognition are used. Literature describes a plethora of experiments related to these multi-modal perceptual phenomena.

If it was possible to provide a salience model describing what the importance of all singular percepts in a multi-modal perceptual situation is, then we could design audiovisual

[1] Moore's Law is the empirical observation made in 1965 by Gordon E. Moore, the later co-founder of Intel, that the number of transistors on an integrated circuit (IC) for minimum component cost doubles every 24 months [moo65]. The number of transistors on an IC, e.g. a central processing unit (CPU), directly correlates with its computational power.

application systems in such a way that they offered an optimum quality / cost ratio. From past experiments it is known that overall perceived audiovisual quality is not directly related to simulation depth. By providing only full simulation depth for those percepts that actually make it through the weighing process described above, and consequently by simply estimating stimuli of lesser importance (instead of simulating them in-depth), the available computing power could be distributed optimally and perceived quality could be increased. This paradigm offers a whole new scope for research with respect to the question of how to optimally use the available computing power – or, in other words, of how to achieve the maximum possible perceived quality at the minimum simulation cost. Such an approach could be referred to as "multi-modal perceptual coding".

This approach provides a multitude of benefits. Not only could it reduce computational costs in audiovisual applications such as teleconferencing systems, interactive computer games and simulation applications, augmented and virtual reality systems, interactive broadcast, and real time rendering and display of 3D audiovisual content in general. It could also prove beneficial for a wider acceptance of object based content schemes like MPEG-4, in which the aggregation of auditory and visual characteristics of an object is immanent. Because of this feature, object based description schemes lend themselves very well for the reproduction of interactive multi-modal content and for multi-modal perceptual coding in general.

In order to arrive at a salience model, being it the basis of any multi-modal perceptual coding scheme, a number of pre-requisites have to be fulfilled. Human perceptual processing has to be evaluated in a systematic way, which calls for an evaluation system capable of rendering and displaying interactive audiovisual content in real time, such as the final application making use of the model would do. As there is no means of directly measuring perceived quality, subjective assessments need to be performed. In order to make the results reliable, a considerable number of test subjects need to participate. Unfortunately, this is both cost-intensive and time-consuming. Hence, this is only possible if the process of testing can be largely automated and the collection of data is fast and reliable.

Whenever conclusions are drawn on the basis of data originating from subjective assessments, these assessments need to be repeatable and verifiable. In the fields of audio quality as well as video quality assessment, recommendations and standards exist that guide through the process of designing, performing and evaluating subjective assessments. These recommendations are not necessarily applicable to quality evaluations of bimodal audiovisual stimuli. On the contrary, it is well known that stimuli perceived in one modality (e.g. auditory) can influence the perceived quality of the other (in this case: visual). This makes it necessary to develop recommendations that consider these potential cross-modal effects.

In audiovisual applications ideally both modalities are stimulated in accordance (time-wise, location-wise, ...). Both visual and auditory receptors as well as the underlying processing system(s) are involved. The multi-modal nature of test subjects' attention entails new challenges for the methodologies to be used in subjective assessments: How to get quality feedback from test subjects without interrupting the multi-modal perceptual flow? How to determine the actual focus of attention in order to identify the most captivating (or salient) attributes? Can this be done at all? What role does interactivity play in the perception of quality?

Once an experiment has been performed and subjective data has been collected, this data needs to be analyzed. The analysis of data collected in such audiovisual subjective

assessments is fairly simple, although a few pitfalls exist. The main concern here is that such data rarely satisfies all of the ANOVA (ANalysis Of VAriance) criteria, thus making its evaluation dependent on other methods of statistical analysis that can be confusing at times. Again, a clear set of recommendations would be helpful for the case.

Finally, the question to which degree these results can be generalized needs to be addressed. This is perhaps the most challenging task in the context, as laboratory situations are not necessarily comparable to real life experiences. Perceptual situations in the lab are always controlled. Most experiments are carefully balanced to keep out unwanted influences from external stimuli, from other modalities, or in general from varying constraints. It is therefore expedient to categorize experiments according to their aims. Often, it is neither necessary nor reasonable to come up with generally valid conclusions on the human perceptual processes. E.g. in the case of this work it is sufficient to draw conclusions valid only in the context of interactive audiovisual application systems of moderate complexity, a field that in itself is complex enough to provide a number of significant challenges.

1.2 Overview of the Thesis

This thesis is organized into five parts. PART I contains the introduction to this work with an overview of the thesis, the contributions of the author to the field, and a list of scientific publications by the author that are related to this work.

PART II gives an overview of the related research. CHAPTER 2 on human perception first discusses visual and auditory perception separately from a physiological point of view. Subsequently, an introduction to the processing of percepts in the human brain and the known mechanisms of perception is given. Questions related to the joint processing of audiovisual stimuli are discussed here. Also, a number of known effects in human bimodal perception are presented. Finally, the question of control over the perceptual process (steering of attention) is discussed along with the presentation of models that try to describe the perceptual process on different levels of abstraction. Chapter 2 concludes with a preliminary posit on human perceptual modeling derived from the scientific literature presented.

CHAPTER 3 introduces computational room acoustics and auralization. It explains the main challenges related to real-time rendering of room acoustics and presents the three most important rendering methods used for room acoustic simulation purposes today, along with their pros and cons for real-time applications. Chapter 3 is completed with an overview of implementations of auditory as well as audiovisual Virtual Environments, the type of application closest to interactive audiovisual application systems as discussed in this work.

CHAPTER 4 brings into focus issues related to interactivity and presence in audiovisual application systems. Effects of latency, input and perceptual feedback are discussed, as well as the possible involvement of the user.

PART III describes in detail the tools that have been generated in the course of this work. After an introduction to the object-based concept and the main features of MPEG-4, CHAPTER 5 describes the I3D MPEG-4 player developed at the Institute of Media Technology, with a strong focus on the TANGA real-time audio engine. Structure, features and integration of the TANGA engine are discussed in the MPEG-4 scene context. This

includes a detailed discussion of the MPEG-4 standardized *Physical* and *Perceptual Approach*es. Implementation of acoustic obstruction and the use of distributed sound sources are shortly presented. Dynamic audio scene graph simplification is introduced and discussed as a means of a unified approach to reduce computational complexity dynamically. Chapter 5 concludes with a detailed documentation of the processor independent multi-threading capabilities implemented in the TANGA engine. This feature is truly unique among current object-based real-time audio rendering engines.

Chapter 6 discusses the remaining elements of the system created for performing subjective assessments of perceived audiovisual quality. These are the so-called Input Device, a micro-controller based hardware box used by the test subjects to control the course of the experiments, and the SALT (Subjective Assessment Logging Tool), a JAVA-based software created to log all events in the course of the assessment. The system structure, along with a description of the communication protocol based on the MIDI (Musical Instrument Digital Interface) standard is presented in detail. Also, further scene control and interaction possibilities offered by the customized (non-standardized) extension of MPEG-4 with a MIDI input sensor is discussed.

After having detailed all technical constraints in the previous chapters, Part IV mainly details the application of these tools in the practical implementation of subjective assessments. Chapter 7 presents a theoretical approach to a first step toward a salience model of human perception in interactive audiovisual applications of moderate complexity, based on the identification of the most important influence factors.

Chapter 8 presents various concepts of measurement, describes the existing standards for auditory as well as visual subjective quality assessments, and discusses the appropriateness of unimodal assessment rules for the bimodal or multi-modal case. This is followed by the presentation of a newly developed categorization of assessments which helps to evade some of the problems associated with the ITU's categorization commonly used until now. The discussion of test methods, procedures and rating scales gives some insights to the main problems with subjective assessments of perceived quality in general. Suggestions for the bimodal (audiovisual) case are presented, and the constraints given for the assessments performed in the course of this work are discussed: experimental design, test material, test room, setup, and test subjects. Finally, the most important fundamentals of statistical analysis along with a basic introduction to the terminology used in statistic analysis are presented. Chapter 8 is completed with the detailed description and analysis of six exemplary audiovisual assessments performed in the course of this work. Conclusions from these experiments are drawn in the respective sections.

Finally, Part V summarizes the work described in this thesis. Chapter 9 gives an overview of the results obtained from the assessments described in the previous chapter. These results are critically reviewed and evaluated within their context. Results that can be regarded as assured facts are identified as such. Results that need further evaluation are discussed. Advancements to the state-of-the-art associated with the work presented here are recapitulated. Conclusions for future implementations of interactive audiovisual application systems are drawn. An outlook is presented that makes suggestions for further work in the field.

1.3 Contributions of the Author

In this work, human audiovisual quality perception in interactive applications of moderate complexity has been investigated from an engineering point of view. Audiovisual quality perception in the given context is a relatively new field of research. Although the methods and standards describing and defining such (often object- and scene-based) applications have been under development since the 1990s, the technologies and techniques necessary for the creation, transmission and reproduction of interactive audiovisual content are only becoming available now. Therefore the tools necessary for systematically scrutinizing the problem had to be created in the first place.

The main contributions of this work consist in:

- An exhaustive overview of the related research on human bimodal audiovisual perception.
- Design, implementation and proof of concept of an object based system (hardware, software) for the evaluation of perceived quality in the given context by means of subjective assessments.
- Conceptual layout and extension of a modular, object based real-time audio rendering engine providing an extraordinarily high amount of flexibility and performance both within and outside of the MPEG-4 scene description framework.
- Proof of concept and implementation of MIDI as a very flexible control protocol for scene interaction in MPEG-4 applications.
- Introduction of a novel categorization of experiments related to audiovisual subjective quality, making the results of these assessments more representative for the respective category of experiment. This leads to a better comparability of experimental results.
- Identification of important factors influencing the overall perceived quality of audiovisual scenes, especially the role of interaction in different modalities.
- Proof of inner-modal influence of interaction / working memory task upon the perceived quality of audiovisual scenes in the given context. Evidence that this statement is not generally valid cross-modally.

1.4 Related Publications by the Author

- Reiter, U.: TANGA - an Interactive Object-Based Real Time Audio Engine, Proc. Audio Mostly 2007 - 2nd Conference on Interaction with Sound, Ilmenau, Germany, September 27-28, 2007, pp 104-109.
- Reiter, U. and Jumisko-Pyykkö, S.: Watch, Press and Catch - Impact of Divided Attention on Requirements of Audiovisual Quality, 12th International Conference on Human-Computer Interaction, HCI2007, Beijing, PR China, July 22-27, 2007.
- Reiter, U. and Weitzel, M.: Influence of Interaction on Perceived Quality in Audiovisual Applications: Evaluation of Cross-Modal Influence, Proc. 13th International Conference on Auditory Displays (ICAD), Montreal, Canada, June 26-29, 2007.

- Reiter, U. and Kühhirt, U.: Object-Based A/V Application Systems: IAVAS I3D Status and Overview, IEEE/ISCE'07, International Symposium on Consumer Electronics, Dallas, TX, USA, June 20-23, 2007.

- Jumisko-Pyykkö, S.; Reiter, U.; Weigel, Chr.: Produced Quality is not Perceived Quality - a Qualitative Approach to Overall Audiovisual Quality, Proceedings of the 3DTV Conference, Kos Island, Greece, May 7-9, 2007.

- Reiter, U.; Weitzel, M.: Influence of Interaction on Perceived Quality in Audio Visual Applications: Subjective Assessment with n-Back Working Memory Task, II, AES 122nd Convention, Vienna, Austria, May 5-8, 2007, Preprint 7046.

- Reiter, U. and Partzsch, A.: Multi Core / Multi Thread Processing in Object Based Real Time Audio Rendering: Approaches and Solutions for an Optimization Problem, AES 122nd Convention, Vienna, Austria, May 5-8, 2007, Preprint 7159.

- Reiter, U.; Weitzel, M.; Cao, S.: Influence of Interaction on Perceived Quality in Audio Visual Applications: Subjective Assessment with n-Back Working Memory Task, Proc. AES 30th International Conference, Saariselkä, Finland, March 15-17, 2007.

- Reiter, U.: A System for Performing Subjective Quality Assessments of Interactive A/V Content, 24. Tonmeistertagung - VDT International Convention, Leipzig, Germany, November 16-19, 2006.

- Reiter, U.: TANGA Updated - A Modular Framework for Real Time Audio Rendering of Object-Based (MPEG-4) Audio Visual Scenes, Proc. 51st IWK - Internationales Wissenschaftliches Kolloquium, Ilmenau, Germany, September 11-15, 2006.

- Reiter, U.: Subjective Assessment of the Optimum Number of Loudspeaker Channels in Audio-Visual Applications Using Large Screens, Proc. AES 28th International Conference, Pitea, Sweden, June 30 - July 2, 2006, pp 102-109.

- Reiter, U.; Partzsch, A.; Weitzel, M.: Modifications of the MPEG-4 AABIFS Perceptual Approach: Assessed for the Use with Interactive Audio-Visual Application Systems, Proc. AES 28th International Conference, Pitea, Sweden, June 30 - July 2, 2006, pp 110-117.

- Steglich, B.; Reiter, U.: Sound Source Obstruction in an Interactive 3Dimensional MPEG-4 Environment, 120th AES Convention, Paris, France, May 20-23, 2006, Preprint 6705.

- Reiter, U.; Großmann, S.; Strohmeier, S.; Exner, M.: Observations on Bimodal Audio Visual Subjective Assessments, 120th AES Convention, Paris, France, May 20-23, 2006, Preprint 6852.

- Reiter, U.: Bimodal Perception Phenomena in Interactive Audio Visual Application Systems: an Assessment Framework, ICOB2005, Workshop on Immersive Communication and Broadcast Systems, Berlin/Germany, October 2005.

- Reiter, U.: Audio Rendering System Design for an Object Oriented Audio Visual Human Perception Assessment Tool, DAFx'05, 8th International Conference on Digital Audio Effects, Madrid, Spain, September 20-22, 2005, ISBN 4-7402-318-1, pp 69-72.

- Reiter, U., Köhler, T.: Criteria for the Subjective Assessment of Bimodal Perception in Interactive AV Application Systems, IEEE/ISCE'05, International Symposium on Consumer Electronics, Macau SAR, China, June 14-16, 2005, ISBN 0-7803-8920-4, pp 186-192.

- Reiter, U., Holzhäuser, S.: An Input Device for Subjective Assessments of Bimodal Audio Visual Perception, IEEE/ISCE'05, International Symposium on Consumer Electronics, Macau SAR, China, June 14-16, 2005, ISBN 0-7803-8920-4, pp 296-300.

- Reiter, U., Schwark, M.: A plug-in based audio rendering concept for an MPEG-4 Audio subset, IEEE/ISCE'04, International Symposium on Consumer Electronics, Reading, UK, September 1-3, 2004, ISBN 0-7803-8526-8, pp 55-60.

- Dantele, A., Reiter, U.: Description of audiovisual virtual 3D scenes: MPEG-4 perceptual parameters in the auditory domain, IEEE/ISCE'04, International Symposium on Consumer Electronics, Reading, UK, September 1-3, 2004, ISBN 0-7803-8526-8, pp 87-90.

- Reiter, U., Körner, F., Kootz, M., and Rüffer, S.: A room acoustics design tool for MPEG-4 conforming scene design, IEEE/ISCE'04, International Symposium on Consumer Electronics, Reading, UK, September 1-3, 2004, ISBN 0-7803-8526-8, pp 49-54.

- Dantele, A., Reiter, U., Schwark, M.: Audiovisual Virtual Environments: Enabling Realtime Rendering of Early Reflections by Scene Graph Simplification, 116th AES Convention, Berlin, Germany, May 8-11, 2004, Preprint 6028.

- Reiter, U.: On the Need for a Salience Model for Bimodal Perception in Interactive Applications, IEEE/ISCE'03, International Symposium on Consumer Electronics, Sydney, Australia, December 3-5, 2003.

- Reiter, U.; Schuldt, M.; Dantele, A.: Determination of Sound Source Obstruction in Virtual Scenes, Proc. AES 24th International Conference on Multichannel Audio, Banff, Alberta, Canada, June 26-28, 2003, ISBN 0-937803-50-2, pp 201-206.

- Dantele, A.; Schuldt, M.; Reiter, U.: Audio Aspects when Using MPEG-4 in an Interactive Virtual 3-D Scenery, Proc. AES 24th International Conference, Banff, Alberta, Canada, June 26-28, 2003, ISBN 0-937803-50-2, pp 335-337.

- Dantele, A.; Reiter, U.; Schuldt, M.; Drumm, H.; Baum, O.: Implementation of MPEG-4 Audio Nodes in an Interactive Virtual 3D Environment, 114th AES Convention, Amsterdam, The Netherlands, March 22-25, 2003, Preprint 5820.

- Drumm, H.; Kühhirt, U.; Rittermann, M.; Reiter, U.: Application Systems for MPEG-4, IEEE/ISCE'02, International Symposium on Consumer Electronics, Erfurt, Germany, September 23-26, 2002.

Part II

Related Research

2 Human Perception

Human perception in real world situations is a multi-modal, recursive process controlled by attention. Stimuli that call our attention are processed with priority, giving them a higher weight in the overall perceptual process. Attention can be focused intentionally or subconsciously. Stimuli from different modalities usually complement each other and make the perceptual process more unequivocal.

Because of its complexity, the human perceptual process cannot easily be explained in a simple block diagram without neglecting important features. A number of models exist, but these only cover certain aspects of the process, depending on the level of abstraction at which the respective model is located.

It is easily accepted that only those stimuli that can actually be perceived by the primary receptors of sound, light, pressure, etc. will contribute to an overall impression (which is the result of any perceptual process). Therefore, the first two sections of this chapter focus on the physiology of visual and auditory receptors. Because one of the main benefits of audiovisual application systems is their ability to easily communicate spatial and temporal relations of objects, emphasis is placed on the perception of spatial attributes in audio and vision.

Auditory and visual stimuli are processed jointly. How exactly this is done and what the mechanisms of joint processing are is evaluated in section 2.3. Here, the mechanisms of information transport to and in the human brain are discussed, and the methods originally used to detect these mechanisms are shortly introduced. The human brain is described from a neurophysiological point of view in order to arrive at a simplified explanation for the mechanisms of joint audiovisual processing.

Especially literature in the field of psychology holds a considerable number of publications on perceptual effects. Although most of these effects stir the interest because they point out some of the deficiencies of the human perceptual system, their knowledge might be useful in explaining certain perceptual phenomena that also occur in audiovisual application systems and that possibly influence the perceived overall quality. Furthermore, it might be possible to take advantage from some of these effects in the design of future audiovisual application systems.

The following section 2.5 on the role of attention in the human perceptual process takes a closer look at the simultaneous perception of more than one perceptual stream. Whether there is a difference or not if these streams are located in the same modality (or distributed across different modalities) is discussed. Neisser's Perceptual Cycle is introduced as one of the more abstract models of the human perceptual process. Selective and divided attention and potential capacity limits of the perceptual system are discussed. Along with this, another model of perception originally presented by Hollier and Voelcker is described. It is more application-oriented than the Perceptual Cycle and potentially lends itself better for a use in technical applications.

The chapter concludes with a preliminary posit drawing conclusions about the limitations of current perceptual models and the difficulties to be encountered when trying to benefit from them in the implementation of audiovisual application systems.

2.1 Visual Perception

2.1.1 Physiology of Visual Perception

The human visual system seldom responds to direct stimulation from a light source. Rather, light is reflected by objects and thus transmits information about certain characteristics of the object. The reflected ray of light enters the eyeball through the cornea as depicted in fig. 2.1. After passing through the cornea and the watery aqueous humor, the photon beam enters the inner eye through the pupil, which regulates the amount of light allowed to enter. The lens focuses the light on the sensory cells of the retina. Located between the lens and the retina is a semi-colorless, gelatinous material called the vitreous humor. Some frequencies of the wave forms contained in a light beam are absorbed by this substance [mur73].

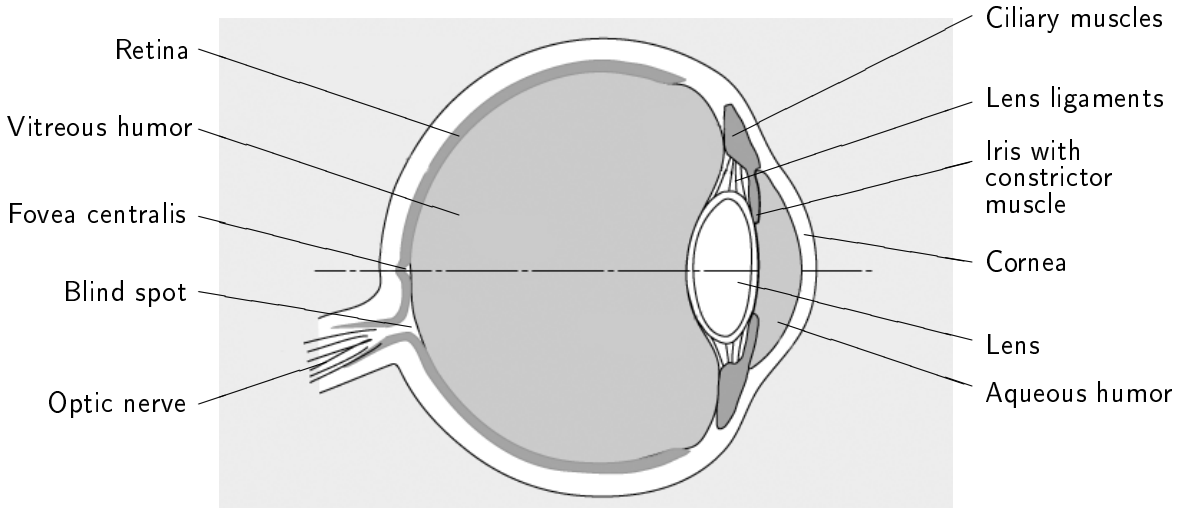


Fig. 2.1: Cross-section of the human eye, *after [w-hyp]*.

The cornea represents the strongest part of the refracting power of the eye, providing about 80% of the total eye's refracting capacity. About $9mm$ in diameter and $4mm$ thick, the crystalline lens provides perhaps 20% of the refracting power of the eye. It is made up of about 20000 thin layers of transparent material. The index of refraction ranges from about 1.406 at the center to about 1.386 in outer layers, making it a gradient index lens. It is pliable, and changes shape to accomplish accommodation for close focusing. When the eye is relaxed and the interior lens is the least rounded, the lens has its maximum focal length for distant viewing. As the muscle tension around the ring of muscle is increased and the supporting fibers are thereby loosened, the interior lens rounds out to its minimum focal length, see fig. 2.2.

The internal layer of the eyeball is made up of the nervous coat called retina. The retina covers the inner back part of the eyeball. This is where the optical image is formed by the eye's optical system. Here, a photochemical transduction occurs: nerve impulses are created and transmitted along the optic nerve to the brain for higher cortical processing, see section 2.3.3 *Visual Pathway and Cortex*. The point of departure of that optic nerve through the retina does not have any receptors, and thus produces a "blind spot". The retina consists of two different types of light-sensitive cells, rods and cones. There are about 6.5 million cones in each eyeball, most of them located in the middle of the retina, in a small dimple about $1.5mm$ in diameter called the fovea or fovea centralis. It is the

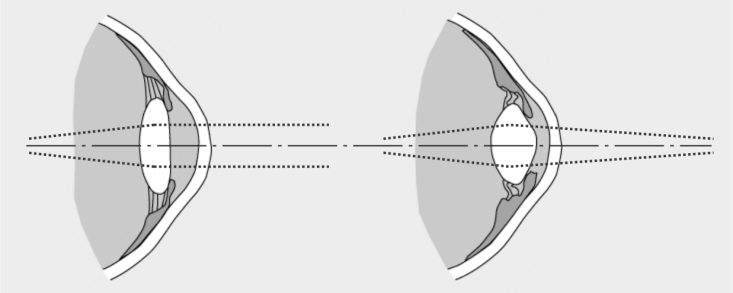


Fig. 2.2: Accommodation of the human eye. On the left, relaxed muscles result in maximum focal length for distance viewing. On the right, muscle tension loosens the supporting fibers and the lens rounds out to minimum focal length, *after [w-hyp], modified.*

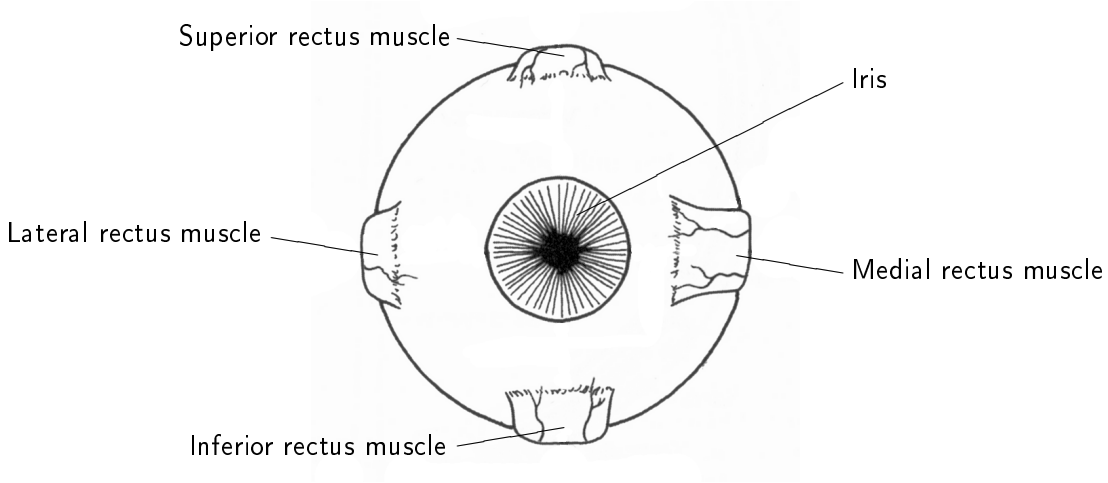


Fig. 2.3: Diagram of the right eyeball, seen from the front. Four of six muscles attached to the eyeball are visible. *After [sne98].*

center of the eye's sharpest vision and the location of most color perception, performing in bright light, but being fairly insensitive at low light levels. Located around the fovea centralis are about 120 million rods. They are mainly responsible for vision in dim light and produce images consisting of varying shades of black and white. The acuity over most of that range is poor, and the rods are multiply connected to nerve fibers, so that a single nerve fiber can be activated by any one of about a hundred rods. In contrast, cones in the fovea centralis are individually connected to nerve fibers [sne98].

The eyeball is situated in the orbital cavity, a location that protects it and provides a rigid bony origin for the six extrinsic muscles that produce ocular movement. Fig. 2.3 shows four of these muscles seen from the front. When the visual system focuses on a certain object, then the optical axes of both eyes are adjusted toward it. The sensation of tension in the muscles serves as an indicator for the distance the object is away.

2.1.2 Visual Perception of Space

The direction of a visually perceived object corresponds directly to the position of its image on the retina. But a visual stimulus also 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 sources

or cues of depth.

Monocular Cues of Depth

That depth perception is still possible with only one eye can easily be demonstrated by viewing a scene first with both eyes and then with only one. Whereas there is a definite change in the three-dimensional appearance of the scene, we are still able to judge depth with one eye in a fairly accurate way. In the following, the most important cues will be explained.

Interposition If an object partially covers another object, the blocked object is perceived to be located behind the blocking object, see fig. 2.4. This cue is a very dominant one and mostly supersedes other, conflicting cues. Yet it provides only information about the relative depth between objects and not about the actual distance to the viewer.

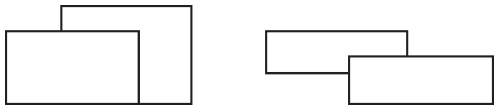


Fig. 2.4: Interposition cue: two examples for an object concealing another object.

Size If the size of an object is known, then the distance it is located at can be deduced. Fig. 2.5 shows two squares physically different in size and thus creating an apparent depth effect in which the larger square seems to be closer to the viewer than the smaller. If we want to use this cue to make a statement on the relative size of two objects as an indication of depth, then we assume that the two objects are really of the same physical size. Judgment of depth fails when this assumption does not hold true, or when familiarity with the size of the two objects is not given. Murch [mur73] describes a number of experiments performed by Gogel, Hastorf and Ittelson on the subject. Considering the results of these experiments, evidently there are two size cues of monocular depth perception: relative size and familiar size.

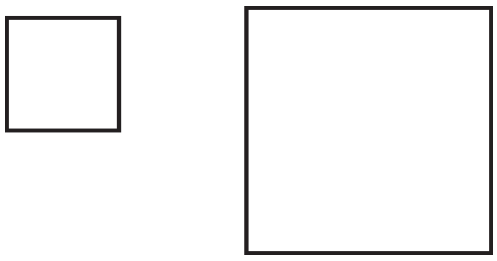


Fig. 2.5: Apparent depth produced by squares of unequal size. *After [mur73].*

Linear Perspective The optical mapping of a lens results in parallel structures appearing to converge. Orthogonal areas or contours are turned and compressed if they are not located directly in front of the observation point. Linear perspective is seldom found in nature but rather common in man-made constructions. It can be used to enhance

the apparent distance of an object by suggesting parallel structures when in reality they converge, see fig. 2.6.

Fig. 2.6: Apparent depth produced by suggesting parallel structures: house with pathway.

Accommodation The varying muscle tension in accommodation (see fig. 2.2) could deliver hints on the distance that an object in focus is located away from the viewer's eyes. It is unsolved whether the information on accommodation is a feedback from detectors in the muscles or a copy of the commands sent to the muscles from the brain. Also, it is not clear how this information can be interpreted correctly with varying levels of light in the environment. Wade and Swanston conclude that "the question of whether, and to what extent, accommodation actually does play a part in determining perceived distance remains controversial" [wad01].

Other Monocular Mechanisms Other monocular mechanisms of depth perception include the evaluation of contrast, clarity and brightness of an object with respect to other objects in the field of view. Light and shadow provide cues of the relative position of objects, similar to interposition cues. Texture gradients can give hints about position and extension of objects' surfaces.

Binocular Cues of Depth

The monocular visual field in most vertebrates is around 170° in extent. In humans, the presence of the protruding nose decreases the visual field. In more highly evolved species (such as humans), the eyes face forward, such that their combined visual field is much smaller than the added monocular visual fields. When the two eyes are used together, a binocular visual field emerges that overlaps by about 120° [ste00]. Because the two eyes are located at an interocular distance of around 60 to 70mm, they perceive a retinal image of the object from two distinct vantage points. These differences between the monocular views lead to the perception of stereopsis, the binocular form of depth perception, see fig. 2.7. It allows us to be much better at figure-ground segregation, avoiding collisions with looming objects, and accurately navigating through our environment than would be possible with one eye alone.

Convergence When focusing an object located at a close distance, the human eyes turn inwards. The physical distance to the point on which the eyes are converged can be determined from the interocular distance and the deviation angle of the eyes. As with accommodation, it is not clear whether the information on ocular convergence stems from detectors in the ocular muscles (see fig. 2.3) or directly from the brain areas which control the eyes' movements [wad01].

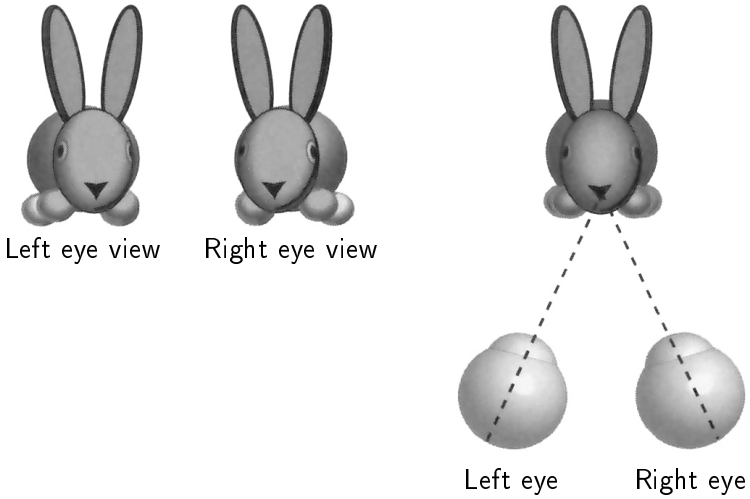


Fig. 2.7: Left and right eye view from different vantage points, leading to a combined view with perception of stereopsis, *modified after [ste00].*

Disparity When both eyes fixate on a point at given distance F, its image falls on the fovea of each eye. Points located nearer to or farther away from the viewer fall on non-corresponding points of the retina. If these disparities are not too great they can be processed in the visual system and provide cues for relative distance [wad01].

Non-visual Cues of Depth

A major concept of recognition of objects is perceptual constancy, the fact that known objects are recognized as the same objects despite variations in their appearance. Constancy also plays a major part in the perception of depth, of slant, and of size [mur73]. It is therefore involved in the processing of nearly all visual cues of depth.

(Self-)Motion in the Perception of Depth All the depth cues mentioned here 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 over an important role in depth perception. Motion parallax refers to 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 renders cues in the monocular case.

2.2 Auditory Perception

The ear is a truly amazing organ. Von Békésy [bek57] describes it as being

> "... so sensitive that it can almost hear the random rain of air molecules bouncing against the eardrum. Yet in spite of its extraordinary sensitivity the ear can withstand the pounding of sound waves strong enough to set the body vibrating."

2.2.1 Physiology of Auditory Perception

The hearing organ itself is made up of three parts: the outer ear, the middle ear and the inner ear, see fig. 2.8. What is commonly called 'the ear' is the visible part of the organ,

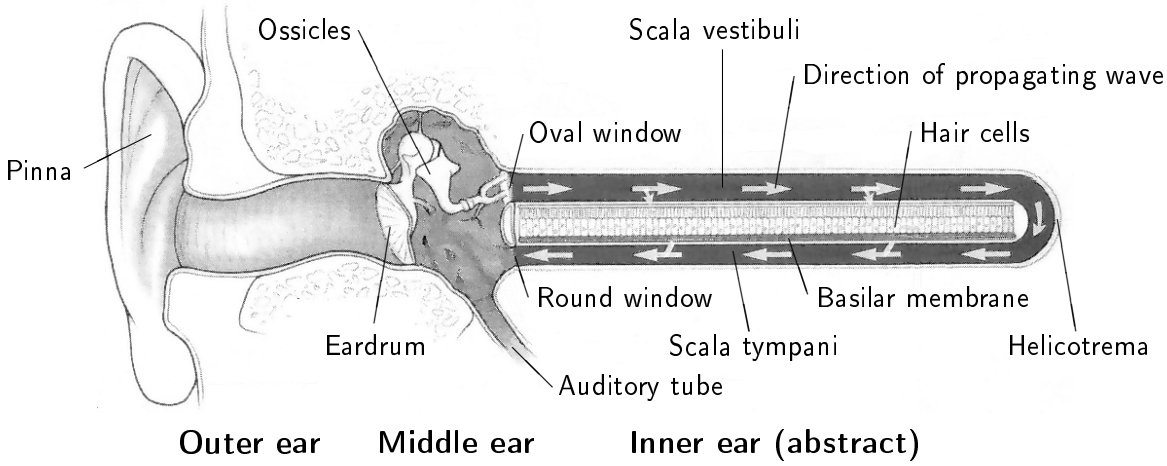


Fig. 2.8: The principal parts of the auditory apparatus with external, middle and internal (abstract) parts of the ear. *Modified after [you97].*

the outer ear. It has the responsibility of capturing the sound wave on the pinna and channeling it to the eardrum. This channel is about $20mm$ long, which corresponds to about a quarter of a wave length of a $4kHz$ sound wave propagating in air. Therefore the transmission from the free sound field to the eardrum works very well in the neighboring frequency range. Fig. 2.9 shows that in most humans with normal hearing the frequency range from $2-5kHz$ is of highest sensitivity [zoz98, zwi99].

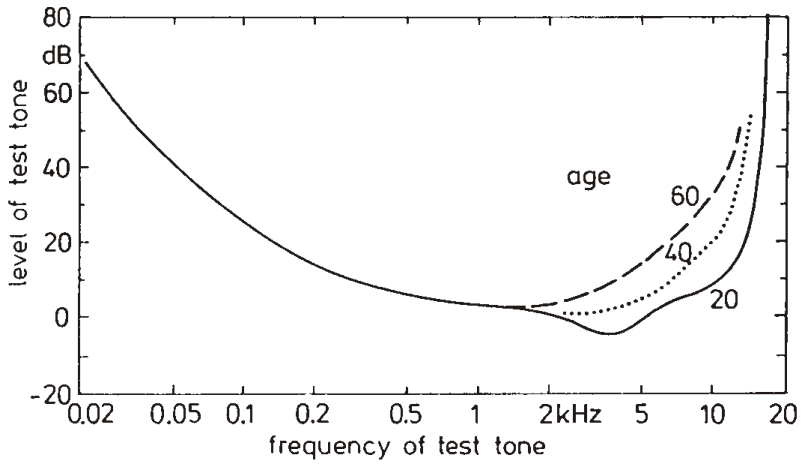


Fig. 2.9: Threshold in quiet as a function of frequency with age as a parameter. *Taken from [zwi99].*

The eardrum is the transition element between outer ear and middle ear. It transforms the incoming sound wave into mechanical vibrations. These get amplified via the ossicles (hammer, anvil, and stirrup), small bones which also serve as impedance transformers and which are connected to the inner ear via the oval window. The middle ear is a normally sealed cave which only opens to the pharynx via the auditory tube, also known as the Eustachian tube after Bartolomeo Eustachi, a 16th-century (c. 1500-1510 to 1574) Italian physician. The function of this tube is to protect, aerate and drain the middle ear.

The inner ear can be thought of as two largely independent organs: the semicircular canals which serve as the body's balance organ, and the cochlea. The latter converts

sound pressure impulses coming in from the middle ear at the oval window into electrical impulses which are passed on to the brain via the auditory nerve. The cochlea is a snail-shell like helical canal in the human temple bone, see fig. 2.10, with a diameter of $1cm$ at the basis and a height of about $5mm$. If stretched, the cochlea's $2\frac{1}{2}$ windings would result in a total length of about $35mm$. It is divided into three fluid-filled canals, see fig. 2.11. The oval window connects the stirrup in the middle ear to one of the canals, the scala vestibuli. The other canal (scala tympani) is connected to the round window, which serves as a pressure valve. The two canals are connected at the far end of the cochlea, called helicotrema. For very low frequencies and atmospheric pressure changes the helicotrema provides an exchange of fluid between the two canals, so that no stimulation takes place.

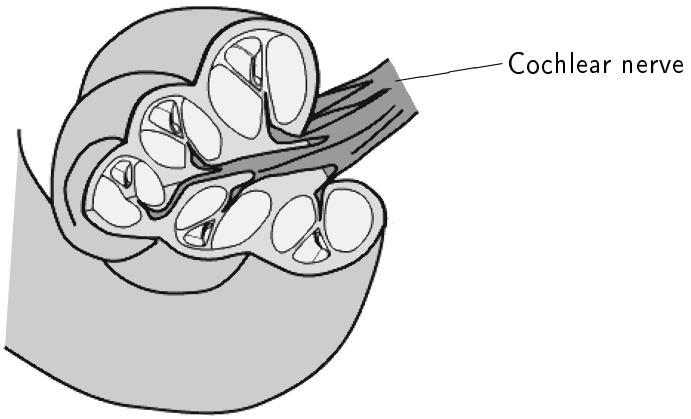


Fig. 2.10: Look-in cochlea structure with cochlear (auditory) nerve, *taken from [w-hyp].*

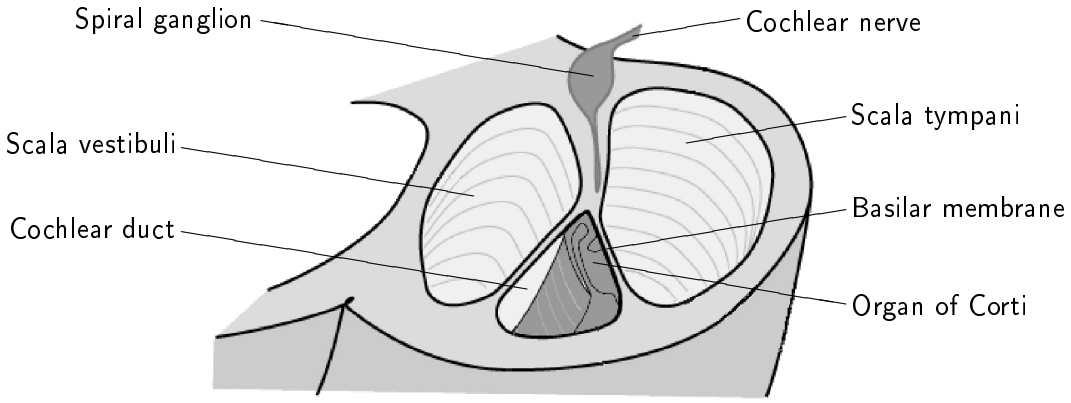


Fig. 2.11: Cross-section of cochlea with basilar membrane, *taken from [w-hyp].*

The third canal (scala media, also called cochlear duct) is of smaller cross-section and separated from the scala vestibuli by the Reissner's membrane. The scala media is separated from the scala tympani half by a bony partition wall (lamina spiralis) and half by the basilar membrane, which carries the sensitive organ of Corti. All three canals are rigid and filled with incompressible fluids with physical attributes similar to those of water. Whereas tympanic and vestibular canals are filled with perilymph (almost identical

to spinal fluid), the third compartment (scala media) is filled with endolymph, differing in terms of electrolytes. If the membranes are ruptured so that there is mixing of the fluids, the hearing is impaired.

The oscillation of the perilymph's molecules (introduced via the oval window by the movement of the stirrup) provokes the basilar membrane and thus the organ of Corti to oscillate vertically. Fig. 2.12 shows that the latter contains four rows of hair cells which protrude from its surface. Individual hair cells have multiple strands called stereocilia, see fig. 2.13. The vertical displacement results in the stereocilia being pressed against the tectorial membrane which can move in response to pressure variations in the fluid-filled tympanic and vestibular canals. There are between 16000 and 25000 hair cells distributed along the basilar membrane.

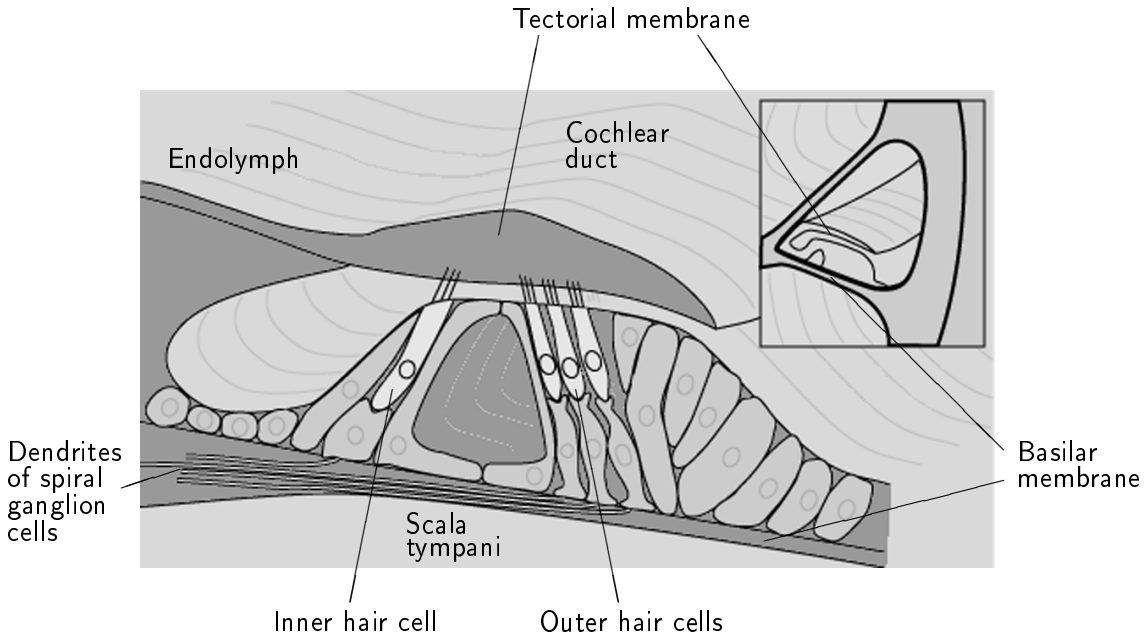


Fig. 2.12: The organ of Corti. Inset is the cross-section of the cochlea with the basilar membrane. *Modified from [w-hyp].*

As fig. 2.12 shows, there is one inner row of hair cells and three outer rows. The outer rows of hair cells are each located on a V-shaped line, fig. 2.13, and almost all are connected to efferent fibers (axons) of the auditory nerve. Each efferent fiber is connected to a significant number of outer hair cells. In comparison, the inner hair cells are located on a nearly straight line, and each inner hair cell is connected to up to 20 afferent fibers[1]. Conversely, this way each afferent fiber is connected to one single inner hair cell and thus related to a certain position in the helical organ of Corti [ter98]. This clearly assigns the main role of sensory reception to the inner hair cells, though inner and outer hair cells need to cooperate: a damage in the outer hair cells impairs hearing significantly, even when the inner hair cells remain fully operational [roe95]. One reason for this can be found in a process called the "cochlear amplifier". As the stirrup and oval window vibration induces a wave which travels toward the apex of the basilar membrane (the helicotrema), each

[1] There are basically two types of neural fibers: afferent (sensitive) fibers, which transmit excitation of the sensory cells to the central nervous system, and efferent (effectory) fibers, which transmit excitation from the central nervous system to the executing organs (e.g. muscles). Besides, there are also fibers with mixed functionality [roe95].

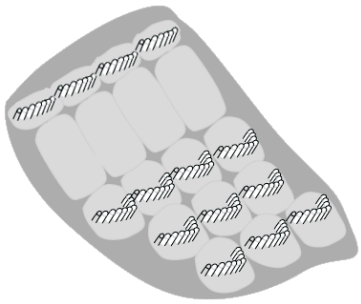


Fig. 2.13: Straight row of inner hair cells and three V-shaped rows of outer hair cells, *after [w-hyp], modified.*

frequency component approaches its cut-off point along the membrane and slows down. At the same time, the outer hair cells sense the basilar membrane motion and inject energy with the correct timing to enhance the vibrations. The complete role of the efferent fiber system is still under research, but it probably also oversees the operation of the inner ear.

2.2.2 Frequency and Directional Hearing / Auditory Localization

If stimulated with a pure tone of given frequency, the basilar membrane oscillates only in a locally limited area. The position of the oscillating area is depending on the frequency of the tone. This means that the basilar membrane provides an area of maximum sensitivity for each frequency, the so-called "resonance area". The lower the frequency of the tone, the closer the resonance area is to the helicotrema, where the basilar membrane is most flexible. The higher the frequency, the closer the resonance area is to the oval window, see fig. 2.8. Therefore, the basilar membrane can be regarded as a frequency-to-location coder, because each frequency can be assigned to a unique local position in the cochlea [roe95].

In 1960, von Békésy [bek60] published the relationship between frequency of a pure sinusoidal tone and location of maximum stimulus on the basilar membrane of an average grown-up human. Zwicker and Fastl rendered von Békésy's findings more precisely [zwi99]. This relationship is illustrated in fig. 2.14. Two interesting facts can be derived: First, the frequency range which is most important for the auditive quality of a (virtual) room (from roughly $20Hz$ to $8kHz$) requires far more than three quarters of the basilar membrane's length. Second, the doubling of a tone's frequency equals to a translation of the resonance area of about $4 - 5mm$ toward the base of the basilar membrane, independently from the original frequency of the tone.

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 fre-

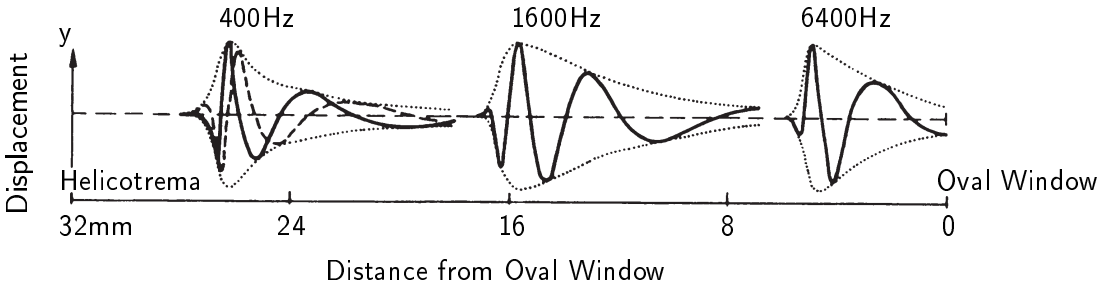


Fig. 2.14: Transformation of frequency to place along the basilar membrane. Three simultaneously presented tones of different frequencies cause traveling waves that reach their maximum at three different places corresponding to the different frequencies. *After [zwi99].*

quencies, more than one wavelength fits between the two ears, thus making the comparison of phase information between left and right ear equivocal [bra05]. Interaural level differences (ILDs)[2] between the two ears are the primary cues for signals with frequencies above $1500Hz$ [bla01]. Regardless of the source position, ILDs are small at low frequencies. This is because the dimensions of head and pinnae are small in comparison to the wavelengths at frequencies below about $1500Hz$, thus not representing 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 torso, head, and pinnae. Each direction of incidence (azimuth, elevation) is related to a unique spectral filtering for each individual. The 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 essential for resolving front-back confusions [bla01]. Pulkki reports that for elevation perception frequencies around $6kHz$ are especially important [pul01].

It should be noted that in everyday situations, localization of sound sources seldom relies on auditory cues alone. Knowledge of the potential source of a sound (e.g., airplane noises from above or crunching shoes from below) aids in the localization process. Of course, visual cues also influence sound localization, see subsection 2.4.1.

2.2.3 Auditory Perception of Space

The auditory perception of space is based on evaluating not one but a number of auditory cues. In order to analyze the acoustical characteristics of an enclosing space, the human auditory system tries to compare the perceived pattern of cues (direct sound, early reflections and reverberation, see chapter 3) with patterns memorized from past audiovisual experiences. The perception of space comprises attributes like depth, size, spaciousness (or envelopment, according to Griesinger [gri99]) and distance, to name a few.[3]

The perceptual impression of different patterns of auditory cues has been studied in-depth by a number of researchers. Its knowledge is crucial for the creation of artificial, yet naturally sounding room acoustic effects. Griesinger claims that the most important attribute in the auditory perception of space is spaciousness or envelopment. He has

[2]Sometimes, *Interaural Level Differences* are also called *Interaural Intensity Differences*, IIDs.

[3]In fact, the problem of denominating in a meaningful way the parameters or attributes that describe spatial quality is not completely solved yet. There is a considerable lack in the clarity of standard definitions of subjective attributes related to auditory perception of space, see e.g. [rum02].

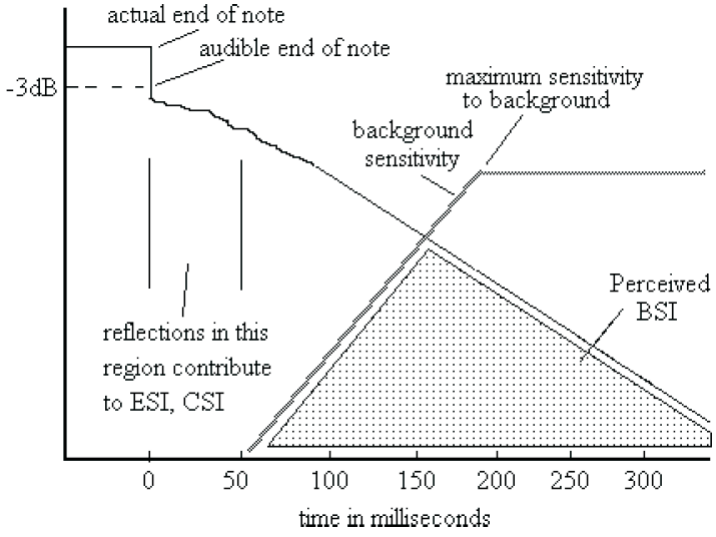


Fig. 2.15: Timeline model of auditory perception of space, *from Griesinger, [gri99].*

published a huge collection of professional articles on the topic of reverberation patterns and the question whether these patterns are perceived as enveloping or not, see e.g. [gri97, gri98, gri99].

Griesinger differentiates between three kinds of spatial impressions, see fig. 2.15. They depend on the type of signal they are caused by:

- Continuous Spatial Impression (CSI): caused by lateral reflected energy interfering with a continuous signal; CSI depends on the ratio between medial and lateral sound.
- Early Spatial Impression (ESI): caused by lateral reflected energy arriving within $50ms$ after the end of an impulsive signal; ESI also depends on the ratio between medial and lateral sound.
- Background Spatial Impression (BSI): caused by impulsive signals (e.g. speech) which are organized into a foreground perceptual stream by the brain; energy arriving $150ms$ after the end of impulsive signals is perceived as BSI if sufficiently spatially diffuse. BSI depends on the amount of spatially diffuse reverberant energy.

Only CSI and BSI are enveloping, with BSI being dominant. ESI is not enveloping, but contributes to the sensation of apparent source width. Early reflections of continuous signals contribute significantly to CSI in the range from $20ms$ to around $100ms$ [gri99]. Griesinger suggests that spaciousness is primarily determined by lateral reflected energy arriving at least $80ms$ after the direct sound. At frequencies above $500Hz$, spaciousness is reported to depend primarily on the spatial diffusion of the reverberant component of the sound. Below $500Hz$ spaciousness is primarily determined by the phase coherence of the direct and the reverberant sound cues [gri97].

2.3 Mechanisms of Cognition / Processing of Percepts

Until now, relatively little is known about the mechanisms of bimodal processing in the human brain. The main questions with respect to audiovisual perception are: At what level

of perceptual processing do cross-modal interactions occur? What mechanism underlies them?

In order to answer these questions, it is reasonable to start looking at the unimodal processing of stimuli. The following subsections will discuss the processing of stimuli perceived via the sensory receptors of vision and audition separately.

2.3.1 Transmission of Information in the Nervous System

Nervous cells or neurons have the task of transmitting information from sender to receiver inside of the body. Therefore they are often organized as long fibers, along which variations of electrical potentials are traveling. These have the form of impulses of small electrical currents. The impulses code the information to be transmitted, e.g. the effect of light upon the receptors of the retina. Whenever a neuron is inactive, there is a difference in potential between its cell body and its environment. This difference in potential is the result of different concentrations of electrically charged molecules (iones) inside and outside of the neuron. It is called the resting potential of the cell.

Certain iones, especially sodium (Na) and potassium (Ka), are able to penetrate the cell walls, thus changing the difference in potential between cell body and environment. Such changes in potential can propagate across the whole of the cell wall, i.e. along the nervous fiber, if they are only strong enough. These propagating impulses are called action potentials. Located at the end of each nervous fiber is a synapse, in which incoming impulses lead to certain bio-chemical reactions. In these, so-called transmitter substances are produced. The transmitter substances travel to the cell walls of neighboring nervous cells and produce new action potentials themselves, once their concentration reaches a certain threshold. This way, external stimuli that impact receptor cells (e.g. light, sound, touch, changes in temperature, etc.) and that lead to chemical or electrical reactions, can be transmitted across many stations until finally reaching their destination in the respective areas of the brain.

Nervous fibers that originate from the receptors and that transmit information to the brain are called “afferent”, whereas “efferent” fibers are those that transmit impulses originating from the brain, e.g. to control the muscles’ activity.

2.3.2 Methods of Examination

Classical methods for neuronal and fiber staining, applied as early as 1880 by Franz Nissl on subcellular components of neurons and glia [ait90], are helpful in determining the paths of individual axons. As axons and cell bodies of the central nervous system cannot exist in isolation from each other, they are heavily connected. By looking at the amount of axons connecting regions in the brain, it is possible to separate and group functional areas. Tracing the nervous fibers can then identify regions related to the processing of certain types of stimuli.

Especially the processing of visual stimuli has been studied in detail, applying methods like the microelectrode recording technique. Development of microelectrodes for recording from neurons started in the 1950s. Microelectrodes are narrow tubes made of electrolyte or metal, surrounded by glass or varnish everywhere but at the tip. With tip diameters as low as $0.1\mu m$ or less they are suited for intracellular recordings of resting, synaptic and action potentials of individual neurons. Yet, *in vivo* recordings from the cortex are difficult because of vascular and respiratory pulsations, which can rarely be eliminated [ait90].

Additionally, micropipettes (glass microelectrodes with tubes) can be used to inject small quantities of dye or tracer into or around cortical neurons. Using this method, the morphology of the recorded neuron can be determined. Extracellular injection allows to identify the afferent and efferent paths of the micro-area from which the recordings are made. Microelectrodes have also been used to make microlesions in the cortex by applying a low positive direct current, thus causing a localized burn in the neural tissue [ait90].

Another method for the study of processing mechanisms in the brain is "the morphological investigation of tissue that has been shown to have been activated by a stimulus" [ait90], also named *functional anatomy*. During increased activity in the brain the local arterioles dilate, as there is increased blood flow to the active area. Substances normally uniformly distributed in the blood will show high regional concentrations in the active area. Also, metabolically active neural tissue consumes high amounts of glucose. If metabolically inert homologues of glucose are provided, these will be taken up and accumulated in the respiring tissue. Both mechanisms can be proven by incorporating positron-emitting (radioactive) atoms into the substances.

2.3.3 Neurophysiology of the Human Brain

The human brain (cerebrum) is part of the central nervous system: it is the central processing unit for stimuli perceived via the sensory receptors of sight (vision), hearing (audition), taste (gustation), smell (olfaction), and touch (taction or pressure)[4]. Its main part, the *telencephalon*, is divided into left and right hemispheres which are connected by the *corpus callosum*, a large bundle of nerve fibers. The two sides of the brain are similar in appearance, and every structure in each hemisphere is mirrored on the other side. Despite these strong similarities, the functions of each cortical hemisphere are different. This lateralization has been most thoroughly studied investigating the processing of language: grammar and word production, both processes that are of mainly linear (sequential) or logical character, are often lateralized to the left hemisphere of the brain. Therefore, a very popular belief is that the right hemisphere is more related to creativity and emotion, a belief that has been vitiated in a meta-analysis of findings from neuroimaging by Wager et al. [wag03].

The *telencephalon* is usually referred to as 'the brain', although the latter also includes the *cerebellum* or 'little brain', a region that plays an important role in the integration of sensory perception and motor output. The cerebellum is connected heavily with the motor cortex which controls muscles' movements and the spinocerebellar tract, which provides proprioceptional information, thus integrating these pathways in order to fine-tune motor movements.

Areas Each hemisphere of the *telencephalon* can be divided into four areas or *lobes*, see fig. 2.16. The *frontal lobes* are attributed to so-called executive functions, i.e. the ability to recognize future consequences resulting from current actions, to choose between good and bad actions (or better and best), suppress and override unacceptable social responses, and find similarities and differences between things or events. They also play an important part in retaining long-term memories which are not task-based. The *parietal lobes* are located posterior to the frontal lobes and superior to the occipital lobes. They integrate

[4] Equilibrioception or sense of balance is often regarded as another physiological sense. It is mainly based on the detection of acceleration.

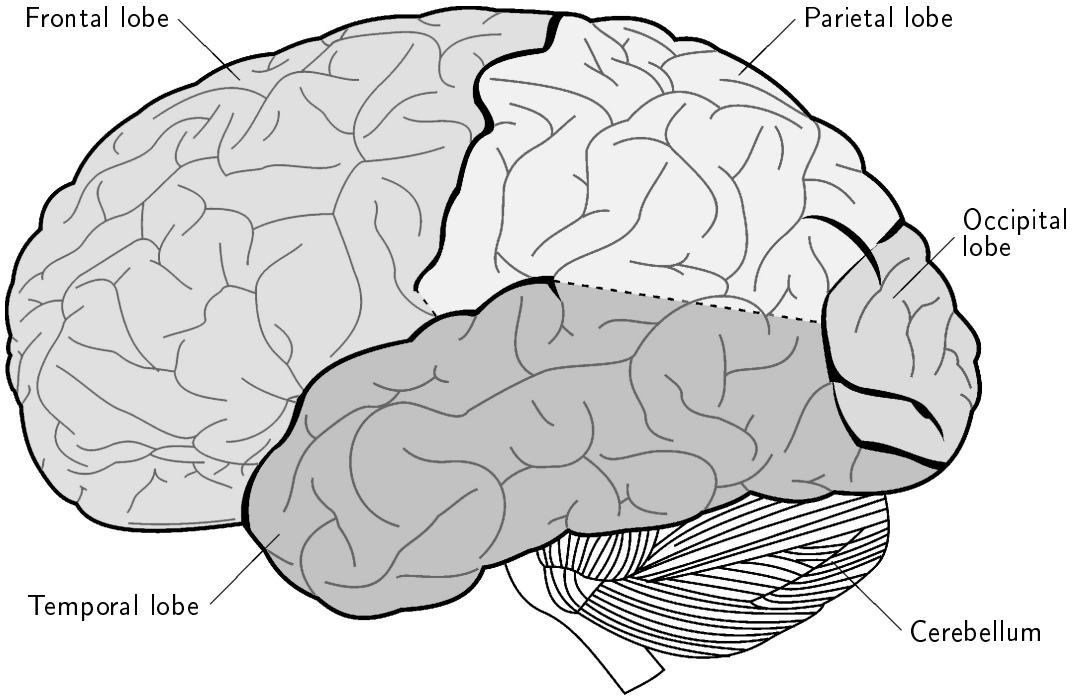


Fig. 2.16: Principal fissures and lobes of the cerebrum, lateral view of left hemisphere. *After figure 728 from Gray's Anatomy [w-gra].*

sensory information from different modalities, especially for determining spatial locations of objects. E.g., the parietal lobes play a vital role in mapping objects perceived visually into (relative) body coordinate positions. Therefore the parietal lobes are of most interest to this work - unfortunately, much less is known about these lobes than the other three pairs of lobes in the cerebrum. The *occipital lobes* are the visual processing center of the human brain. Their function is to control vision and color recognition. Finally, the *temporal lobes* are located at the sides of the cerebrum, inferior to the parietal and frontal lobes. They are involved in auditory processing, in semantics (in both speech and vision) and in memory formation [nol02].

Already in 1909, Korbinian Brodmann published his definition of the cerebral cortex into 52 distinct regions, based on their histological characteristics [bro09]. These areas, nowadays called Brodmann-areas (BA), have later been associated to nervous functions. The most important areas in the audiovisual context are BA17 (Primary Visual Cortex, V1), BA18 (Visual Association Cortex, V2) and BA19 (V3), as well as BA41 and BA42 (Primary Auditory Cortex, anterior and posterior transverse temporal areas (H)), see figs. 2.17 and 2.18.

Visual Pathway and Cortex The optic nerves, transmitting sensory information from each eye, proceed posteriorly and medially to unite at the optic chiasm. There, fibers from the nasal halves of each retina cross over to the opposite hemisphere. Fibers from the temporal halves project to the hemisphere on the same side. The result is that signals from the same regions of the visual field are projecting to the same hemisphere; thus, the left half of the visual field projects to the right halves of the retinas, which in turn send neural signals to the right hemisphere. Departing from the optic chiasm are the optic tracts. The

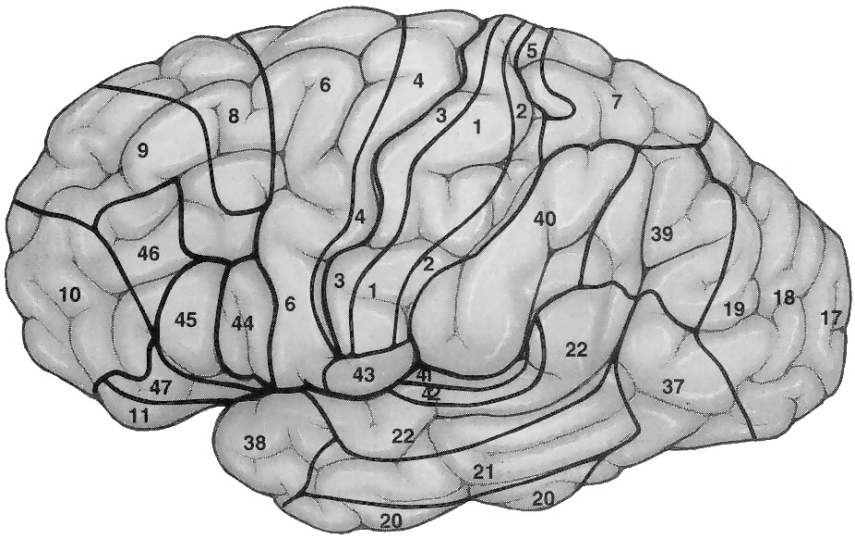


Fig. 2.17: Lateral view of the left hemisphere with Brodmann areas associated to visual (BA17-BA19) and auditory (BA41-BA42) perception. *From [you97].*

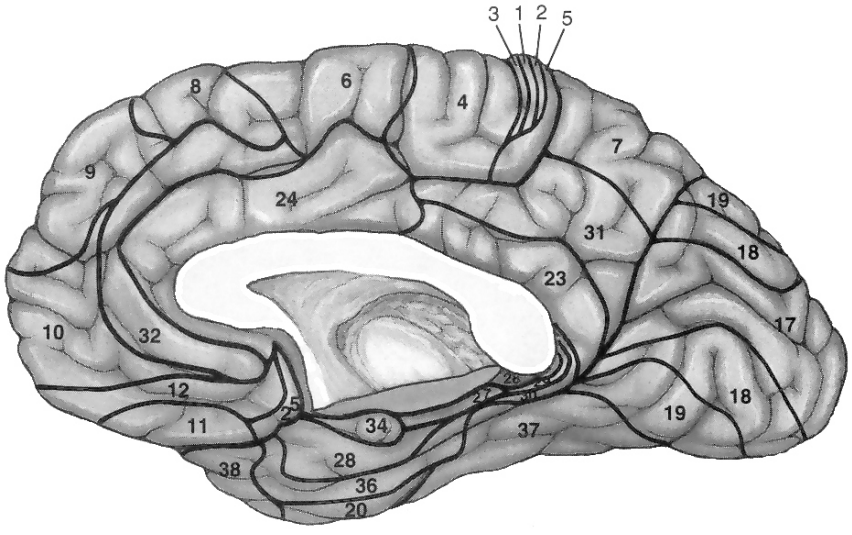


Fig. 2.18: Medial view of the right hemisphere with Brodmann areas associated to visual (BA17-BA19) perception. *From [you97].*

nervous fibers terminate in a subcortical area called the lateral geniculate nucleus (LGN) located in the thalamus [wad01].

The LGN consists of six layers, three of them being connected to the contralateral eye and three connected to the ipsilateral eye. Two of the layers, one being connected to each eye, contain larger neurons that receive their inputs from a class of ganglion cells in the retina that are more sensitive to movement and contrast. The other four layers contain small neurons that receive their input from the numerically larger class of small ganglion cells that are sensitive to form and color. Yet, all information from a given point in the visual field ends up in a column extending through all six geniculate layers [nol02].

Cells from the LGN project via the optic radiation to the primary visual cortex V1 located in Brodmann area BA17. There is a point-to-point relation between the retina, lateral geniculate nucleus, and the primary visual cortex. Impulses from the upper and lower halves of the visual field are located in different parts of the optic radiation, see fig. 2.19, and consequently also project into different areas of the primary visual cortex. This is called retinotopic projection, as the distribution of stimulation on the retina is preserved in the cortex. Yet, the precise separations between adjacent regions of the retina and their projections into adjacent regions in the cortex are not retained. The fovea, being only a small area in the retina (macular area), has a projection into the cortex that is disproportionally large. In turn, the large peripheral areas of the retina are only presented by small areas in the cortex[5], see fig. 2.19. This fig. also shows an overview of the visual pathway.

Although the LGN receives signals from both eyes, these project to different layers of the LGN, see above. Binocular integration first occurs in the primary visual cortex V1 [wad01]. Although the cortical neurons can be excited by stimulation of either eye, they will not mix until reaching V1. Projections from the ipsilateral or contralateral eye remain also separate in V1, which results in a horizontal organization based on ocular dominance. In V1, every $0.5mm$ columns of eye preference alternate between ipsilateral and contralateral eye [hub87]. Fig. 2.20 shows the distribution of areas with different ocular dominance over the surface of V1, unfolded into a plane.

Within each column of ocular dominance, a large variety of contrast-sensitive cortical neurons with different complexity can be found ("simple", "complex" and "hypercomplex" neurons). Most prominent are the orientation selective neurons discovered by Hubel and Wiesel in 1959. These are neurons that "analyze" the orientation of a bar of light projected on the retina. Their firing rate will be highest for a bar of light of orientation matching their preference, and no activity will be observed for bars of perpendicular orientation. These neurons differ in their preferred orientation, and a variety of them exist such that all orientations can be detected. In V1 there are also cortical blobs that respond to the wavelength of the original light stimulus, with input from all layers of the LGN. All neurons are arranged in a repeating pattern, such that within a cortical area of about $1mm^2$, both eyes and all orientations are represented for a small region of the visual field [wad01].

The primary visual cortex V1 seems to separate the pattern of light falling on the retina into discrete features. Apparently, these are retinal location, orientation, movement, wavelength, and the difference between the two eyes. In subsequent cortical processing these features are further differentiated. Therefore, the primary visual cortex has the task of sorting visual information and distributing them to other, more specialized cortical

[5]This is a general scheme found in the cortical projections of all sensory systems: those parts of the sensory surface with greatest sensitivity also have the greatest cortical projection.

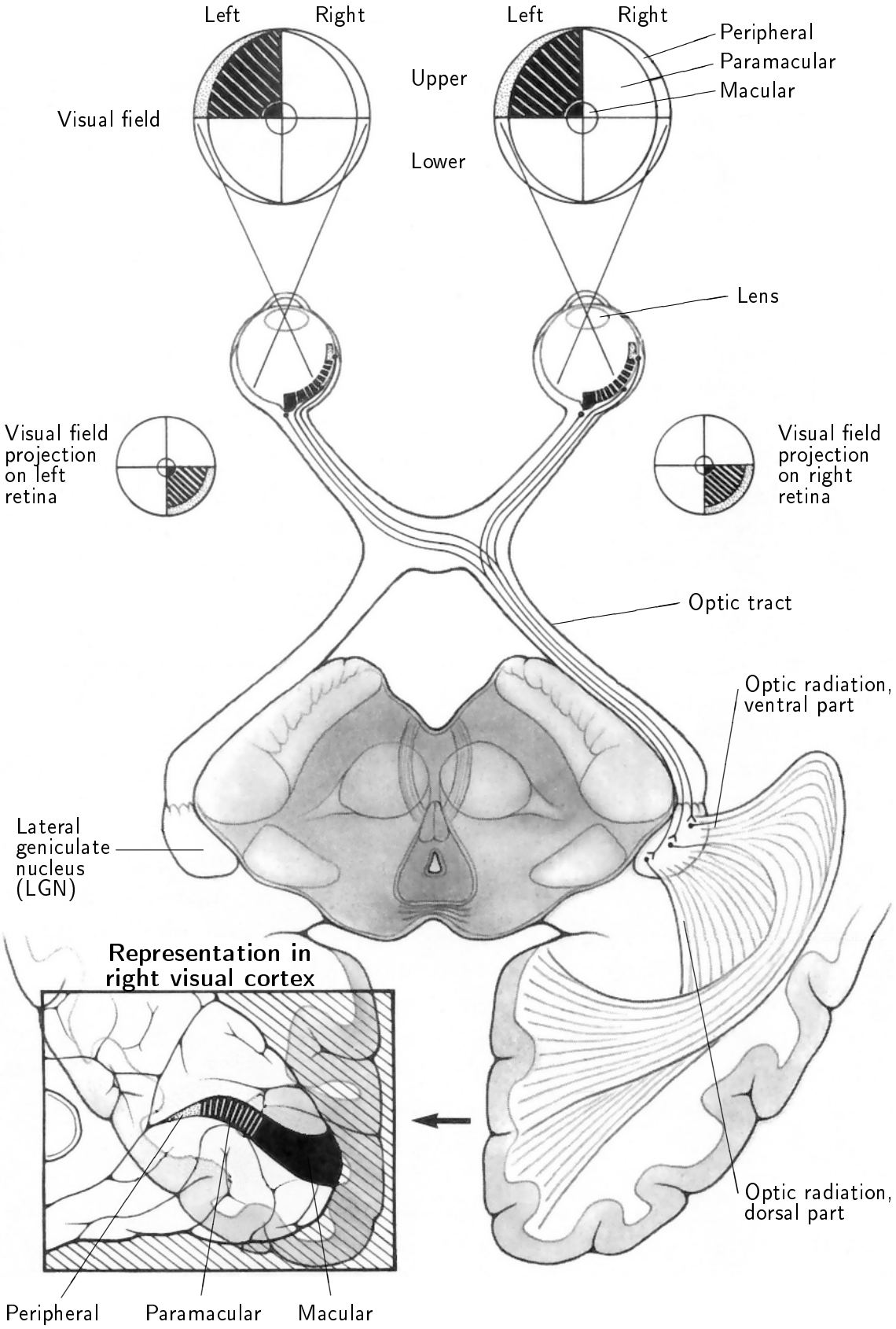


Fig. 2.19: Visual pathways from the eyes to the brain, *modified from [you97].* The ventral part of the optic radiation is located mainly in the temporal lobe, the dorsal part mainly in the parietal lobe.

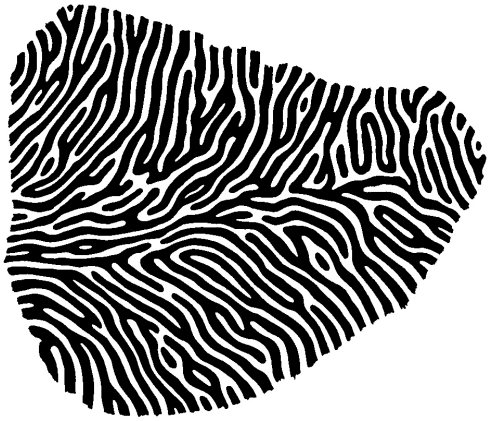


Fig. 2.20: The pattern of ocular dominance columns, represented by black and white areas, across V1. Note that V1 has been unfolded into a plane. Shown here is the pattern of the right hemisphere, *from [hub87].*

areas.

Two visual streams have been identified by Ungerleider and Mishkin in 1982 that originate from the primary visual cortex: the dorsal or parietal stream, passing dorsally to the posterior parietal areas of the cortex, and the ventral or temporal stream, passing ventrally to the inferotemporal cortex, see fig. 2.21. Apparently, the first correlates more to location, depth and movement, whereas the latter is more connected to color, spatial detail and form [nol02, wad01]. Goldstein, on the basis of the experiments performed by Ungerleider and Mishkin, suggests that perceiving the location of an object is attributed to the dorsal stream, whereas the ventral stream determines the object's identity - the *where* and *what* dimensions of vision [gol02].

Wade and Swanston report an interesting revision of the functions served by these streams in terms of action and perception [wad01]: "The dorsal stream is said to be concerned with motor control and the ventral stream with perceptual representation." They argue that the distinction relates to the perception of visual illusions vs. our response to them: whereas our perception might be distorted by such patterns as the Müller-Lyer configuration, see fig. 2.22, our fingers would be adjusted correctly to the appropriate physical length of the two sticks rather than to their perceived inequality. According to Wade and Swanston this indicates that action and perception are processed separately from each other.

Apart from the classical projection pathway from the eyes to the brain via the LGN, fig. 2.23 also shows the - in evolutionary terms - older visual pathway to the superior colliculus. Rather than only synapsing in the LGN, the optic tract also connects to the superior colliculus. From there, the collicular fibers connect to three nuclei which are responsible for the control of eye movements. The superior colliculus also receives input from the primary visual cortex V1.

Although the basic mechanisms of sensory information transmission are well understood, a detailed understanding of how visual input is processed and interpreted is still missing. Especially the transition from neuronal reaction to perception, i.e. the process of attaching a meaning to the stimulus, remains unexplained.

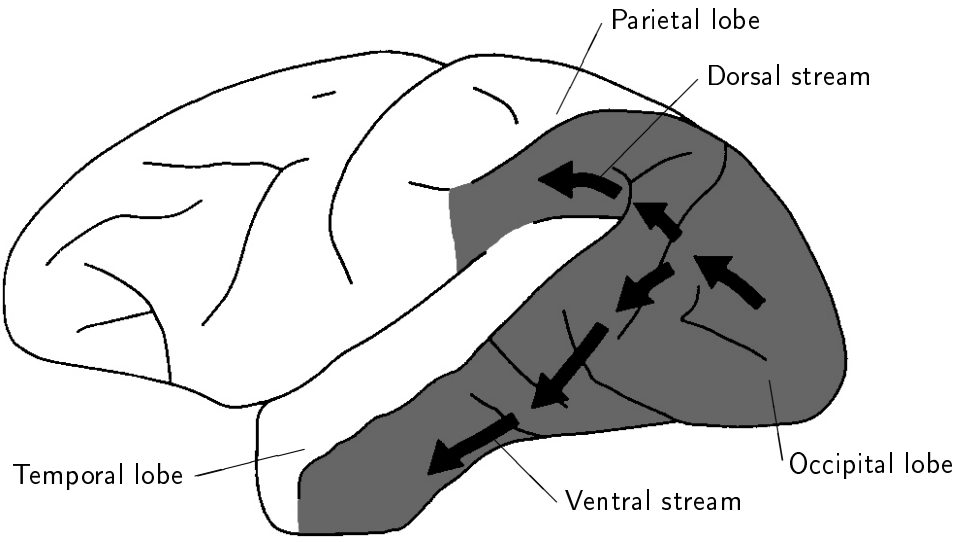


Fig. 2.21: Dorsal and ventral visual streams, simplified, *after [wad01].* Both streams depart from the primary visual cortex V1. Along the stream, information is passed on across a number of synapses, represented by the arrows.

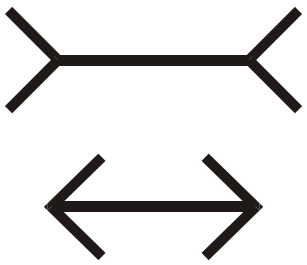


Fig. 2.22: The Müller-Lyer configuration causes distorted perception of length.

Auditory Pathway and Cortex Because of a bilateral representation of auditory impulses on each side of the brain, the central auditory pathways are different from other ascending sensory paths. As fig. 2.24 shows, the left and right auditory pathways are interconnected at various points in the processing chain: nuclei on opposite sides of the brain are connected to each other. Three groups of nuclei are located along the auditory pathways between the cochlear nuclei and the inferior colliculus. They are the superior olivary nuclei, the nuclei of the trapezoid body (not shown), and the nuclei of the lateral lemniscus.

The cochlear nerve terminates in the dorsal and ventral cochlear nuclei. The superior olivary nucleus then receives input from both the ipsilateral and the contralateral cochlear nuclei. The olivary nuclei play a key role in the localization of sound in space, as they combine and compare stimuli originating from the left and right ear on a low processing level. Originating from the medial superior olivary nucleus (MSO) is the so-called timing pathway, which transports interaural phase difference cues to the lateral lemniscus and the inferior colliculus (IC). Similar connections exist that originate from the lateral superior olivary nucleus (LSO), which possibly transport interaural intensity cues [sch06]. Although many interconnections between the nuclei have been described in literature, lots of details of the auditory pathways are still unknown [gol02].

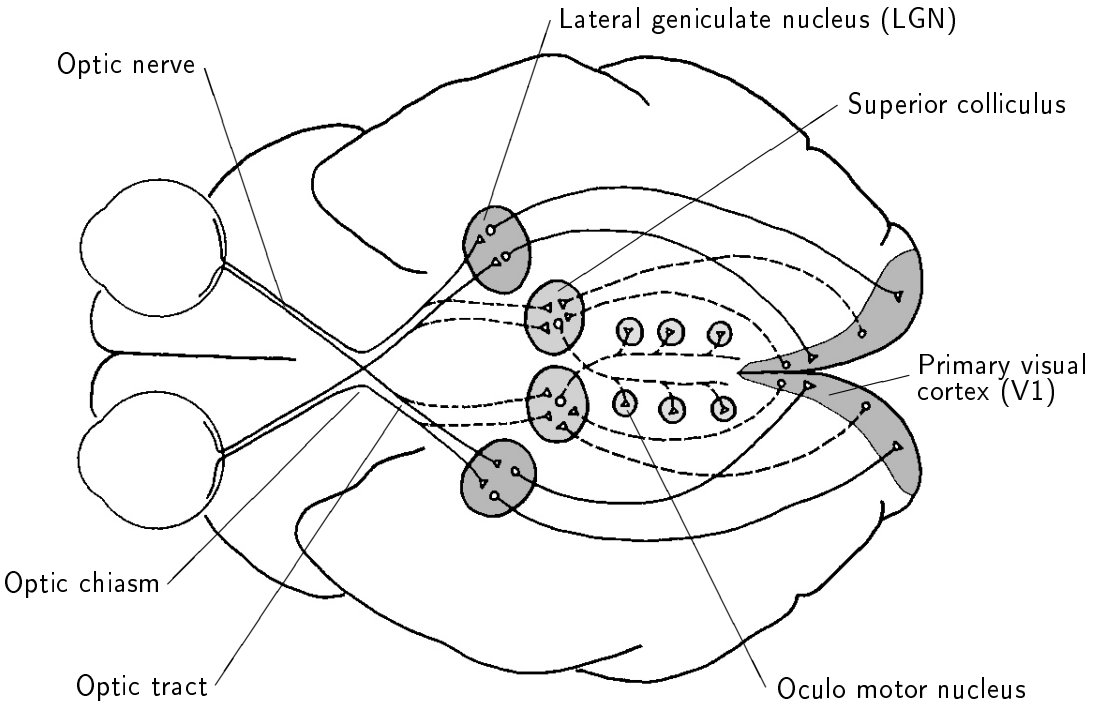


Fig. 2.23: Classical (solid line) and evolutionary older (dotted line) pathways, *modified from [wad01].* In the older pathway the optic tract also projects to the superior colliculus.

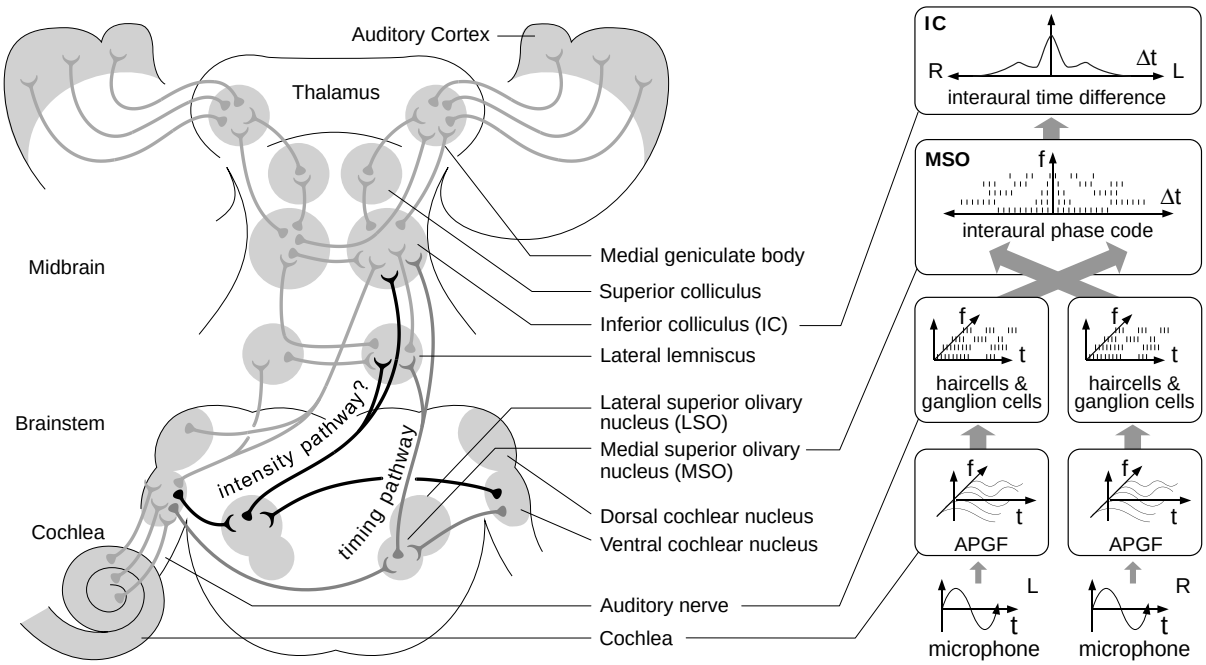


Fig. 2.24: Main afferent components of the auditory pathways. On the right hand side a technical representation of the components' functionality is suggested. *From [sch06].*

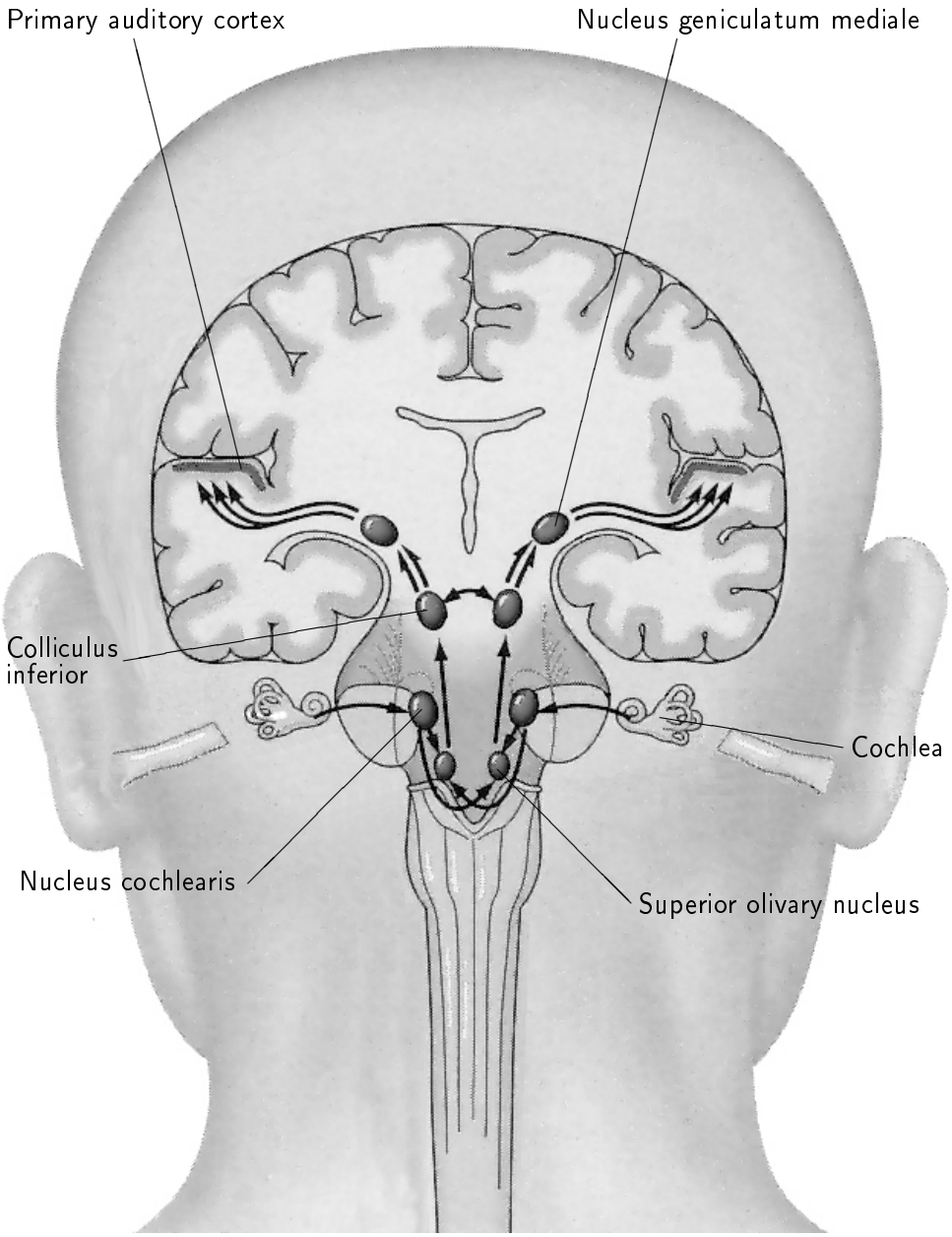


Fig. 2.25: Schematic view (strongly simplified) of the auditory pathways. *From [gol02], modified.*

Fig. 2.25 shows in a simplified view the location and elements of the auditory pathways. Apart from the nerve fibers that pass on signals from the cochlea to the auditory cortex (afferent components), there are also corticofugal fibers that transmit signals back from the brain to the cochlea. These signals are related to the active nonlinear feedback that our peripheral hearing system is assumed to use for an increase of the perceivable dynamic range. Although it is not clear in detail how such a system acts, Zwicker and Fastl deduce the basic structure from the functional behaviour observed, see [zwi99].

Superior and inferior colliculus located in the midbrain pass on the cues to the medial geniculate body located in the thalamus, from where the nerve fibers connect to the primary auditory cortex. It is located in the transverse temporal gyri of Heschl, on the dorsal surface of the superior temporal gyrus and buried in the lateral fissure, see figs. 2.16 and 2.25.

The auditory cortex consists of three main areas: the core area AI that contains the primary auditory cortex, the secondary auditory cortex and the associative auditory cortex (Brodmann Areas BA41 and BA42). Tonotopic localization is believed to exist in the primary auditory cortex, with high frequency tones represented posteromedially and low frequency tones anterolaterally [you97]. However, Aitkin reports that in AI lesion experiments cats were able to discriminate frequencies even after complete bilateral lesions of the AI. Yet, frequency detection was impaired if frequency was "tied to other attributes of an auditory stimulus such as its location in space [...] It may not therefore be appropriate to study frequency discrimination without relating the frequency to location in space, or to other parameters such as intensity, temporal pattern, duration and modulation" [ait90].

2.3.4 Joint Processing of Audiovisual Stimuli

After looking at the perception and processing of auditory and visual stimuli separately, this subsection will present an overview of studies performed with the aim of identifying the mechanisms of bimodal audiovisual processing. As already stated in the introduction to section 2.3, the main question here is at what level of perceptual processing do cross-modal interactions occur? Where, in which regions of the human brain, does the processing of audiovisual stimuli take place? Are these different from the areas of processing for unimodal (auditory or visual alone) stimuli?

If the latter question should be answered positively, then most of the plethora of experiments that have been performed to evaluate human perceptual processing would have been done in vain because they would not be transferable to everyday situations. These experiments usually try to simplify the experimental setup as much as possible in order to diminish or exclude side-effects and potentially influencing variables. They are performed using simple stimuli, mostly presented unimodally, in clinical laboratory environments that separate test subjects from the "real world". This separation from real world influences is very relevant when looking at aspects of (selective) attention. These aspects will be discussed further in section 2.5.

In fact, Goldstein reports an experiment performed by Graziano and Gross in 1995 in which they had found evidence for the existence of bimodal neurons [gol02]. Experimenting with a monkey, they found neurons in the cortex that fired whenever the hand of the monkey was touched lightly. The same neurons also fired when presented with certain visual stimuli: they responded whenever there was visual movement within a certain area around the hand, even without the hand being actually touched. Yet, when the hand was held such that it could not be seen by the monkey, the neurons did not react to approximation any more but only to actual touch.

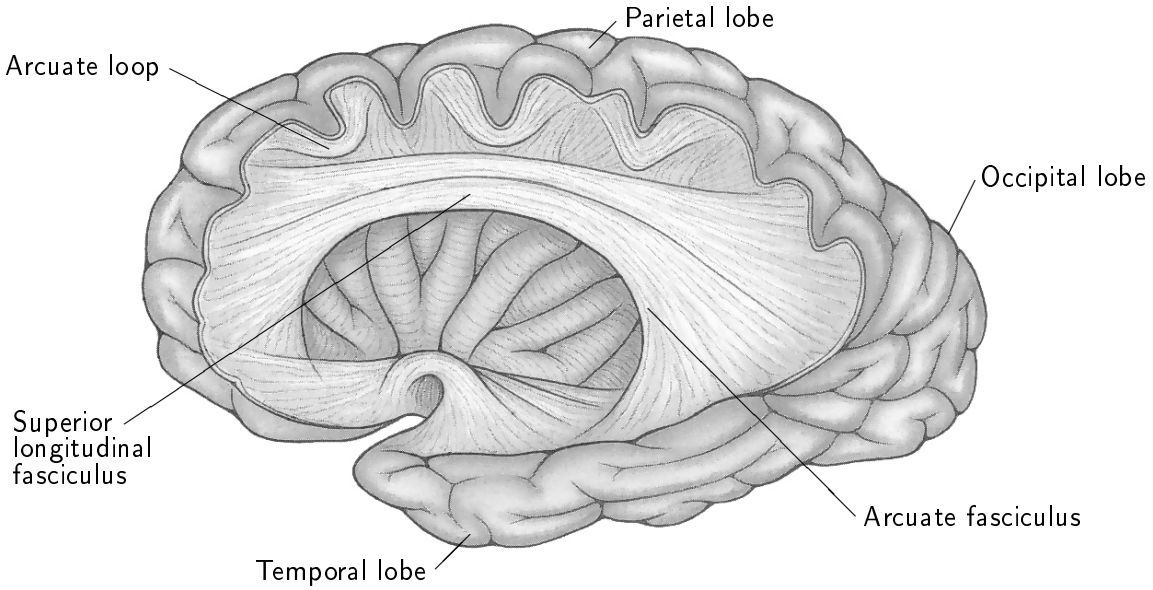


Fig. 2.26: The major association bundles connecting functional areas of the neocortex, dissected from lateral aspect. *From [you97], modified.*

Graziano and Gross call these types of neurons body-centered neurons (opposed to retina-centered visual neurons), because apparently they serve to indicate the relative position of hands or body surface with respect to visual objects. They correlate vision and touch and provide information necessary for interaction with our environment.

However, it is common ground that this type of bimodal neuron cannot be the sole mechanism for bimodal or even multi-modal processing. This is because in natural environments the perception of multi-modal stimuli is not the exception but the rule. If multi-modal processing was only performed through multi-modal neurons, then their number would have to be much greater. Therefore the question to be answered here is how the neurons of the highly specialized cortical areas for visual, auditory, and other sensory perception communicate.

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 which are oriented perpendicular to the surface - the so-called cortical columns. 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, e.g. the superior longitudinal fasciculus, 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. Fig. 2.26 shows that they occupy most of the space underneath the cortex.

As can be seen, there are many such connections between different functional areas of the neocortex, such that information can be exchanged between these and true multi-modal processing can be achieved. Yet, how is the information coded that is to be shared between functional areas? How is context-dependent binding or segmentation of stimuli achieved? This is a topic that is discussed very controversially, as it also plays an important role in the perception of unimodal stimuli. Goldstein gives an example of a red, rolling ball entering our field of view. Locally distinct neurons are activated by either movement,

shape or color. Subsequently, dorsal and ventral streams are also activated. However, we perceive one singular object, not separate "rolling", "red color" or "round shape" [gol02].

Until now, it is unclear how the processing of multitude characteristics of an 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. The most prominent theory, suggested by Singer and Engel et al., assumes that visual objects are represented by groups of neurons [gol02]. 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.

Verschure and König [ver99] give an overview of the role of biophysical properties of cortical neurons in binding and segmentation of visual scenes that can be transferred to the processing of multi-modal scenes. Yet, the *binding by synchrony* theory has left doubts with respect to the interpretation and processing of the synchrony code. Klein et al. 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" [kle03]. They have found evidence that the coding of relevant natural sounds can be described by a sparse spectrotemporal coding scheme that is similar to the sparse representation of images, a computational description that has already lead to powerful algorithms for image denoising, compression and preprocessing for image recognition [kle03]. Many others argue that binding can be explained by (selective) attention [gol02]. Attention is discussed in section 2.5.

Summing up, the binding problem needs to be resolved before we can explain the connection between neuronal reaction and perception. Yet, this connection might be the key to understanding multi-modal processing from a neurophysiological point of view. Binding may be further evaluated by comparing event-related potentials (ERP) to behavioral results, as suggested by Thompson et al. [tho01], or by applying functional magnetic resonance imaging (fMRI) techniques as done by Martinkauppi et al. [mar00], Cohen et al. [coh97], and Owen et al. [owe05] in the analysis of working memory tasks.

2.4 Known Effects in Human Bimodal Perception

This section describes a few well-known effects that frequently occur when humans are presented with multi-modal stimuli. Some of these effects are well documented, though not necessarily all are well understood. Far more of these effects exist than can be described here. For a broader overview refer to e.g. [mur73].

2.4.1 Visual Dominance

The dominance of visual stimuli over other modalities is very often accepted as being given naturally. In fact, looking at our everyday experiences we might be inclined to accept this posit without further discussion: because "seeing is believing", often we think that we tend to trust our eyes more than the other senses. Yet, this appraisement is often due to the fact that in the real world we seldom are confronted with 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 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 contradicting sensory information and see what the generally dominating modality is - if there is any. In fact, 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 in conflict. One model suggests that the signal typically most reliable dominates the competition completely and thus "cashes in on the complete ballot": the judgment is based exclusively on the dominant signal. In the context of spatial localization based on visual and auditory signals, 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 stemming 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. The 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. The *MLE* model is also known as the Kalman filter in engineering literature [kal61].

Battaglia et al. in [bat03] report that several investigators have examined whether human adults actually combine information from multiple sensory sources in a statistically optimal manner according to the MLE model. They further explain that "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." Looking at it this way, visual capture is just a special case of the MLE model: one 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. in [bat03] describe an experiment designed to find out 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 indicates 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 information perceived via the auditory channel more and more, as suggested by the MLE model. Yet most notably, according to their analysis, test subjects seemed to be biased to use visual information to a greater extent than 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 (and thus overestimating the use of auditory information).

Shams et al. report experiments in which visual illusion was induced by sound, thus resulting in the auditory cue outweighing the visual cue [sha00, sha02]. They presented test subjects with flashes of light and beeps of sound: when 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 due to cross-modal perceptual interactions rather than 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 provoked the human perceptual system to rely on other modalities.

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. tested subjects for contradicting localization cues and were presented with a bias toward the visual percept, Shams et al. tested subjects for temporal variations of cues and were presented with a bias toward the auditory percept. This might indicate 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 3 to 2 degrees for sinusoidal signals coming from a forward direction [zoz98, ter98, zwi99], the resolution of the visual system is at least 100 times as high, about 1 min of arc [how82]. On the other hand, the time resolution of the auditory system allows to resolve the temporal structure[6] of sounds as close as $2ms$ [zoz98], whereas the human visual system can be intrigued to belief in a continuously moving object when presented with only 24 sampled pictures per second of the continuous movement.

2.4.2 Ventriloquism

Ventriloquism is the art of making one's voice appear to come from elsewhere. Its best known implementation today is probably the puppet player engaged with his hand puppet in a discussion about some trivia. In ancient times the effect of ventriloquism was exploited by the Roman and Greek oracles [con00]. Its working mechanism is based on the fact that we expect spoken words to be uttered by a speaker moving his lips. When the speaker does not move the lips, then we search for some alternative sound source: in the case of the hand puppet we gladly accept the moving lips of the puppet as a sound source. In the case of the oracle, people accepted the opening in a stone, though static, as the source of the spoken words.

But even in everyday life we regularly experience the effect when watching TV and movies. The voices seem to emanate from the actors' lips rather than from the actual sound source (the loudspeakers). As it seems, the ventriloquist effect is a clear case of visual capture as defined in section 2.4.1.

Alais and Burr [ala04] have shown that the ventriloquist effect cannot be explained fully by the visual capture model. They have conducted a number of experiments similar to those by Shams et al. [sha00, sha02], where test subjects were presented with flashes of light ("blobs") and beeps of sound ("clicks"). Whenever visual localization was good, vision indeed dominated and captured sound. For severely blurred visual stimuli, however, the reverse held: sound captured vision. For less blurred stimuli, neither sense dominated and perception followed the mean position. Additionally, precision of bimodal localization was usually better than in the single (auditory or visual) unimodal presentation – this interesting observation is further discussed in section 2.4.5. Alais and Burr conclude that their results are "well explained not by one sense capturing the other, but by a simple model of optimal combination of visual and auditory information" [ala04].

Kono et al. have also evaluated visual image effects on sound localization [kon02]. They

[6] This resolution relates to monaural hearing and is not to be confused with binaural resolution as e.g. used in localization tasks.

point out that in contrast to most other studies approaching the area of audiovisual interactions from a psychological viewpoint, their aim was to approach those interactions for the purpose of an application to actual audiovisual environments. Furthermore, they concentrated on simultaneous audiovisual percepts presented in the peripheral visual field, a perceptual situation very common in everyday real-life situations. Test subjects were presented with a movie of a moving police patrol car and its siren sound. Kono et al. conclude that visual capture is stronger in the beginning of the presentation than in the end, an observation they call the "beginning effect". This suggests that visual capture is also depending on the temporal context of the percepts, and that – regarding visual capture as a special case of the MLE model – audiovisual perception in general might be time dependent.

McGurk and MacDonald in the mid-70s published their findings on "hearing lips and seeing voices", another prominent example of auditory illusion induced by visual percept [mcg76]. The so-called McGurk effect e.g. describes the situation in which an auditory $//ba//$ is paired with a visible $//ga//$, which is perceived as an auditory $//da//$. The reverse pairing, a visual $//ba//$ and an auditory $//ga//$, tends to produce a perceptual judgment of $//bga//$. The visual input obviously changes the auditory experience, not only with respect to localization, but also to perceived content.

2.4.3 Synchrony

Synchrony between multi-modal percepts is something given in natural environments. Apart from a few exceptions like the thunder arriving significantly later at the perceiver than the flash (due to the limited traveling speed of sound and varying with distance), we are used to congruent stimuli. Un-synchrony of multi-modal stimuli is therefore rather a technical problem than a perceptual one. Yet, as we are going to evaluate human perception in interactive application systems (as the title of this thesis suggests), it is well worth looking at synchrony from a perceptual point of view.

In the field of perceptual psychology, interaction between the aural and visual modalities is well documented. A rather large number of singular effects have been scientifically researched. Because in these experiments often extremely simple auditory and visual stimuli are used, it is often hard to extrapolate the results for more complex media applications. Therefore it is interesting to look at the field of film music investigations. A number of models for film music perception have been proposed. Lipscomb and Kendall e.g. suggest that there are at least two implicit judgments made during the perceptual processing of the movie experience: an association judgment and a mapping of accent structures [lip99]. According to their model, the mapping of accent structure is determined by the consistency with which important events in the musical score coincide with important events in the visual scene - the synchrony between the two modalities. When test subjects were presented with extremely simple auditory and visual stimuli, accent alignment played an important role in the determination of ratings of effectiveness. As the stimuli became more complex, the importance of accent structure alignment appeared to diminish and the association judgment assumed a dominant role. This means that apparently synchrony loses its importance with growing complexity of the audiovisual stimuli.

Alais and Burr have looked at the "flash-lag" effect (FLE) cross-modally [ala03]. The FLE is easiest demonstrated on a computer screen showing two spots of light, one translating across and the other briefly flashed in vertical alignment with the first. Despite being physically aligned, the brief flash is perceived to lag behind the moving spot. Alais and

Burr have shown that also briefly presented auditory stimuli lag behind continuous auditory movement, no matter whether the auditory movement is represented by a frequency sweep or by a translating sound source. They also claim that when spatial motion is used, the FLE can occur cross-modally. Interestingly, the implied temporal lags for the cross-modal case are reported to be smaller than what is observed for the unimodal auditory version. Alais and Burr conclude that there might be a discrepancy between internal brain timing and external stimulus timing [ala03].

The FLE can also offer an explanation for an observation made during the pilot tests to one of the subjective assessments performed in the course of working on this thesis: in an experiment done to determine the optimum number of loudspeakers to be used in interactive application systems of moderate complexity presented on large screens, test subjects claimed that visual and aural stimuli were out of sync when these performed very fast rotational movements. In these situations the virtual loudspeaker showed up only for a very short period of time and then left the field of view / the large projection screen again, while the sound coming from that virtual loudspeaker was panned continuously in a circular array of real loudspeakers surrounding the test subject. For a detailed description of the experiment itself see section 8.7.

Wenzel has published a number of papers on the topic of system latency, see e.g. [wen98, wen99, wen01], the most prominent incarnation of potential un-synchrony in real time audiovisual application systems. She defines the total system latency (TLS), or end-to-end latency, as "the time elapsed from the transduction of an event or action, such as movement of the head, until the consequences of that action cause the equivalent change in the virtual sound source location" [wen98]. In the assessments performed, Wenzel found out that latency had to be at least $250ms$ to be readily perceived by the test subjects. Also, the accuracy with which test subjects performed a localization task was generally comparable for the shortest ($34ms$) and longest ($500ms$) latencies tested. This suggests that listeners are able to ignore latency during active localization. In audiovisual systems of more than basic complexity, where e.g. also room acoustic rendering plays a role, latency is not solely related to changes in sound source location, but also to other (room) acoustic parameters. As Wenzel points out, it "may be that the localization task per se is not the most sensitive test of the impact of latency in a virtual audio system" [wen01]. Latency with respect to interactivity issues is further discussed in section 4.1.1.

In technical systems used for reproducing audiovisual content it is considered mandatory to play back synchronized auditory and visual stimuli. Interestingly, the detection thresholds for un-synchrony are not temporally symmetrical. Hollier and Rimell have performed a number of experiments with a focus on audiovisual communications systems to examine this temporal asymmetry with different types of stimuli [hol98, hol99]. They compared a talking head audiovisual scene with a bouncing pen scene and an audiovisual stimulus in which an ax hit an object one single time. They conclude that the general trend in error detection asymmetry is apparent for all stimulus types. Furthermore, the distinctness of the ax stimulus results in greater probability of detection than for the pen stimulus. For the talking head stimulus the error detection rate is consistent with the other stimuli when audio lags behind the video, but greater than either ax or pen stimuli when the audio leads the video. Apparently, test subjects compared the artificial stimuli presented in the lab with real life experiences. In real life, due to the physical nature of different traveling speeds of sound and light, audio can never lead the visual percept.

Hollier, Rimell et al. in [hol99] conclude that the perception of synchrony is content de-

pendent and suggest a very rough draft of a synchronization perceptual model. The model tries to combine objective descriptions of audio and video events with rules concerning the perception of synchronization errors in order to provide an objective prediction of the perceived error. The audio descriptors suggested are a sliding $20ms$ RMS energy, as well as signal peak and decay parameters. The video descriptors consist of motion vectors and a persistence parameter specifying the subjective importance and its decay over time. Yet, it remains unclear how the subjective importance might be determined.

The findings on synchronization error detection asymmetry are also reflected in the recommended synchronization thresholds given in ITU-T J.100, which is $20ms$ for audio lead and $40ms$ for audio lag [itu100]. The recommendation suggests these fixed figures for all types of television content and is intended to ensure that synchronization errors remain imperceptible for all possible varieties of content.

This relatively small threshold means that the human perceptual system is generally quite sensitive to errors in synchrony, a capacity we can observe ourselves when watching sluggishly synchronized movies or shows on TV. Errors in lip-synchronization can easily be detected, especially when audio leads the visual percept. This observation contrasts with the film music perception model by Lipscomb and Kendall, and it puts into perspective the findings of Wenzel. Apparently, the perception of synchrony is strongly context-sensitive.

2.4.4 Impact on Perceived Quality

The influence of cross-modal interaction on perceived audio or visual quality has been researched in a number of studies. Rimell et al. as well as Hollier and Voelcker of British Telecom Laboratories in Ipswich / Suffolk have conducted experiments in the area of speech quality perception [rim98, hol97]. They have tested the hypothesis that the audio and video content of multimedia presentations are highly interrelated. They also hypothesized that cross-modal interaction is greater for "talking head" material than for other video sources, based on the fact that humans pay great attention to the eyes and mouth of a talker as part of normal one-to-one communication. Both hypotheses tested positive. Rimell et al. conclude that the "quality of one mode affects the perceived quality of the other mode and a single mode should not be considered in isolation" [rim98].

Zielinski et al. have reported about the division of attention between the evaluation of audio quality and involvement in a visual task. In their experiment, the visual task consisted in playing a computer game. While playing, test subjects were requested to rate the audio quality of the game's background music. The quality deterioration consisted in applying low pass filters to the audio material with cut-off frequencies between $16kHz$ and $13kHz$. The main research question was formulated as "To what extent does involvement in a game change the results of evaluation of audio quality?" [zie03]. They report that involvement in a visual task may significantly change the audio quality evaluation results obtained. Yet, the effect was found to be subject-specific and rather small but significant when averaging the results across all participants.

In the same context, Kassier et al. have taken the experiment described by Zielinski et al. further to include time-variant changes (drop-outs) in the audio quality [kas03]. They compared two situations, one in which test subjects passively watched a static picture, and another one in which subjects played a computer game. They report that involvement in the computer game had a very small but significant overall effect (+ 3%) in the rating of the audio quality. Yet, as already seen in the preceding experiment, the effect was highly dependent on the test subject itself (subject-specific). They also report that active

involvement in the visual task decreased the consistency of the audio quality rating.

The rather unclear outcome of the experiment might be due to the low number of participants. Kassier et al. themselves state that "partial violation of the second ANOVA assumption indicates the need to test larger groups of subjects in future experiments" [kas03].

2.4.5 Superadditivity

Superadditivity occurs when the perception accuracy with two or more sources of information is greater than predicted by the sum of accuracy measures for the individual sources [ber96]. The main research area where superadditivity has been studied is speech perception. Bernstein and Benoît give an overview of speech perception in [ber96], presenting the main affirmative proposition in the title of that paper: "For speech perception by humans or machines, three senses are better than one". As an example they describe an experiment in which the face of a talker is combined with an acoustical signal presenting the talker's voice fundamental frequency ($F0$). Whereas by itself the fundamental frequency is not intelligible, when it is combined with visible speech performance it is typically enhanced by roughly 20-40 %. In contrast, the average performance for visual speech by adults with hearing is only approximately 20-30 % words correct in sentences.

It can be expected that superadditivity also plays a role in other perceptual tasks involving stimuli presented in more than one modality.

2.4.6 Motion Sickness

Motion sickness or *kinetosis* is usually provoked by contradictions in the multi-modal percepts. These contradictions can stem from an agitation of the fluid in the semicircular canals of the inner ear, while the visual percept communicates an apparent lack of movement. Typically, motion sickness occurs when the perceived motion is not self-controlled - the driver of a car never suffers from motion sickness, whereas reading a book in the lap during a car drive increases the danger of motion sickness. Motion sickness can provoke symptoms like fatigue, vertigo, nausea and headache.

Very closely related to motion sickness is simulator sickness, which is usually provoked by visual percepts indicating self-movement, while the inner ear does not perceive any acceleration. Whereas in real-world situations motion sickness can be avoided by solving the input conflict (e.g. by looking out of the window or closing the eyes), the input conflict in simulators can usually not be solved by the subject itself. It is therefore most important to watch test subjects in audiovisual subjective assessments such as those described in section 8, because the simulator sickness threshold can vary greatly between individuals.

2.5 Attention in Human Perception

When being confronted with an increased load 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. All of them are usually referred to as *attention* [pas99].

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. 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?

2.5.1 Perception of Multiple Streams

Eijkman and Vendrik conducted one of the earliest studies on the perception of bimodal stimuli in 1965. 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 [pas99]. Other studies, e.g. by Shiffrin and Grantham (1974) and by Gescheider, Sager and Ruffalo (1975), also support these results for presentations of short bimodal stimuli [pas99].

The stimuli presented in the auditory and the visual modality were not contextually related in the study of Eijkman and Vendrik, so they constituted what could be called separate perceptual streams. Yet, detection of increments in the duration of the same stimuli were 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. [sha00, sha02] already discussed in subsection 2.4.1.

On the other hand, other studies combining auditory and visual discrimination tasks observed modest but considerable decrements in terms of performance, when test subjects were confronted with bimodal stimuli in comparison to unimodal ones. To give an example, Tulving and Lindsay in 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 [pas99]. All of these studies differ from others in that they involve magnitude judgments rather than categorical judgments. Actually, the performance of test subjects in the bimodal case might have been limited by the difficulty in maintaining a standard in memory against which to judge the inputs, rather than by the influence of a second modality itself.

2.5.2 The Perceptual Cycle

Neisser's model of the *Perceptual Cycle* describes perception as a setup of *schemata*, *perceptual exploration* and *stimulus environment*. These elements influence each other in a continuously updated circular process, see fig. 2.27. Thus, Neisser's model describes 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 the knowledge about our environment. They are based on previous experiences and are located in the long term memory. Neisser attributes them to generate 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 [keb94, gol02, far03].

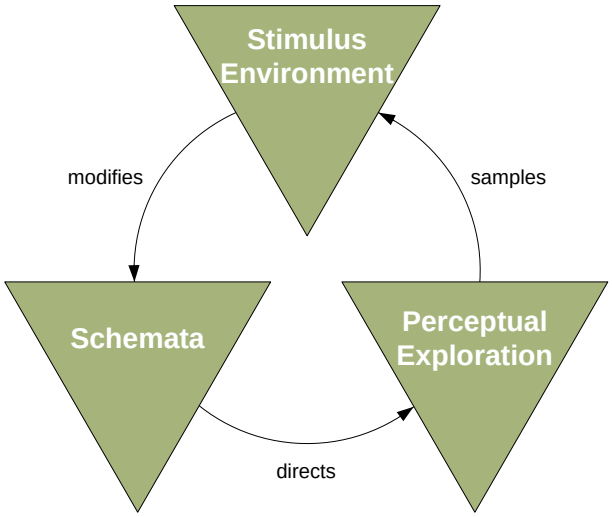


Fig. 2.27: Neisser's Perceptual Cycle, *after [far03], modified.*

The differences in schemata present in the human individual cause the same stimulus to provoke different reactions between subjects. Following Neisser's model, especially new experiences (those that cause a modification of existing schemata) are likely to generate a higher load in terms of processing requirements for the percepts.

2.5.3 Selective Attention

The 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.

An unmanageable number of studies have tried to identify and describe the strategy that is actually used in the human perceptual process. Pashler gives an overview and identifies two main concepts of attention [pas99]: 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.

Fig. 2.28 shows an example of a perceptual model originally developed by Shiffrin and Atkinson that is based on exclusion [mur73]. Ineffective stimuli are those that do not pass the selective attention process because they are considered irrelevant.

The second concept construes the limited processing resource (rather than the filtering device) as attention. It suggests that when attention is given to an item, it is perceptually analyzed. When attention is allocated to several items, they are processed in parallel until the capacity limits are exceeded. In that case, processing becomes less efficient. A perceptual model based on resource allocation could look similar to Shiffrin and Atkinson's (fig. 2.28), but with all potential stimuli being effective and a larger amount of lost input due to capacity constraints.

Neither of the two concepts can be ruled out by the many investigations performed up to now. Instead, assuming either the gating or the resource interpretation, all empirical results can be accounted for in some way or other. As a result, it must be concluded that both capacity limits and perceptual gating characterize human perceptual processing. This

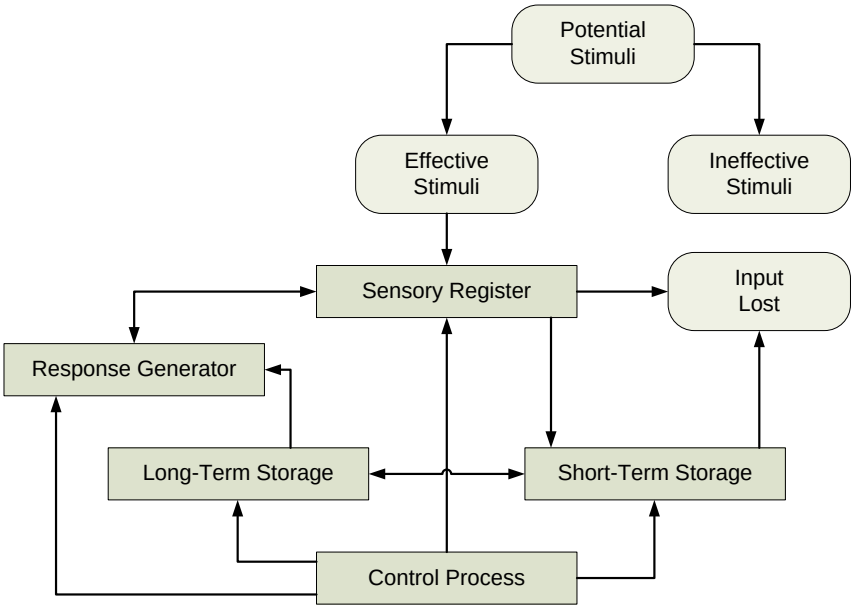


Fig. 2.28: Shiffrin and Atkinson's model of the perceptual process. *After [mur73].*

combined concept is termed *controlled parallel processing* (CPP). It claims that parallel processing of different objects is achievable but optional. At the same time, also 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 if 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 that are associated with different sensory modalities, or whether a unified multi-modal attention system exists. Are visual and auditory attention the same thing? Investigations have shown that humans are capable of selecting visual stimuli in one location in space and auditory stimuli in another [pas99].

Spence et al. have examined the effect of expecting a stimulus in a certain modality upon human performance [spe01]. 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 vs. a 75% majority of targets located in one modality). They report that reaction times for targets in the unexpected modalities were slower than for the expected modality or no expectancy at all. They further note that shifting attention from the tactile modality was taking longer than shifting from the auditory or visual modality. The results show that performance depends not only on what actually happens, but also on what is anticipated. Yet, they note that a faster response time for the most likely modality in their study 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 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 [ala99]. 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" (compare section 2.3.4).

Treismann and Gelade suggest, based on a significant number of experiments, that

"Visual attention, like a spotlight or a zoom lens, can be used over a small area with high resolution or spread over a wider area with some loss of detail [...] attention can either be narrowed to focus on a single feature, when we need to see what other features are present and form an object, or distributed over a whole group of items which share a relevant feature" [tre80].

These findings are a clear indicator of attention actually controlling the perceptual process. They can not clearly answer the question whether there is one multi-modal attention or whether attentions are associated with modalities. There are, however, indicators that favor the latter.

2.5.4 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 [pas99]. 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. Sections 8.9, 8.10 and 8.11 present experiments that contribute to the verification of this claim.

In an overview article, Lavie examines the capacity limits in selective attention [lav01]. 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, gating of stimuli takes place such that 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 specifically on perceptual capacity limits rather than on limited information contained in the relevant stimuli. This makes the maximum-likelihood estimation model described in section 2.4.1 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.

Lavie concludes that "capacity limits play a major role in selective attention, and that even purely psychological concepts of capacity can lead to clear and testable predictions for neural activity in the brain when considered in terms of a perceptual load model" [lav01]. Such a load model has been introduced by Lavie earlier [lav95], see fig. 2.29, but unfortunately is not quantifiable yet. Its usefulness for technical applications is therefore still very limited.

2.5.5 From Attention to Perceptual Quality - the BTL Perceptual Model

Hollier and Voelcker from British Telecom Labs have presented what can be regarded as a first step toward a multi-modal perceptual model in an application-oriented context

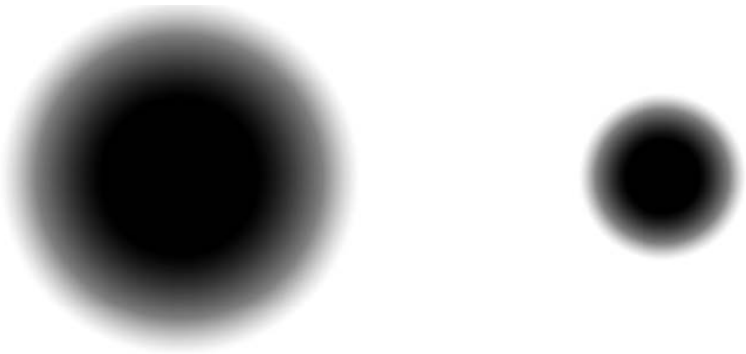


Fig. 2.29: Lavie's perceptual load model interpreted for visual perception. *Left*: Hypothesized size and shape of visual attention in a low central task load condition. Attention is broadly distributed over the visual field. *Right*: Hypothesized size and shape of visual attention in a high central task load condition. Attention is constricted toward the central visual field of fixation. *From [pak05].*

[hol97a]. The proposed multisensory perceptual model builds upon the concept of sensory and perceptual layers for the individual senses and introduces cross-modal dependencies and the influence of task. The three main components of the model consist in the sensory models, the cross-modal model, and the scenario-specific task model. Fig. 2.30 shows a diagram of the proposed model.

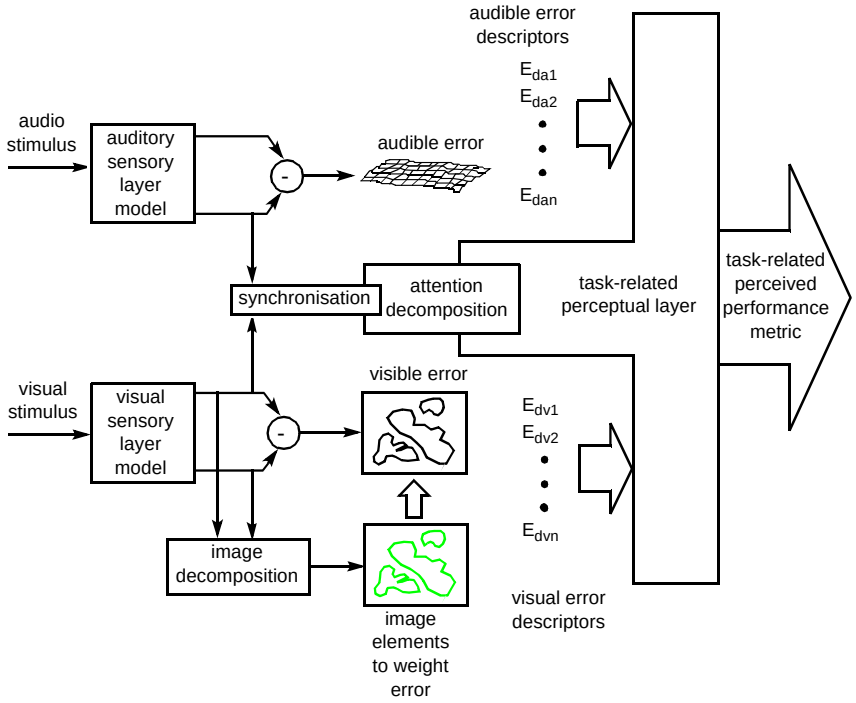


Fig. 2.30: The BTL multisensory perceptual model proposed by Hollier and Voelcker. *From [hol97a].*

The model does not consider the full extent of human perception and cognition, but rather tries to establish and model the gross underlying low-level dependencies between the modalities. Auditory and visual sensory layers are modeled separately using well established sensory layer models known from perceptual coding. According to Hollier and

Voelcker, the resulting unimodal error descriptors can be designed to map to different subjective opinion scales, thus enabling them to predict overall opinions. They are subsequently weighted according to four factors:

- high-level cognitive preconceptions associated with the task
- attention split
- degree of stress introduced by the task
- experience of the user.

Additionally, auditory and visual stimuli are compared cross-modally for timing and quality balance aspects between the modalities.

Hollier and Voelcker argue that multi-modal applications, by their very nature, are more complex and varied than unimodal ones (like e.g. a simple telephone call). Hence, the influence of the task increases in such applications. Yet, it remains unclear how task could be categorized and how its influence could be measured. Apparently, they recognize this as a major problem in applying their model when saying that "it is likely that the definition of task for the model will tend to be more towards the coarse end of the spectrum" [hol99].

2.6 Preliminary Posit

The mechanisms of human multi-modal perception are, until now, not very well understood. Approaches in several fields of research such as neurophysiology, psychology, natural sciences in general, and engineering have been undertaken to explore these. As a result of these efforts, a lot of knowledge on specific aspects of perception exists. Because of the sheer complexity of the involved building blocks and the many factors that potentially influence the processing, a *unified* model of the perceptual process does not exist.

Depending on the focus and background from which the topic is addressed, a number of models (e.g. Neisser, Shiffrin and Atkinson, Hollier and Voelcker) have been suggested on different levels of abstraction. These try to identify the basic elements involved in perception, as well as their interrelationship.

The main problem with these models is that they are difficult or impossible to quantify. A perceived overall quality metric is hard to implement because of the abstract concepts behind these models. The most promising approach in terms of usefulness for actual technical implementations seems to be the one by Hollier and Voelcker. This model originates from the field of transmission of traditional audiovisual content (video with accompanying sound), i.e. neither effects of interactivity nor real-time rendering problems and the related specific quality attributes are addressed.

To make such a model applicable to the field of interactive audiovisual application systems, key influence factors have to be identified and, if possible, their impact has to be categorized or even quantified. This is, at the current state of art, only possible by performing subjective assessments among test subjects. The system used for performing the tests should resemble the actual application and its features as close as possible.

3 Computational Room Acoustics and Auralization

In rooms, sound not only travels on a direct path from the sound source to the listener's ear. Instead, sound is additionally reflected at the room's walls and at larger objects in the room, such that so-called indirect sound paths are created. For large reflectors (walls or objects) with plain surfaces, geometrical reflection can be assumed, and the frequency characteristics of the reflected sound is influenced by the material characteristics of the reflector. The density of these reflections builds up over time, while the acoustic energy contained within each sound path is diminished with each reflection (and also due to distance-dependent air absorption) along the sound path.

Real sound sources usually radiate sound not only into one direction. They have a certain frequency-dependent radiation pattern that determines amplitude- and frequency-wise how sound emanates into different directions. Real reflectors usually are neither necessarily plain nor smooth; objects can be small, jagged, consisting of arbitrarily distributed surface materials, and so on. They may cause scattering and diffraction effects, such that sound paths tend to get "smeared" with time. This results in a diffuse sound field that, along with the direct sound and the early reflections, creates a very complex acoustic impression of a room.

Looking back at the history of acoustics, one easily notices that the basic idea of describing the behavior of sound waves propagating by means of mathematical equations is quite an old one. Already in the sixth century B.C. Pythagoras described that halving the length of a string obviously doubled its pitch. This might be regarded as the first attempt in describing mathematically an acoustic wave propagation. As simple as the description of the phenomenon reported by Pythagoras might seem to be, the solution of the wave equation not on a string but in a room of arbitrary shape can be arbitrarily complicated. This "brute force" approach is not only a very challenging one in terms of mathematical skill necessary, but it also has the disadvantage of being much too exact for the usual needs of an acoustician: the amount and quality of input data (e.g. the description and characteristics of walls and objects in the room) as well as of output data can not be handled with reasonable effort.

Therefore a number of alternative ways to compute a room's acoustic behavior have been developed. Most of these methods describe the interaction of (static) sound source, (static) room and (static) sound receiver as a linear time invariant (LTI) system[1]. According to systems theory, such an LTI system is characterized completely by its impulse response (IR). The Room Impulse Response (RIR) is defined as the sound pressure in a room that results from an excitation of infinitesimal duration and infinite power, the Dirac impulse δ. The RIR is usually measured as a function of the time after the arrival of the first direct sound from the source to a listener's location. The interaction of sound source and walls results in reflections contributing to the overall acoustic room impression contained in the RIR. The RIR also contains information on the distribution of sound energy over frequency. It needs to be transformed into the frequency domain to make this information accessible. A specific RIR is valid only for the one position in the room it has been

[1] "Static" in this context means that neither position / location nor other characteristics of source, medium, room, or receiver are modified in the course of computation.

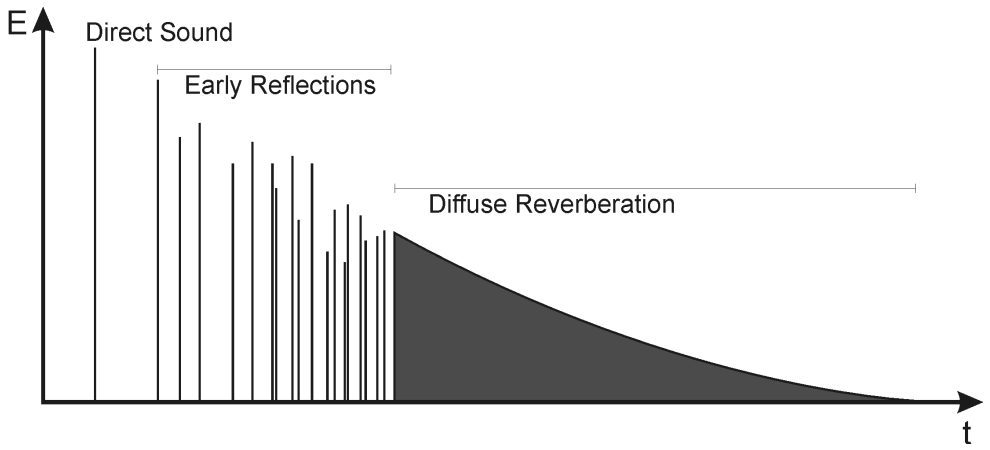


Fig. 3.1: Schematic energy view of a room impulse response (RIR), subdivided into three sections: direct sound, early reflections, and diffuse reverberation.

measured or calculated at. Yet, a general statement on a room's acoustic behavior, like e.g. its reverberation time T_{60}, can be derived from it.

For an analysis, the RIR can be divided into a number of temporal sections along its abscissa, see fig. 3.1. The direct sound arrives first, its temporal delay being directly connected with the distance between source and receiver and the speed of sound c in the medium (air). Then the early reflections (ER) follow, usually within $10 - 100ms$ after the direct sound. Very early reflections (within the first $10ms$ after the direct sound) tend to change the timbre of the direct sound. Reflections coming in between the first $10 - 50ms$ after the direct sound contribute to the subjective loudness level of the signal [zoz98]. The directional impression at a certain receiver position in the room is usually determined by single reflections, which have directions of incidence clearly perceivable. Apart from the discrete reflections there can also be some diffuse or dispersed portions of signal in the later early reflections part of the RIR. These can be broadened and therefore overlap to form an early diffuse part which is statistically distributed over all directions of incidence. This deformation of the original impulse is due to the frequency dependent characteristics of the sound source, the walls, and the sound propagating medium [hei94].

The early reflections' temporal density grows continuously over time, with each reflection being reflected over and over again. Whereas the reflections' density grows, the amplitude of the RIR becomes smaller, because each reflection of the sound diminishes the energy contained within. This section of the RIR is the diffuse reverberation (DR) part. Since Moorer it is known that this can be approximated as a random signal (white noise) of exponentially falling sound pressure (or linear falling sound pressure level) [moo79]. Here, a constant and even distribution of sound energy over the room's volume is assumed. Therefore, the laws of statistic reverberation theory can be applied [cre78, kut91]. Subsection 3.2 shows that the concept of dividing the RIR into different temporal sections represents a major advantage when it comes to real-time room acoustic simulation.

Although visual cues are a predominant part of our perception, they alone are not sufficient to create believable artificial worlds. In analogy to the process of visual rendering of a virtual world called *visualization*, the term *auralization* is used to describe the process of creating a sound impression in conformance with the virtual world. As well as visualization in this context is three-dimensional, auralization is the simulation and reproduction of the three-dimensional sound present at the listener's location. In the 1993 AES journal's

November edition, Kleiner proposes a definition of the term more universally valid:

> "Auralization is the process of rendering audible, by physical or mathematical modeling, the sound field of a source in space, in such a way as to simulate the binaural listening experience at a given position in the modeled space" [kle93].

Another definition commonly used describes auralization as the reproduction of an acoustic surrounding or space A in a different acoustic surrounding B. Therefore, auralization can provide an impression of how an acoustic signal like music or speech would sound if replayed in a room different from the one the listener is actually located in. In order to achieve this, the acoustic properties of surrounding A need to be measured or modeled (e.g. in the form of a RIR as described above), applied in some way to the acoustic signal (e.g. by convolving the audio signal with the RIR), and subsequently rendered audible in surrounding B. Of course, surrounding B may consist of a set of headphones, so this second definition directly matches the one given by Kleiner. Yet, without violating the definition proposed, one can also imagine an auralization system reproducing the sound of the modeled space via a setup of loudspeakers placed in a real room.

In fact, a large number of systems have been proposed for the auralization of rooms, see section 3.3. These are either based on headphones or on loudspeaker setups comprising different numbers and positions of speakers. They all have in common that they try to imitate the original incident sound of direct and indirect sound paths as closely as possible. This means that the complex system of sound source(s) with frequency- and amplitude-dependent radiation patterns, reflectors of different materials, shape, size, and distances that is found in the original situation is reproduced by sound emitters that may have completely different characteristics than the original sound sources, are located at different positions than the original ones, and are situated in a room with acoustical characteristics of its own.

Reproduction via headphones almost always involves a convolution of the reverberated sound with Head Related Transfer Functions (HRTF) to attach a direction of incidence to the signal. If the listener's head is not kept in a fixed position, head tracking is necessary to avoid in-head localization when the listener performs rotational head movements. Reproduction via loudspeakers can be based on different panning laws and needs a 3D setup of speakers if elevated sound sources or non-vertical reflectors (like floor and ceiling) are to be considered. Both approaches can render valid results. Whereas a loudspeaker-based system increases in complexity with each additional loudspeaker signal to be created, HRTF-based systems require listeners to wear a device on their head (a headphone with head tracker). Because the HRTF-convolved signal is related to the current head movement of the listener, such systems can only be used for single-user applications.

3.1 Exact Simulation vs. Real-Time Rendering

The traditional field for the application of room acoustic simulation is the room acoustic design process for architectural purposes, like the acoustic design optimization of concert halls, lecturing rooms, and more generally spaces where the room acoustic properties play an important role. There is a large number of tools available to compute the RIR of such rooms. Their complexity ranges from easy-to-use freeware tools limited to the computation of the first early reflections in a shoe box shaped (rectangular) room to full-featured software suites like *CATT-Acoustic* [w-cat], *ODEON* [w-ode] or *EASE* [w-ada].

These allow to simulate a near arbitrarily complex room design combined with features for the auralization of the room's "predicted" sound. The latter software even provides a database for different brands and models of loudspeakers commonly used in public address (PA) situations. Therefore, not only the room's characteristics can be rendered audible, but also the sound source's properties like frequency dependent angle of radiation can be taken into consideration in the simulation.

In these applications, there is a clear focus on the achieved realism of the simulation. It is therefore performed with the highest accuracy possible, normally for a number of typical listener positions in the room. Yet, each simulation pass is independent from the others: the calculation of different RIRs is limited to static situations in the room. Movement of source(s) or receiver are not considered, as well as there are no changes in the geometry of the room or the walls' acoustic properties during simulation time.

In interactive audiovisual application systems, the possibility for the user to influence the virtual scene (a more detailed definition of the term *interactivity* is given in section 4.1) is one of the key functionalities offered. Therefore, a transition from a static to a dynamic system is necessary for the description of the scene. All dynamic systems are inherently time variant. Thus the description of an interactive environment in terms of the room impulse response is, strictly speaking, not valid any more.

Apart from that, there are a number of other factors that influence the design decisions necessary to implement such a room acoustic simulation environment or, more generally speaking, an auditory virtual environment. First of all, the computation has to be performed in real-time for interactive applications. Therefore the degree of room acoustic detail that can be considered in the simulation is limited by the computing power available. Whereas the classic approach is based on the physical characteristics of the room (like geometry, material of walls, etc.), the *Perceptual Approach* emanates from the human perception of spatial audio and room acoustical quality[2]. If the simulation is based on the physical approach, then there are a number of different simulation methods available, each with specific pros and cons which have to be weighed against each other. These will be described shortly in the following subsection.

3.2 Rendering Methods

As explained in section 3, the early reflections in the room can be calculated separately from the diffuse reverberation (see fig. 3.1). This is very advantageous, because in the ER part most information on the room is already contained. As it is the most critical part of a room's impulse response in terms of quality and fidelity of the simulation, most of the computing power available is usually invested here. For the case of interactive audiovisual application systems, the question of how to distribute computing power available is discussed as one of the central questions of this work in chapter 7. The question of how to model the room acoustic impression is crucial for the quality that can be achieved under certain conditions of real-time acoustic rendering. Therefore the most prominent methods are discussed in the following.

There are a number of different methods known for the computation of the ER part. These different methods can be regarded as components of the overall room simulation model and thus are interchangeable. Fig. 3.2 shows how ER and DR parts can be used to jointly form a complete room acoustic description in a signal processing schematic. The

[2]The *Perceptual Approach* is described in section 5.6.

DR calculation can either be fed solely by the original dry sound or additionally by the ER calculation output to make the DR part more dense.

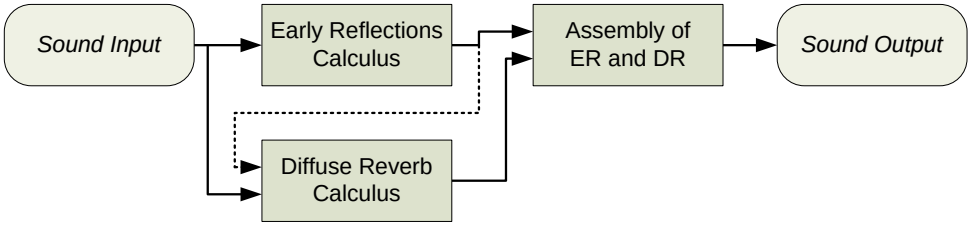


Fig. 3.2: Schematic view on Early Reflections (ER) and Diffuse Reverberation (DR) parts, regarded as components in a signal processing schematic.

3.2.1 Image Source Method

The most popular method for the computation of early reflections is the image source method. The basic idea is to substitute each reflection of the sound source at each wall with a so-called *mirror* or *image source*. Fig. 3.3 shows the direct sound path r_1 from sound source S to receiver R. The reflection of sound source S at the wall W generates a secondary sound path r_2. Because sound propagates at finite speed ($c \approx 340m/s$ at $20°C$), and sound path r_1 is shorter than r_2, the direct sound will arrive at the receiver before the reflected sound. The receiver does not know that both sounds originally stem from the same source, so the reflected sound path r_2 can be substituted by a secondary sound source S'. Its position is defined by the (geometrical) reflection of the original source S at the wall W. Subsequently, the wall W can be disregarded. Because both sound paths r_2 originating from S and S' have the same length, no further delay processing has to be performed. Of course, the wall's acoustic characteristics like (frequency dependent) absorption α have to be accounted for. Therefore, the new image source S' is assigned the acoustic power of the original source S minus the wall's absorption. The new sound source S' is called *image source of first order*, because its original sound path contained *one* reflection at a wall.

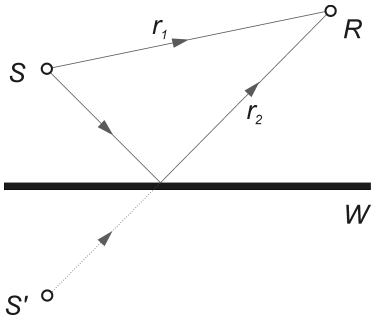


Fig. 3.3: Basic principle of image source method.

For sound sources located in closed rooms, a sound path may be reflected more than once, see fig. 3.4. Thus, image sources of higher order come into existing. They stem from the reflection of a (image) source already reflected at another wall. Furthermore, each wall contributes to the number of image sources to be constructed. Therefore, the more complicated the room's geometry, the more image sources have to be considered.

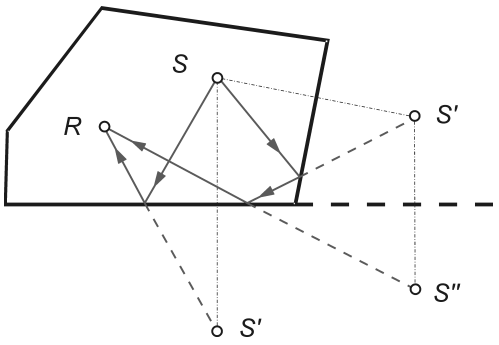


Fig. 3.4: Sound source S in a closed room generating image sources of first (S') and second (S'') order.

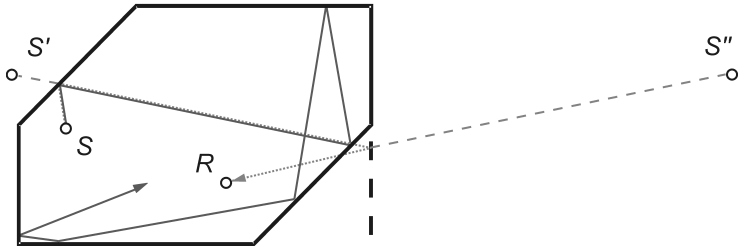


Fig. 3.5: No contribution of image source S'' at receiver position R, though S'' can be formally constructed.

In general, not all formally existing image sources are visible for all receiver positions. Although image sources can be constructed up to a given quantity (or order), their contribution may not be relevant for each and every receiver position in the room. Therefore the simulation has to verify whether there is an existing sound path going from the original source via the walls' reflections to the receiver position. Only if this condition is met the image source is counted as being "visible" for the receiver position under examination. Actually, for rooms of irregular geometry this "visibility check"[3] is the most time consuming part of the simulation of early reflections via the image source method. The visibility problem is displayed in fig. 3.5. Whereas both image sources S' and S'' formally exist, S'' is not visible from the receiver position R and therefore does not contribute. The dotted line represents the formal sound path after constructing the image source of second order S'', the solid line is the actual sound path.

As the computational complexity is growing exponentially with the order of image sources to be computed, it is not efficient to use the image source method for the generation of (late) reverberation tails. Also, the image source concept is not a really good model for more than a low number of reflections, because diffraction and surface scattering very soon dominate the specular reflections in real rooms.

3.2.2 Ray Tracing Method

The ray tracing method proceeds from the assumption that a sound source S sends out numerous sound particles with a given distribution of angular deviation at the time t. The

[3]The visibility check is sometimes also called "audibility test" [sch06a] to make clear that the propagation of sound is under examination. Yet, using the term "visibility check" is more precise because effects of diffraction are not taken into account here.

propagation paths of these sound particles are traced along all the reflections occurring at the walls, see fig. 3.6. With each bounce against a wall the location of the collision, as well as the weakening of the particle's energy by means of absorption, and the direction of the reflection are calculated. Therefore even diffuse reflections at walls of rough surfaces can be considered. Finally, the incoming sound particles have to be counted at the receiver position. This is done by using so-called *counting volumes* of spherical or cubic shape. Whenever a sound particle crosses the counting volume, its energy, direction of incidence and distance covered (resulting in delay time) are determined. Therefore the ray tracing method can be split algorithmically into three parts: radiation, propagation and reception of sound.

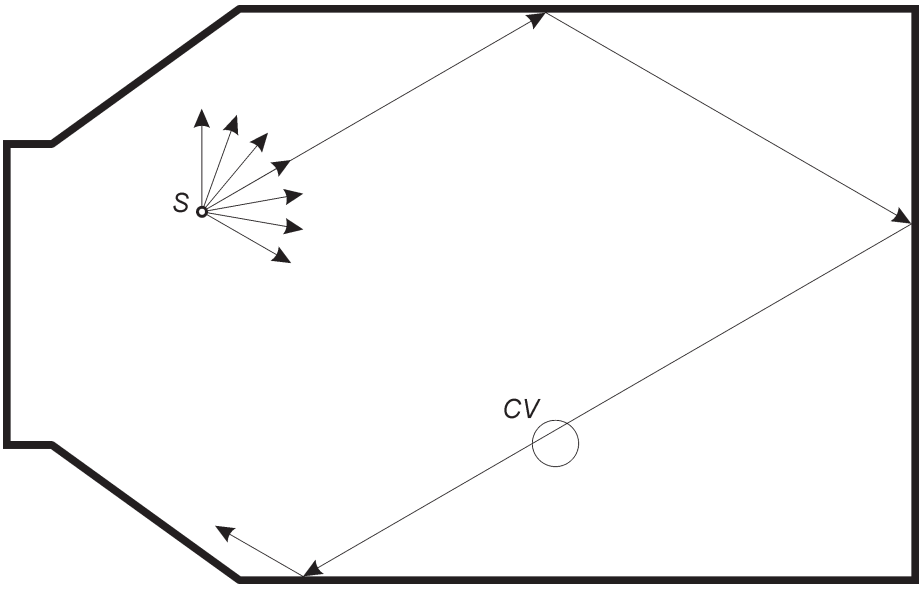


Fig. 3.6: The ray tracing method: sound source S radiates sound particles, trace of sound path, and reception of particle in counting volume CV.

For practical reasons, the radiation of the sound particles is usually described in spherical coordinates. Fig. 3.7 shows the first octant of the sound source with azimuth angle φ of range $[0, 2\pi]$ and polar angle θ of range $[0, \pi]$. For an omni-directional source, the angular distribution of start directions can be of deterministic, half-deterministic or stochastic character. Especially for highly symmetric rooms (e.g. shoe box shaped ones), the first possibility bears the risk of producing unrealistic regular ray paths [vor89]. These easily result in wide gaps between singular incoming particles in the reflectogram[4]. Fig. 3.8 shows the projection of start directions of an omni-directional source onto a surface equivalent to the unity sphere for the three possibilities, with roughly 2500 particles each.

Algorithmically, these start directions are usually organized in a vector field. A flow diagram of a typical ray tracing algorithm can be observed in fig. 3.9. Marked are the radiation part (light gray), and the reception part (dark gray) of the algorithm. Unmarked boxes are part of the propagation calculation. In the exemplary algorithm presented here, the energy of the reflected particle is decreased according to the degree of absorption at each collision. Another common way to account for the loss of energy of the sound

[4] A reflectogram is a visual representation of the incoming reflections at the receiver's location, much the same as the RIR. It displays the amplitude of reflections over time and, sometimes, additionally over frequency.

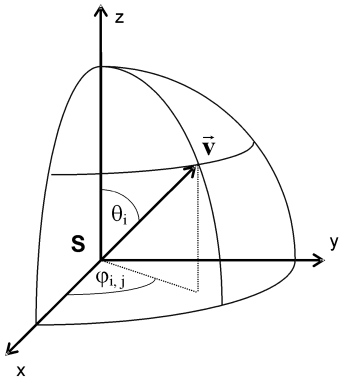


Fig. 3.7: Coordinate system for start directions of sound particles, first octant of sound source.

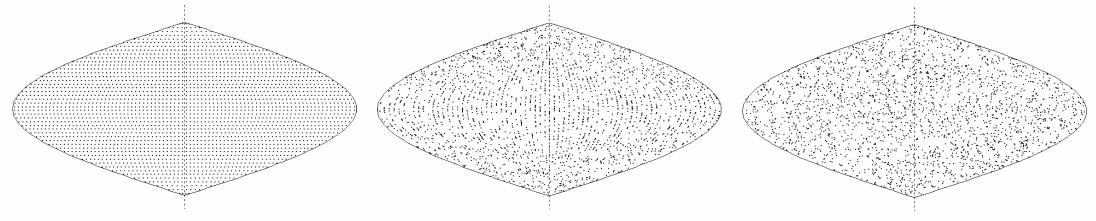


Fig. 3.8: Distribution of start directions of an omni-directional source: deterministic (left), half-deterministic (middle), and stochastic (right).

particles over time is to annihilate a particle with a certain probability upon each wall hitting. Both methods have particular advantages. According to Vorländer [vor88], the first is better suited for long impulse responses with high resolution in time, the latter is preferred for the computation of room acoustical quantities such as *reverberation time* or *definition*.

In order to display the reflectogram resulting from a simulation by means of the ray tracing method, a temporal resolution equaling the one of the monaural temporal resolution of the human ear is necessary, which lies in the order of $\Delta t = 10ms$. Vorländer states that for a reasonable precision around 12000 particles have to be traced. But if the method is used for the simulation of a binaural RIR which allows for the auralization of the simulated room, then the resolution must be in the order of the binaural temporal resolution of the human ear. This is at least a factor of 100 higher than the monaural one: therefore a temporal resolution of around $100\mu s$ for the RIR is necessary. If, again, the accuracy of the RIR shall be reasonably high to maintain the original features of the simulated room (Vorländer defines this as a maximum deviation of $3dB$ in amplitude at $200ms$), then at least 100 times more particles are needed for the simulation [vor88].

The propagation part of the ray tracing method is interesting in that the reflection of particles can be assumed as being geometrical or diffuse. Totally diffuse reflections from a rough surface can be described by Lambert's cosine law of diffuse reflection, originally formulated for the reflection of light: The luminous intensity of a light emitting area is proportional to the cosine of the scattering angle. Transformed to the field of acoustics, this means that the intensity of the sound which is scattered in a direction characterized by its angle ϑ is proportional to the cosine of this angle ϑ [kut91].

The reception of sound particles can be performed using spherical or cubic counting

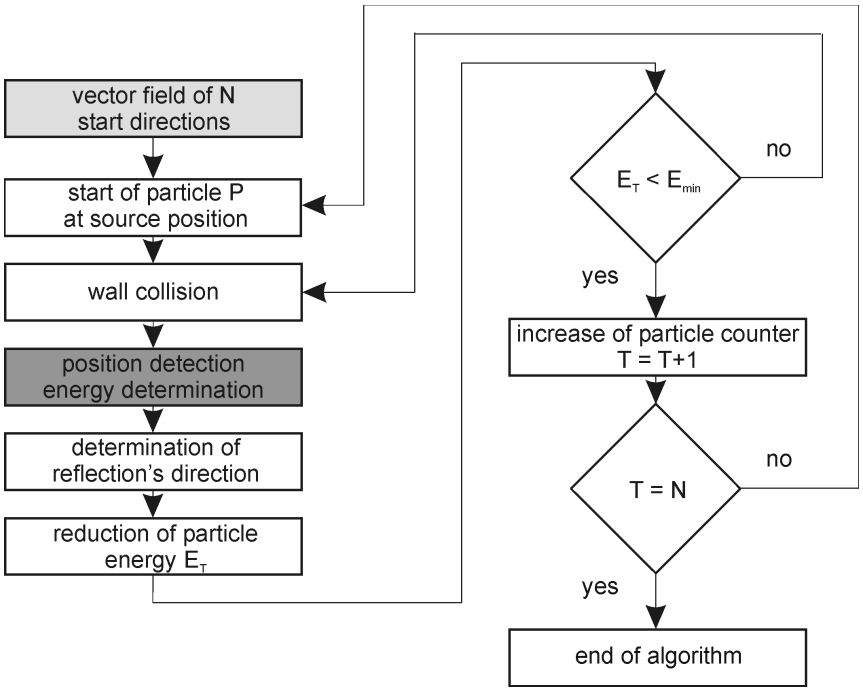


Fig. 3.9: Flow diagram of a typical ray tracing algorithm.

volumes. Whenever a particle crosses the counting volume, its energy contribution is determined. One of the problems to be solved here is a proper balance between number of particles and size of the counting volume. For small counting volumes, a large number of particles need to be sent out, because the probability of particles crossing the counting sphere or cube is proportional to its volume. Therefore, the reflectogram might show gaps if the number of particles is not high enough. On the other hand, bigger counting volumes which allow fewer particles to be sent out decrease the accuracy of the simulation for the receiver positions in question.

In fact, none of the past or current auditory environments are based on the ray tracing method, see sections 3.3 to 3.4. This is mostly due to the high amount of particles and the related amount of computational power necessary to get consistent binaural auditory impressions. On the other hand, this is also due to the high degree of algorithmic complexity of the ray tracing method. In comparison, the image source method with its straight forward approach can be implemented easier for rooms of regular geometric shape (which dominate in the real world), and first results can be achieved faster and with less manual fine tuning of the algorithm.

3.2.3 Beam Tracing Method

The beam tracing method is similar to the ray tracing method in that it also traces the rays of sound particles sent out from a sound source. Yet, it does not pursue arbitrarily directed rays, but only those that collide with the boundaries of a reflecting surface. In the two-dimensional projection of a virtual room always two boundary rays form what is called a "beam" of sound: a pyramidal shaped area or cone. Beam tracing sometimes is also referred to as cone tracing.

Funkhouser et al. have developed a beam tracing method for real-time rendering of

distributed acoustic spaces [fun98, fun99, fun04]. They divide their approach into four phases. The first phase is the *spatial subdivision phase*, in which spatial relationships inherent in the set of polygons describing the acoustic environment are precomputed. These relationships are then represented in a cell adjacency graph data structure that

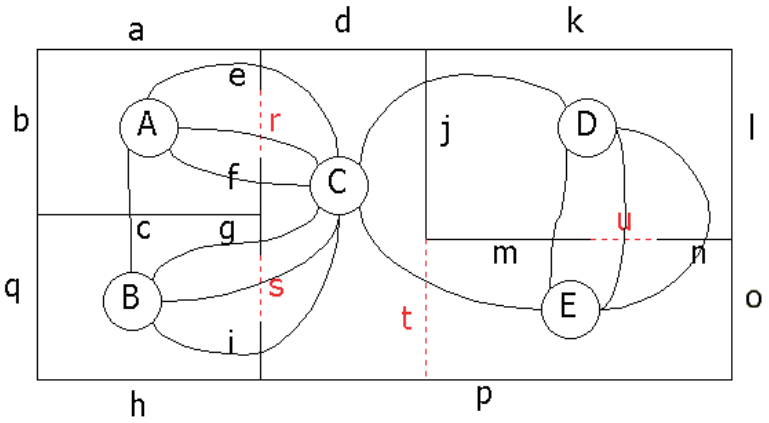


Fig. 3.10: Spatial subdivision of an acoustic space. Here, a Binary Space Partitioning (BSP) algorithm is used to separate space C from E. Subsequently, a cell adjacency graph is constructed containing all reflectors (lower case letters). *After [fun04].*

supports efficient traversals of the space, see fig. 3.10. The second phase is the *beam tracing phase*. The algorithm recursively follows beams of transmission, diffraction, and specular reflection through space for each audio source present in the space. The result of this phase is the beam tree data structure, see fig. 3.11. For stationary sound sources and stationary reflectors the beam tree can be computed a priori. The third phase is the *path generation phase*, in which the propagation paths from each source to the receiver are computed via lookup into the precomputed beam tree data structure. Whenever the receiver's location changes, the propagation paths need to be updated. During the final *auralization phase*, a spatialized audio signal is rendered by convolving anechoic source signals with impulse response filters. Funkhouser et al. derive the filters from the lengths, attenuations, and directions of the computed propagation paths.

To make the calculation of the beam tree more efficient, Foco et al. propose to compute the beam tree in the dual space [foc03]. They claim to significantly speed-up the construction of the beam tree by avoiding space subdivision at all. Their method allows them to dynamically recompute the beam tree as the sound source moves.

Fig. 3.12 (left) shows a typical situation: two beams are sent out from the sound source s toward the boundaries of reflectors $r3$ and $r4$. Reflector $r3$ partially covers reflector $r4$ for the given location of sound source s. After the beam has been sent out from the sound source, it is traced recursively along its propagation path. Whenever the beam crosses an object (a reflector), a geometrically correct (specular) reflection is computed. This can be done by mirroring the original source s at the reflector, similar to the image source method. Fig. 3.12 (right) shows the mirrored source s' and the beam resulting from the reflection at reflector $r4$.

The resulting beam is subsequently split into a number of sub-beams. These are delimited by the rays crossing the boundaries of the active reflectors in the new propagation

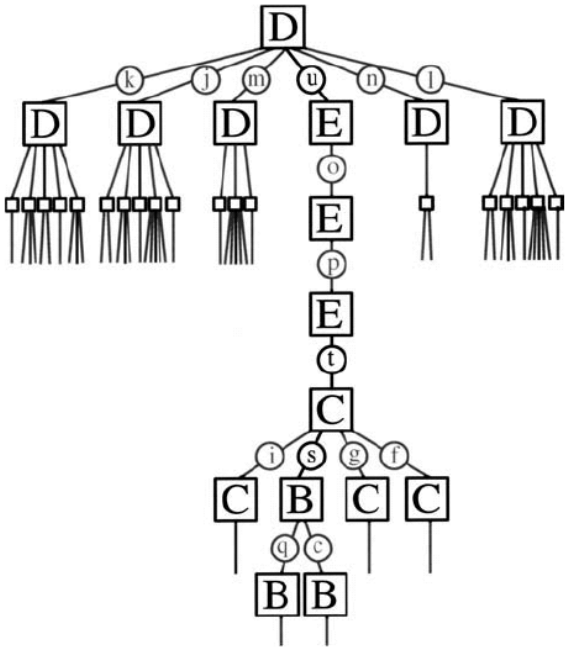


Fig. 3.11: The resulting beam tree based on the acoustic space shown in fig. 3.10. The detailed part of the beam tree shows relevant entries for source location in sub-space D and receiver location in sub-space B. *Taken from [fun04].*

area.[5] When these operations are performed recursively on the beams, each beam covers a so-called visibility area. Because these areas are constructed recursively, they can be organized in a tree-like structure to form the beam tree. Again, for stationary sound sources and stationary reflectors the beam tree can be computed a priori. It contains a map of the acoustic space with detailed information of which sub-beam covers which area. Auralizing the space then mainly consists of crossing the beam tree and "harvesting" the contributions from those sub-beams that cover the receiver's location. Crossing the beam tree can be organized as a depth-first-search, a well known problem in computer science for which efficient implementations exist.

The dual space transform, the main innovation introduced by Foco et al., is obtained by converting coordinates given in the world space according to a simple rule: starting from the straight line equation (3.1),

$$y = ax + b \tag{3.1}$$

the coordinate system is changed in such a way that the new abscissa results from the steepness a, and the new ordinate results from the former crossing b of the y-axis, see fig. 3.13. Therefore, a ray (a straight line) as described by equation (3.1) given in world space is transformed into a point (a, b) in the dual or parameter space. A point in world space transforms to a straight line in the parameter space. This means that the dual of a source is a ray. A reflector, which is a segment in world space, transforms to a beam-like region, see fig. 3.14, and vice versa.

The performance improvement of the dual space computation compared to the calculation in world space is, according to Foco et al., quite apparent. Whereas in world space the complexity depends on the total number of reflectors n, in dual space only the visible

[5] An active reflector is defined as the part of a reflector directly "illuminated" by a source.

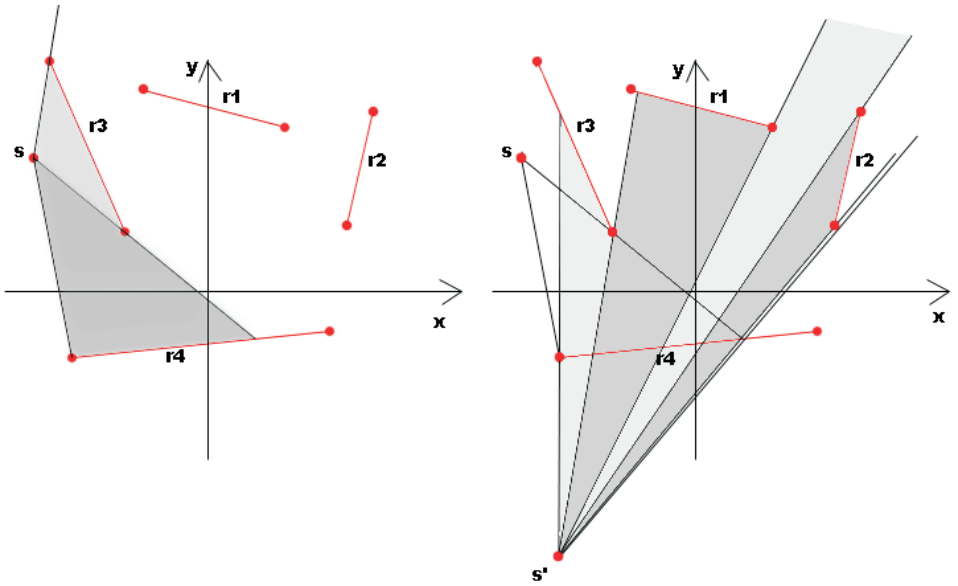


Fig. 3.12: Left: Two beams departing from the sound source s toward the reflectors $r3$ and $r4$. Note that reflector $r3$ partially covers reflector $r4$ for the given location of sound source s. *Right:* Original (s) and mirrored (s') source with beam resulting from reflection at reflector $r4$. The resulting beam is subsequently divided into a number of split beams (here: five split beams) according to the active reflectors it crosses. *After [foc03].*

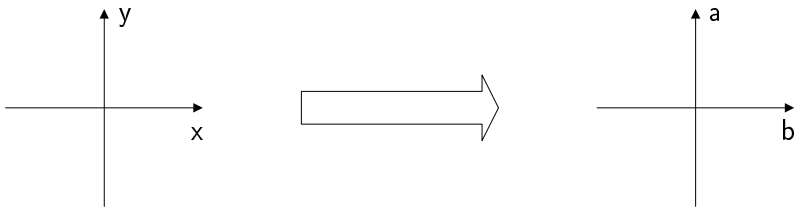


Fig. 3.13: The dual space transform, see equation (3.1), as suggested by Foco et al. in [foc03].

reflectors m ($< n$) are considered. Therefore the computing time necessary to rebuild the beam tree in the world space depends quadratically on the number of reflectors, whereas in dual space the computing time grows almost in a linear way. The full details of the beam tracing algorithm in dual space can be found in [foc03].

Antonacci et al. have further developed the beam tracing approach in dual space by including the simulation of sound diffraction [ant04a, ant04b, ant04c]. Thus, the dual space approach is comparable in accuracy to the space partitioning approach introduced by Funkhouser et al. [fun04], while allowing for updates of the beam tree at interactive rates.

The beam tracing method in general, using the space partitioning approach or applying the beam tree calculation in the dual space, is a very promising approach for real-time room acoustic simulation problems whenever the acoustic space is of more complex shape than a shoe box.

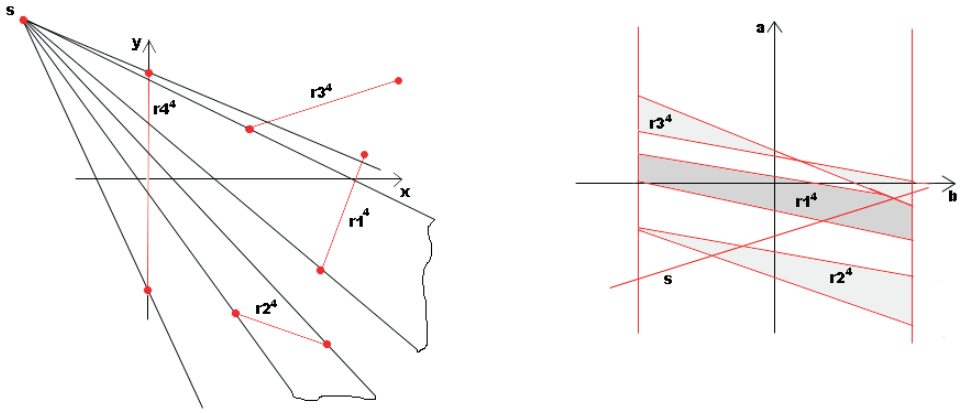


Fig. 3.14: Beam tracing in the dual space. *Left:* Source and reflectors in the world space. Note that, compared to fig. 3.12, the world space has been normalized relative to reflector $r4$ for easier computation in the dual space. *Right:* The same situation represented in dual space. *After [foc03].*

3.2.4 Other Methods for the ER Calculation

Although not exactly a method only for the calculation of early reflections, radiosity has been adapted from the field of computer graphics where it is used to compute the distribution of light in virtual scenes. The basic principle of the radiosity method is that the propagation of sound in a spatial enclosure is modeled as the radiation of sound from an omni-directional source to all the planes that enclose the space. The sound then re-radiates from each plane to all other planes of the spatial enclosure until the intensity of sound drops to a preset limit. This radiation of sound is calculated until a significant portion of the sound is dissipated, or until the reverberation time T_{60} is reached. The radiosity propagation model provides a relatively fast method to calculate an energy response graph at a given listener location for an omni-directional sound source. The number of radiations does not increase exponentially but is based on a constant cyclical iteration, see [mah99]. One major drawback is that an algorithmically simple implementation can only account for omni-directional sound sources. Furthermore, in situations where the planes of the spatial enclosure are highly absorbent, the resolution of the energy response graph is significantly reduced because the number of re-radiation computation cycles is reduced.

There are a number of other methods, mostly variations of the approaches presented here, that cannot be discussed further. They mostly consist in solutions for very specific room acoustic simulation problems, often making them computationally intensive and thus unattractive for real-time calculations.

3.2.5 Rendering of Diffuse Reverberation

In 1970 Schroeder already suggested that the calculation of the late reverberation could be separated from the computation of the early reflections [sch70]. This means that the late reverberation part of the RIR is exchangeable as a signal component, see fig. 3.2. Therefore, early reflections and diffuse reverberation can be computed separately. Also, different methods for the calculi might be applied and combined. Since the work of Barron [bar71], this separation has been common usage. In the preceding paragraphs the most important methods for the calculation of early reflections have been introduced. Now, an

overview of the most prominent methods for the generation of the late reverberation part of a RIR will be given.

In 1962 Schroeder described the usage of a comb filter and an all pass filter for the generation of artificial reverberation [sch62]. Both are single-delay Infinite Impulse Response (IIR) filters, see fig. 3.15. The structure of an all pass filter is similar to the comb filter, but it contains an additional feedforward path. Yet, this simple addition is sufficient to give

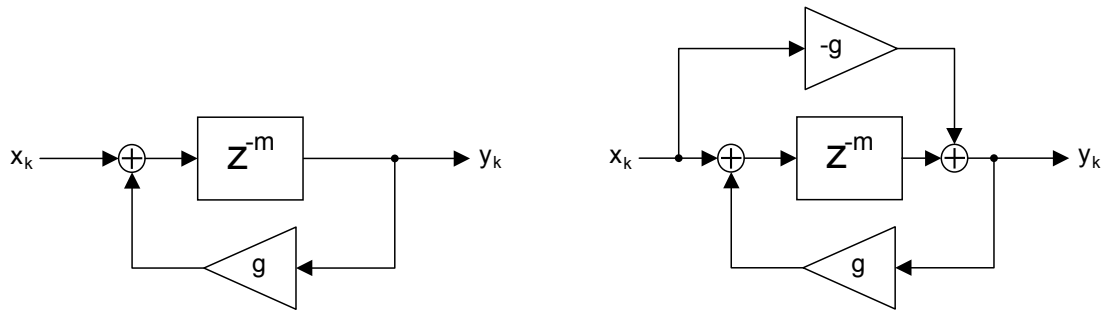


Fig. 3.15: A comb filter, left, and an all pass filter, right. Their structures only differ in the additional direct path.

the all pass filter a flat frequency response for a particular amplitude of the initial pulse in the time response. The all pass filter eliminates the strong coloration of sound caused by the comb filter, unless the input signal consists of short transients. Unfortunately, for these the comb filter coloration persists, as described by Jot [jot91].

To increase the echo density while avoiding coloration of the sound, Schroeder proposed two combinations of the above unit filters: a series of all pass filters and a parallel assembly of comb filters. The first produces a build-up of the echo density along the impulse response, but unnatural colorations of sound are still present in the response to short transients [moo79]. The latter yields a perceptively better sounding artificial reverberation whenever the frequency response exhibits a sufficiently large number of peaks per frequency range, such that the resonances from all comb filters overlap [jot91].

Since the publication of Moorer's article "About This Reverberation Business" in the Computer Music Journal in 1979 [moo79], we know that the late part of a RIR can be approximated by shaping white noise signals in such a way that their slope corresponds to the decay of RIRs measured in natural acoustic spaces - that means that their sound pressure decreases exponentially over time. Because the white noise signal corresponds very much to the stochastic nature of the late reverberation part of a naturally recorded RIR, such synthetic reverberation is perceived as sounding very natural.

Moorer went about generating his synthetic impulse responses by "shaping unit-variance Gaussian pseudo-random sequences with an exponential of the desired length" [moo79] and adding an impulse at the beginning for the direct sound. He noticed that a certain balance between impulse height and strength of the decay (about 12 : 1) was necessary for a natural sound. He then convolved the synthetic impulse response with unreverberated music signals and compared the result to actual concert hall recordings. By selectively filtering the impulse responses before convolution he was able to achieve the desired decay slopes at various frequencies.

When looking at Moorer's approach closely, one notices that there is no separate computation of early reflections. Instead, the direct sound is followed immediately by the diffuse reverberation part. This results in a well sounding, yet generic room impression. Specific

acoustic features of a room, like the amount of lateral energy, or a typical ER pattern (from which we are able to e.g. determine our own location in a well known room), cannot be reproduced using Moorer's approach. Furthermore, the computational load using a convolution operation (or two or more convolutions for stereo or multichannel rendering) over the whole reverberation time is rather high.

Jot and Chaigne in 1991 formulated an advanced approach for efficiently creating the late reverberation part of a RIR by means of using feedback delay networks [jot91]. In their paper they discuss how to produce real-time artificial reverberation that is perceptively indistinguishable from real reverberation. More specifically, Jot and Chaigne enhance the parallel comb filter structure suggested by Schroeder in such a way that the tonal colorations of the individual comb filters disappear even in the response to short transients. The resulting "reverberant filter" is designed in two steps: first, the design of a so-called "reference filter" with non-decaying eigenmodes (resonant frequencies) and a suitable reverberator structure, and second, the subsequent design of a so-called "absorbent filter" providing frequency dependent reverberation time control and a tone corrector stage.

All these approaches neglect the fact that a single reverberated monophonic signal cannot provide a convincing room impression. What is missing in monophonic signals is the apparent impression of width and depth of a sound source, as well as the envelopment caused by the reflections coming from the walls surrounding a listener. Simply raising the number of reproduction channels (loudspeakers used for the reproduction) does not solve the problem, as these signals need to be uncorrelated to provide a room impression. Therefore, the number of channels actually processed by the reverberation algorithm needs to be increased, resulting in an increase in computational demand proportional to the number of reproduction channels.

Gardner in 1992 presented a system capable of rendering six channels of artificially reverberated sound in real-time [gar92]. His "Realtime Multichannel Room Simulator" (RMRS) is based on an idea formulated earlier in 1985 by Vercoe [gar92]: the use of nested all pass filters rather than all pass filters in series as in the Schroeder reverberator in order to achieve an exponential buildup of echoes. The suggested nesting of filters consists in embedding an all pass filter into the delay element of another all pass filter.

Fig. 3.16 shows the basic structure of a nested system. It is the same as for an all pass filter, with the inner delay element being replaced by an arbitrary function $F(z)$.

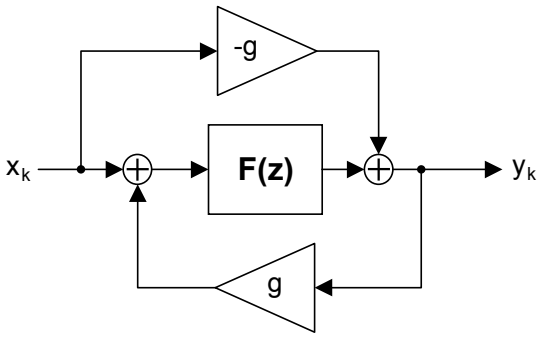


Fig. 3.16: Basic structure of a nested system.

The nesting of filters as suggested by Gardner is shown in fig. 3.17. The echoes generated by the inner all pass filters (here only one nesting level is shown) are recirculated to their inputs via the outer feedback path (here: g_2). Therefore, the number of echoes that are

generated in response to an impulse increases over time rather than remaining constant as in Schroeder's approach.

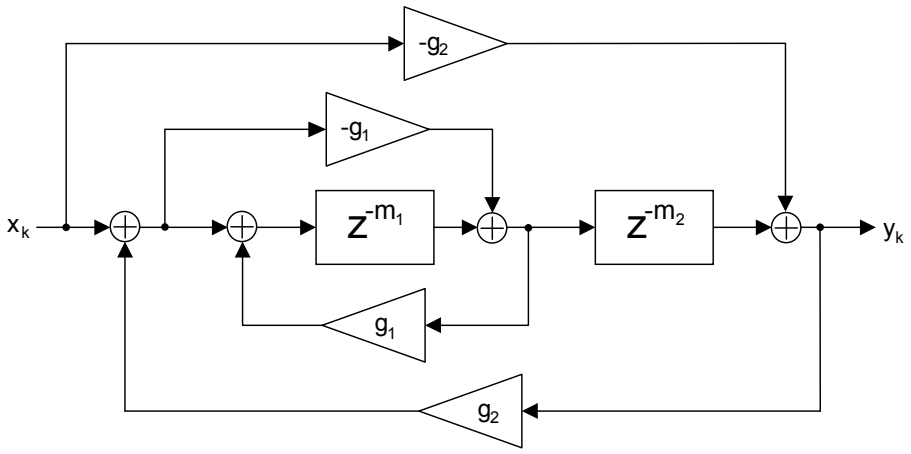


Fig. 3.17: Structure of a single nested all pass system. The all pass around delay element z^{-m_1} is "embedded" into element z^{-m_2}.

One nice effect of using nested all pass filters is that there is no need to worry about stability. Contrary to comb filters, all pass filters are immanently stable and cascading or nesting does not change that particular quality. Yet, tonal coloration can also occur in all pass filters: although they pass all frequencies equally in the long term, the short time frequency analysis performed by the human ear "can detect momentary coloration, and thus allpass systems can sound buzzy, or have a metallic ring" [gar92]. Gardner also points out that it is important not to take the output tap from the interior of a multiple nested all pass filter, because such a system is highly resonant. Output taps should always be taken from locations between cascaded all passes.

The work of Gardner showed that when another all pass is inserted into the outer all pass, the impulse response of the system can take on an entirely new character and loose its buzziness and metallic artifacts. Fig. 3.18 shows Gardner's generalized allpass reverberator. Some of the output of the all pass system is fed back to the input through a moderate delay, thus increasing the echo density and smoothing out the buzzyness and metallic sound coloration. Additionally, the low pass filter in the outer feedback path can simulate the effect of air absorption. Each all pass filter itself may be a cascaded or nested form. Note that the system is no longer all pass, because of the outer comb and low pass filters as well as the multiple output taps which may cause phase cancellation.

3.2.6 Assembly of ER and DR

Once it is agreed upon computing the early reflection part and the diffuse reverberation part of a RIR separately (and assuming that we do not perform the simulations for the pure mathematical joy of it), it is necessary to reassemble both parts to form one resulting RIR. It is easily understood that the two parts need to fit in such a way that the joint between ER and DR parts is not perceivable as such. Yet, as the diffuse reverberation part will most likely stem from a generic reverberator approach not related in any way to the actual room simulated for the early reflections part, the two parts will need to be adapted level- and frequency-wise.

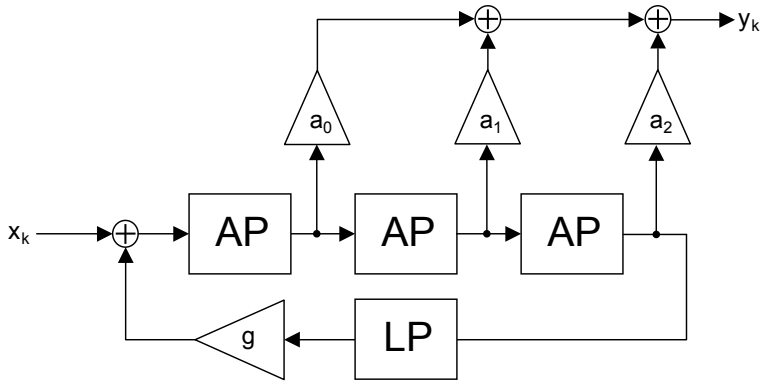


Fig. 3.18: Gardner's generalized all pass reverberator with low pass filtered feedback and multiple weighted output taps. *After [gar92].*

Gardner gives an overview of a methodology developed for the RMRS, which can be adapted also for other systems. He suggests to match the decay slope of the diffuse reverberation part with the maximum energy point of the early reflections part. This way, the gain factor of the diffuse reverberation part is raised (or lowered) such that the linear projection of the diffuse reverberation slope backwards in time passes through the point of maximum early reflection energy [gar92]. The diffuse reverberation slope can easily be determined from the room specification by computing the reverberation time T_{60}. The maximum early reflection energy can be determined from the early reflection calculation itself.

Frequency-wise Gardner offers a very elegant solution to the problem of matching early reflection and diffuse reverberation tonal coloring. By feeding the early reflections into the diffuse reverberator filter network, the overall diffuse reverberation blends seamlessly with the early reflections. At the same time, the echo density of the late reverberation part is further increased [gar92].

3.3 Auditory Virtual Environments

A number of auditory virtual environments have been developed in the past. This section introduces the most important ones and gives an overview of the actual developments and restrictions. The focus here is mainly on the reproduction setup of audio, not on the simulation algorithms themselves, as for most of the systems the algorithms used for the room acoustic simulation have not been published or described properly.

Convolvotron - NASA Ames In 1991, Elizabeth Wenzel of the NASA Ames Research Center and Scott Foster of Crystal River Engineering presented a demonstration system at the SIGGRAPH 91 "Tomorrow's Reality" program [fos91]. This system allowed real-time interactive simulation of simple room audio dynamics. Its basic component was the *Convolvotron* DSP board, which could do real-time 3D positioning for sound that was otherwise not positioned in space via headphones. The DSP board could be placed in a PC running a DOS operating system and provided enough computing power to render the direct sound paths of four sound sources simultaneously. Its scenario update rate was around $33Hz$, with an internal latency of $32ms$ for a filter order of 256. Its sampling rate was fixed at $50kHz$, and the system provided a C language user interface.

In order to achieve a positioning of sound, the Convolvotron simulated the phase and amplitude differences between left and right ear that would exist if the sound came from the position where the Convolvotron placed it. The parameters necessary to place a sound in 3D had been recorded before with the help of a dummy head with a small microphone in each ear. Sounds had been played at various positions around the head and the amplitude and phase differences between the sounds arriving at the microphones had been recorded. In other words, the *Head Related Transfer Functions* (HRTF) of the dummy head had been determined for various positions and stored in a database. These were transformed into *Head Related Impulse Responses* (HRIR) to allow the Convolvotron to do a two channel real-time convolution in the time domain. The convolution of an audio signal with these pairs of impulse responses placed the audio signal virtually around the listener. The Convolvotron allowed the user to manipulate the position of sounds and hear the result in real-time. Only the direct sound path was simulated, no room acoustic simulation could be performed.

HURON - Lake Technologies Lake Technologies' Huron is a proprietary hardware / software DSP platform for audio applications [w-lak]. It is commercially available for use in VR systems. It provides a modular architecture allowing for a variable amount of computational power available in a unit. The hardware platform consists of a PC and a custom backplane supporting DSP and audio input/output cards providing 2 to 512 channels of digital audio signal processing. The software is based on a modular architecture which enables the use of algorithms related to VR applications. Currently available algorithms can be used for positioning and moving of sound sources and listener in real-time, head tracking, binaural rendering and rendering for higher loudspeaker numbers (up to 50). No room acoustic rendering modules are available, although Lake Technologies offer a convolution algorithm used for low latency filtering which could possibly serve as a basis for the rendering of spatial impressions and reverberation.

Realtime Multichannel Room Simulator - MIT Media Labs The Realtime Multichannel Room Simulator (RMRS) developed by Gardner at MIT Media Labs in the early 1990s is a DSP based system that can render the simulated reverberant field of an arbitrary polyhedral room in real-time [gar92]. Sound is reproduced by six loudspeakers located around the listener, with each output signal being computed by one Motorola 56001 digital signal processor. Early reflections are generated by an image source method algorithm, which determines a finite impulse response filter per output channel. Diffuse reverberation is computed using infinite impulse response reverberators based on nested and cascaded all pass filters, see subsection 3.2.5.

Because the loudspeakers are located on a horizontal plane, all image sources not on that plane are made horizontal by setting their elevation angle to zero, while their azimuth angle and distance from the listener is kept. The amplitudes of these sources are then scaled by the cosine of the elevation angle, such that much elevated sources will be ignored, whereas sources with no elevation from the horizontal plane are unaffected.

Apart from introducing three different nested all pass filter structures with increasing complexity (for small, medium sized and large rooms, respectively), the RMRS also makes use of a concept described in subsection 3.2.6: the separated computation of early reflection and diffuse reverberation and subsequent combination of the two. Fig. 3.19 shows an overview of the simulation procedure introduced by Gardner.

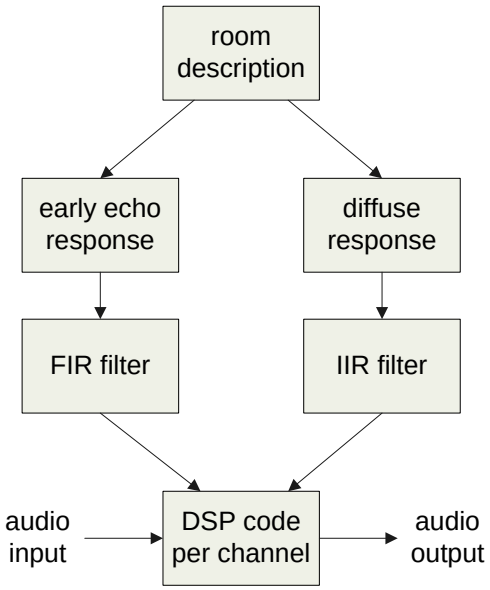


Fig. 3.19: Overview of the RMRS simulation procedure. *After [gar92].*

SLAB real-time auralization tool - NASA Ames Research Center The SLAB system is a real-time virtual acoustic environment rendering system developed mainly by Joel D. Miller and Elizabeth Wenzel at the NASA Ames Research Center. Since 2000 it has been continuously developed [wen00] and is available as an open-source software system for the Microsoft Windows platform. Binaural spatial audio rendering is based on a HRTF database and reproduction is done via headphones. It is a software-based successor to the Convolvotron DSP system.

The SLAB system provides an Application Programming Interface (API) for access to the actual audio rendering processes. Acoustic scene parameters can be manipulated in real-time, such as sound source, listener, and wall locations. Also configurable is the number of FIR taps used for the rendering process. The system can be enhanced in functionality by writing so-called Render Plug-Ins. They can be exchanged while rendering, thus allowing for direct comparison of different rendering strategies. Additionally, a signal generator can produce different test signals for typical alarm sounds. Typically, the room acoustic effect rendered consists of an image source model with six first order reflections for a maximum of 4 direct sound sources located in a single acoustic space [mil02]. According to Miller and Wenzel, the system's main purpose is to study the human spatial hearing characteristics.

Spatialisateur / MPEG-4 Audio Perceptual Approach - IRCAM Spat (short for "spatialisateur" or "spatializer") was originally developed by Jean-Marc Jot and Olivier Warusfel of IRCAM (Institut de Recherche et Coordination Acoustique/Musique) in Paris, France, in the 1990s. Spat is a software tool that allows for managing the spatial dimension of music without dealing with acoustical or geometrical parameters of the enclosing space. This is realized by providing a set of descriptive parameters that are mapped to their physical counterparts. Unfortunately, not all of these parameters are orthogonal, so that changing a parameter sometimes leads to unwanted results. Spat is based on the feedback delay networks created by Jot and Chaigne [jot91] for the simulation of diffuse reverberation,

see subsection 3.2.5. The Spat formalism is part of the MPEG-4 audio standard as the so-called "MPEG-4 Audio Perceptual Approach" [14496-1]. A detailed description is given in section 5.6.

Spat is also provided as a library of objects for the MAX/MSP graphical music creation environment and is thus mainly used by musicians wanting interactive control of a particular room effect or sound position.

SCATIS - Ruhr-University Bochum, Aalborg University, Head Acoustics SCATIS was a European project developing an auditory and haptic virtual reality system in the early 1990s. For the auditory modality, participants in the project were Ruhr-University Bochum / Germany, Aalborg University / Denmark, and Head Acoustics Aachen / Germany. SCATIS presented auditory and haptic stimuli in an environment that allowed users to approach virtual sound sources, touch them and manually move them toward arbitrary positions in the virtual space. Up to 64 direct and indirect virtual sources were computed in parallel using a cluster of 80 Motorola 56002 digital signal processors (DSPs). The sampling rate was $48kHz$, with an update rate of $60Hz$ and a system latency of $60ms$. Sets of HRTFs with a resolution of around 11° in azimuth and 22° in elevation were measured to provide binaural cues to the users. Off-line linear interpolation between these sets was used to improve available system resolution to below 2°. The system's main field of application was psychoacoustics research [bla00, dje00].

The SCATIS system has been ported to run on a high-performance PC environment. The system is now called IKA-SIM, and performance is reported to be comparable to that of the DSP-based implementation [nov05].

BRS / Binaural Room Scanning - Studer, IRT The Binaural Room Scanning (BRS) system was developed jointly by Studer AG, Switzerland, and Institut für Rundfunktechnik (IRT) Munich, Germany, in the late 1990s. BRS overcomes the problem of static directional cues when using headphones for the reproduction of binaural head-related audio: as soon as the listener rotates the head, an immediate in-head localization occurs, because the directional cues contained in the audio rotate in accordance to the listener's head. BRS relies on a head tracker to detect the azimuth of the listener's head. As soon as it is turned, the directional cues are updated. I.e. the binaural HRIRs used to convolve the dry audio signal with, are changed to match the direction of incidence of the virtual sound source.

The BRS system was primarily designed to simulate via headphones the reproduction of multichannel audio by a surround loudspeaker setup in a high-quality listening room, thus enabling highest quality listening conditions in acoustically unprepared spaces. Therefore, the virtual sound source positions correspond to the positions of the loudspeakers as defined in the ITU-R BS.775 standardized multichannel setup [itu775], see fig. 3.20. The system does not generate room acoustic impressions by simulation or artificial reverberation algorithm, but relies on measured impulse responses from real rooms. These binaural room impulse responses (BRIRs) are measured using a dummy head mounted on a rotating table in the center of the multichannel loudspeaker setup. An impulse response "fingerprint" is taken every 6° in azimuth for every loudspeaker and converted to corresponding spectra. The high azimuthal resolution of the head tracker of about 0.015° [w-pol] allows for a precise frequency domain interpolation of the binaural room spectra for loudspeaker positions in between the fingerprints [pel02].

Pellegrini describes a binaural Auditory Virtual Environment using the BRS system

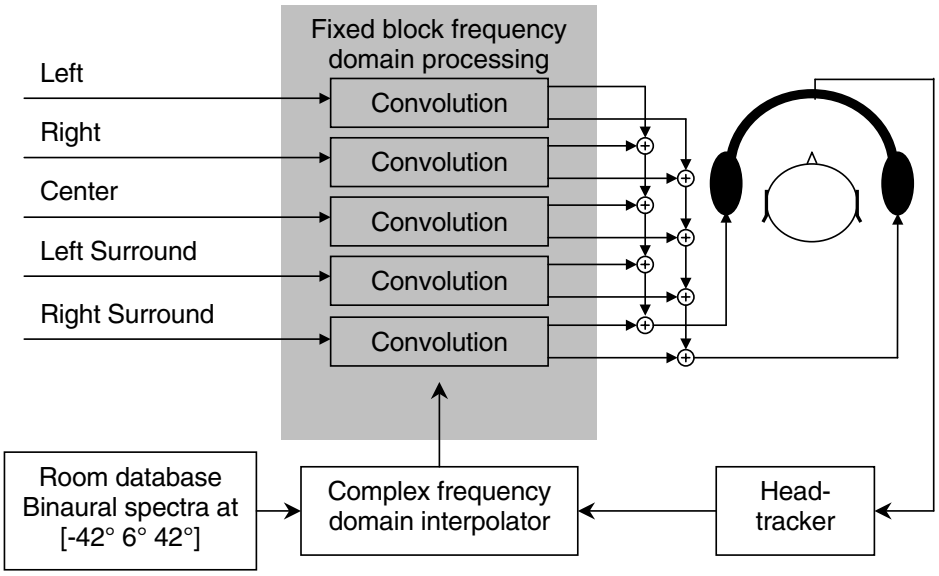


Fig. 3.20: Block diagram for the BRS system. *From [pel02].*

[pel02]. It implements the real-time room acoustic rendering on a total of 24 SHARC processors (Analog Devices ADSP21062) with a processing power of 40 MFLOPS each. The system is capable of computing up to 40 reflections, and additional diffuse reverberation in real-time if only one sound source is rendered. The number of reflections that can be computed decreases with higher numbers of sound sources. Fig. 3.21 shows a signal processing block diagram of the system.

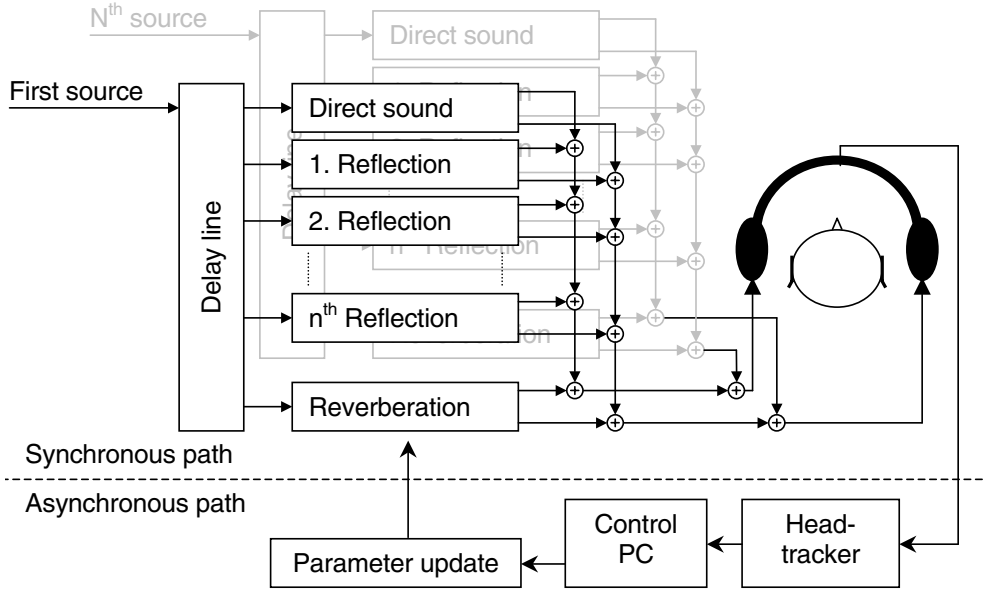


Fig. 3.21: Block diagram for an auditory virtual environment based on the BRS system as described by Pellegrini. *From [pel02].*

CLAM - Universitat Pompeu Fabra CLAM ("C++ Library for Audio and Music") is a software framework for research and application development in the audio and music

domain [ama04]. It has been developed at the Universitat Pompeu Fabra, Barcelona, Spain, since around the year 2000 and is available as an open source, platform independent tool. Functionality currently comprises input / output of sound, processing, storage, display, MIDI interfacing, XML serialization services, a visualization module as well as multi-threading handling [w-cla].

IEM Cube - Institute of Electronic Music and Acoustics The IEM cube developed and installed at the Institute of Electronic Music and Acoustics, Graz, Austria, is a system for the reproduction of 3D sound via a hemispheric arrangement of 24 loudspeakers. It is based on the principles of ambisonics and can be used to reproduce three-dimensional sound fields from ambisonics recordings of third order or higher. It is mainly used for the research of higher order ambisonics and related algorithms. At the same time, the system is also used as a live tool for the music composer, allowing the artist to freely position up to 50 channels (sound sources) in 3D space [w-iem].

WFS - Wave Field Synthesis Wave Field Synthesis (WFS) is an approach based on the Huygens-Fresnel principle stating that wave fronts of a propagating wave can be recreated by spatially distributed secondary sources. This principle was advanced by Berkhout in 1988 to form the theoretical foundations of the WFS [ber88]. One of the main benefits of WFS is that a correct acoustic impression is not limited to a "sweet spot" area as in traditional multichannel loudspeaker setups. Instead, it is reproduced correctly within all of the created wave field, the so-called listening area, spanned by a huge array of loudspeakers. The positioning of virtual sound sources outside as well as inside the listening area is possible [hul04].

A number of systems based on the WFS theory have been developed. The most prominent are the IOSONO system [w-ios] originally developed by Fraunhofer IDMT Ilmenau, Germany, which is currently based on up to 192 loudspeakers, and the WFS system by Sonic Emotion [w-son]. The IOSONO WFS system is installed at a number of venues, e.g. a cinema in Ilmenau, Germany, where it is mainly used to enlarge the sweet spot - the system is used on a track-related basis to simulate a large 5-channel loudspeaker setup. It can also be used on a source-related basis with currently up to 32 moving sound sources. Basic room acoustic simulation algorithms are implemented. A number of studies have shown that a combination of WFS with visual displays is possible [mel03, mel06, spr06].

3.4 Audiovisual Virtual Environments

DIVA - Helsinki University of Technology The DIVA (Digital Interactive Virtual Acoustics) system developed at Helsinki University of Technology, started around 1995, aimed at providing a virtual acoustic space to a user who could move freely inside a room [sav99a, lok01]. Direct sound and first reflections were computed in real-time to create a 3D sound image. In DIVA this was done by applying ITDs (Interaural Time Differences) and HRTFs (Head Related Transfer Functions) to the direct and image sound sources to allow for a correct impression of the direction of incidence [sav99b]. Late reverberation was added by feeding the early reflections into several parallel feedback loops which contained a delay line, a comb-allpass filter and a lowpass filter each [vää97]. Fine tuning needed to "be done by ear in advance" [w-div]. The computing was performed on SGI workstations which communicated across a UDP socket. Thus, the computation of (image) source and

listener positions was separated from the actual audio processing. Fig. 3.22 presents a system overview. In theory, the listening space could be arbitrarily complex, but only the largest surfaces were taken into account in the simulation process.

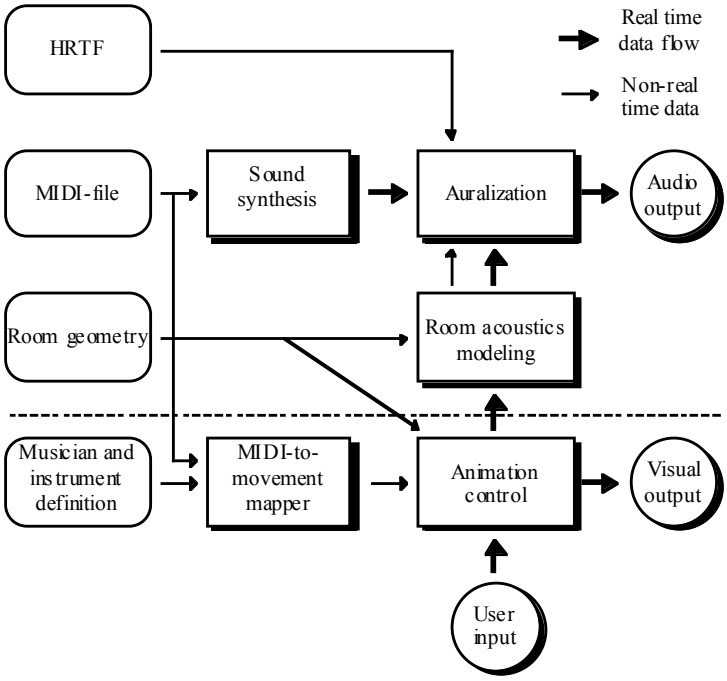


Fig. 3.22: The DIVA system information flow graph, *from [huo96].*

The visual rendering of the virtual space was also performed on SGI workstations and the graphical output was presented on computer monitors. One application demonstrated frequently to the public was a case study of a user acting as a conductor for a virtual orchestra. The four virtual players played their instruments with animated fingerings according to the strokes of the conductor which were analyzed in real-time using a motion tracking device connected to the system. At the same time, a second user was able to move around in the virtual space with continuously changing acoustic and visual impression generated in real-time [sav97].

EVE - Helsinki University of Technology EVE (Experimental Virtual Environment) is a rear-projection based virtual reality system in which users are surrounded by screens of approximately 3 by 3 meters. Stereoscopic images are projected onto these screens and they are viewed through stereo shutter glasses. Because of head tracking applied in EVE, the user can move around in the stereoscopic three-dimensional space projected onto the screen similar to a real-world experience. While wearing a head tracking device, the user can also interact with the virtual environment by means of other tracked sensors.

EVE is the successor to the DIVA system presented in the last paragraph and therefore incorporates a number of technologies already available in DIVA. The hardware that EVE uses for audio processing is built around a dual-processor PC computer running a Linux operating system. Custom software is used for the acoustic modeling, sound source panning, and equalization filtering. The software is split into a low-level signal processing library and a higher level dynamic signal processing/routing toolkit. The low-level signal processing is done in a sample-by-sample fashion. The high-level signal processing allows

for the use of customized plug-in style algorithms to be controlled, modified or exchanged during runtime [ilm01]. A set of 15 loudspeakers plus subwoofer is used for the actual reproduction of audio [w-eve].

The EVE system is used in a number of projects, among them the Uni-Verse project supported by the EC's sixth framework programme. The Uni-Verse Sound Renderer is the most advanced successor of the HUT audio renderer originally used in DIVA. Based on the EVE audio software, the Uni-Verse sound rendering process follows the well-known source–medium–receiver model as described in [beg94]. This means that functionality incorporates algorithms for reading files and streams (source), for simulating air absorption and distance attenuation by means of IIR filters, for creating distance delays for the image sources, and for creating artificial late reverberation (medium). Finally, the panning of sound sources can be done using HRTFs for binaural reproduction via headphones or via Vector Base Amplitude Panning (VBAP) [pul97, pul01] for reproduction via the EVE loudspeaker setup (receiver). The audio rendering is performed on a dedicated computer which receives update commands from the interactivity module via a customized protocol. A performance test was reported in 2006 in which a second order image source model (total of 37 sources at $44.1kHz$) using VBAP panning was causing between 50% and 70% processor load in a static listening setting (no interaction), depending on the audio block size. Audio block size was varied between 256 and 16 samples length, respectively [uni06]. Unfortunately, the authors of that report do not mention the type of CPU used for the performance test.

RAVEN - RWTH Aachen RAVEN (Room Acoustics for Virtual ENvironments) is a system for interactive, high-quality and real-time room acoustic rendering in virtual environments at RWTH Aachen, Germany. It is part of the cave at the Virtual Reality Center Aachen and has been installed in 2004. It provides a stereoscopic 360° view projected onto four rectangular lateral screens plus a floor projection. The dimensions are $3.60m \times 2.70m \times 2.70m$ with a resolution of $1600 \times 1200pixel$ per screen. The stereoscopic projection is done with two digital LCD projectors per screen and with circular polarization. Visual rendering is performed on a number of networked workstation computers [vrc06].

RAVEN itself is based on the principles of geometric room acoustics simulation. It uses a hybrid approach which combines the advantages of the deterministic image source method for the computation of early reflections with a stochastic ray tracing method for the calculation of the diffuse late reverberation. The latter is a unique approach for real-time room acoustic simulations, as the diffuse reverberation is usually computed with the help of a generic filter network, see section 3.2.5. Using binary space partitioning (BSP) algorithms, this hybrid approach is qualified for real-time room acoustic rendering under certain conditions of complexity.

GPAC - Ecole Nationale Supérieure des Télécommunications GPAC is an open source multimedia framework for research and academic purposes [w-gpac]. It consists of an MPEG-4 capable multimedia player called *Osmo4* and a multimedia packager called *MP4Box*. Additionally, some server tools are currently being developed. GPAC is cross-platform compatible.

The Osmo4 player is considered one of the most advanced 2D MPEG-4 players available. 3D scenes can be reproduced based on OpenGL back-ends, but the audio features only

provide basic functionality: multichannel support, mapping and mixing of channels, and basic synchronization between audio and video. No room acoustic simulation is provided, and the audio "engine" is static in the sense that it does not provide the flexibility of a plug-in structure.

4 Interactivity Issues and Presence

4.1 Interactivity Issues

The term "interactivity" is hard to find in dictionaries, but e.g. Merriam-Webster's Online Dictionary explains the adjective *interactive* like this:

> "involving the actions or input of a user; especially: of, relating to, or being a two-way electronic communication system (as a telephone, cable television, or a computer) that involves a user's orders (as for information or merchandise) or responses (as to a poll)" [w-mew07]

The web site of Interactivity Consultants has a definition of interactivity that is more specific to the meaning the term has in this work:

> "interactivity n.
> an attribute or functionality, intentionally designed into man-made objects, physical, or virtual environments, characterized by the ability to sense accurately, then respond or react dynamically and intelligently to movement, gestures, expressions, or changes in human bodily or psychological states and intentions, changes in geographic location, changes in environmental condition, or any combination. Such dynamic intelligence may be achieved by the use of scripting or programming, embedded microcontrollers, sensors, GPS, haptics, and network connections to other systems and data. Ideas or concepts for new applications exhibiting interactivity are traditionally communicated to other people through user scenarios." [w-int]

Many such user scenarios are comprised in the interactive audiovisual application system as introduced in the beginning of this work: The MPEG-4 player that might be used as a presentation system for pre-recorded audiovisual content, as well as for non-linear object-based content, as an enhanced teleconferencing system, as a personalized news broadcast receiver, and so on[1].

The concept of interactivity has been defined by Lee et al. based on three major viewpoints: technology oriented, communication-setting oriented, and individual oriented views [lee05, lee07]. Here, the technology-oriented view of interactivity is adopted. The "technology-oriented view of interactivity defines interactivity as a characteristic of new technologies that makes an individual's participation in a communication setting possible and efficient" [lee07].

Steuer holds that interactivity is a stimulus-driven variable which is determined by the technological structure of the medium [ste92]. 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" - in other words, the degree to which users can influence the target environment. He identifies three factors that contribute to interactivity:

[1] Further explanations regarding the object- and scene-concept of MPEG-4 are given in chapter 5.

- *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. They are very briefly discussed in the following sections.

4.1.1 Latency

Latency is one of the main concerns in interactive application systems. 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. As these components may interact in more than one way, a system's end-to-end latency may possibly vary over time. Miller et al. suggest to include measurements of the mean, standard deviation, and range to completely characterize this parameter [mil03].

Meehan et al. report a study in which they tested the perceived sense of presence (see section 4.2) for two different end-to-end latencies in a Virtual Environment [mee03]. 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. They 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 rooms.

MacKenzie and Ware conducted the first quantitative experiments with respect to effects of visual latency [mac93]. Participants completed a Fitts' Law target acquisition task[2] 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. They report that the threshold at which latency started to affect the performance was approximately $75ms$. This effect also was dependent on the task difficulty: the harder the task was, the greater the adverse effect caused by latency was.

Wenzel has published a number of reports about the impact of system latency on the dynamic performance in virtual acoustic environments with a focus on localization of sound sources [wen98, wen99, wen01]. 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 (or end-to-end latency) is higher than around $60ms$ for audio only presentations [wen98]. 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 completely [wen01].

[2] Fitts' Law is a model of human movement in human-computer interaction and ergonomics. It predicts that the time required to move to a target area is a function of distance and size of the target.

Nordahl examined the impact of self-induced footstep sounds on the perception of presence and latency [nor05]. Interestingly, for audiovisual feedback in a Virtual Environment (VE), the max. 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 $60.9ms$ vs. $41.7ms$). Nordahl explains this with attention being focused mainly on the visual rather than the auditory feedback in the audiovisual case.

Looking at these experimental results, it is difficult to draw a conclusion on the maximum allowed latency for interactive audiovisual application systems. Apparently, the perception of latency as such depends on the system, the task, and the content that is displayed. Furthermore, measuring end-to-end or total system latency is not a trivial task. A general recommendation would be to keep latency as low as possible within any such system, preferably below $50ms$.

4.1.2 Input and Perceptual Feedback

Perceptual feedback is the response that a system provides to the user's input. In the system described in this work, perceptual feedback is provided in the auditory and visual domains. Input provided by the user can, in the general case, consist of any kind of signal accepted by the system for controlling it: speech, gesture, haptic control, eye tracking, etc. Here, input can be provided by a limited number of mainly haptic input devices described in section 5.3.8.

Input and perceptual feedback are related to Steuer's *mapping* factor [ste92]. Although only haptic input and audiovisual output are provided by the system described in this work, the mapping features are very extensive, see section 5.3.

Steuer's *range* factor is related to the kind of interaction that is offered by the audiovisual application. This depends strongly on the goal of the application itself. In an entertainment application users might expect a different range of interaction than in a news broadcast application. Again, section 5.3 provides an overview of the possibilities that the MPEG-4 based system described here offers in terms of interactivity *range*. Some examples of different interactivity ranges are provided in chapter 8 in the description of the subjective assessments that have been performed (sections 8.7 - 8.12).

4.2 Presence

Closely related to interactivity is presence. Presence in interactive audiovisual application systems or Virtual Environments is often described as the feeling of "being there" [lar03] that generates involvement of the user. Lombard and Ditton define presence in a broader sense as the "perceptual illusion of nonmediation" [lom97].

According to Steuer, 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 [ste92]. Steuer has found vividness (ability to technologically display sensory rich environments) to be the second fundamental component of presence. Along the same lines, Sheridan assumed 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 [she94].

Other factors have been found to be determinant for presence - these depend on the theoretical concept applied by the researcher. Kuschel et al. give an overview in [kus07]:

> "Slater defined external factors (technology related) and internal factors (perception related) to successfully generate presence. Witmer and Singer proposed a four-factor categorization consisting of control-, sensory-, distraction- , and realism factors [wit98]. Lombard and Ditton additionally named the willingness to suspend disbelief and prior experience of the operator as well as the form of the target reality as influencing factors of presence [lom97]."

Ellis points out that presence is not necessarily the ultimate goal of every interactive audiovisual application system [ell96]. He holds that communication effectiveness can be far more important than presence, especially in situations "where the medium itself is not the message". This very interesting point makes reference to one of the major problems with technological advancements in the field: the question whether a certain task is actually suited for being performed with the help of an interactive audiovisual application system. Ellis states that the key question to ask "may not be whether the users 'feel' present in the remote or synthetic environment but whether they can accomplish the tasks they accept" [ell96].

Part III

Tools

5 IAVAS I3D Player as Rendering Platform

The main parts of the IAVAS I3D MPEG-4 player date back to the IAVAS project funded by the Thuringian Ministry of Science, Research and the Arts between 2002 and 2004. A group of eight scientists, including the author of this thesis, were involved in the development of the software. In the course of this thesis' work, the foundations for the TANGA real-time audio rendering engine were added in 2005 by Mathias Schwark and the author of this thesis [rei05a]. Subsequent developments and additions of functionality were added by the author until 2007, with the help of a number of students' projects which the author has supervised.

The I3D is based on the IM1 (Implementation Model One) reference implementation of the MPEG, the so-called core module [14496-5]. Originally, the software uses the MPEG-4 file format to read and display content specific to its part. For testing purposes, IM1 uses the ITU-T standards H.263 for video and G.723 for audio. Each part of the reference software (Systems, Visual, Audio) is self contained and does not interoperate with any other part [w-m4if].

The I3D combines all parts of the reference software and makes them functional across these parts. Large parts of the code have been re-designed in the IAVAS project for better performance. It has been enhanced significantly to include more codecs as e.g. MP3, H.263, and H.264. Other formats can be integrated easily because the *libavcodec* part of FFmpeg, a collection of software libraries that can record, convert and stream digital audio and video in numerous formats, is included in the I3D. The *libavcodec* itself is a collection of codecs for audio and video material [w-ffm].

5.1 MPEG-4 Systems

To explain the requirements for the architecture of any MPEG-4 player, be it implemented as PC software or be it a hardware device, an excerpt from the introduction to the ISO/IEC 14496-1 standard is reproduced here:

> "The information representation specified in ISO/IEC 14496-1 describes the means to create an interactive audio-visual scene in terms of coded audio-visual information and associated scene description information. The entity that composes and sends, or receives and presents such a coded representation of an interactive audio-visual scene is generically referred to as an "audio-visual terminal" or just "terminal". This terminal may correspond to a standalone application or be part of an application system.
>
> The basic operations performed by such a receiver terminal are as follows. Information that allows access to content complying with ISO/IEC 14496 is provided as initial session set up information to the terminal. Part 6 of ISO/IEC 14496 defines the procedures for establishing such session contexts as well as the interface to the delivery layer that generically abstracts the storage or transport medium. The initial set up information allows, in a recursive manner, to locate one or more elementary streams that are part of the coded content

representation. Some of these elementary streams may be grouped together using the multiplexing tool described in ISO/IEC 14496-1.

Elementary streams contain the coded representation of either audio or visual data or scene description information or user interaction data. Elementary streams may as well themselves convey information to identify streams, to describe logical dependencies between streams, or to describe information related to the content of the streams. Each elementary stream contains only one type of data.

Elementary streams are decoded using their respective stream-specific decoders. The audio-visual objects are composed according to the scene description information and presented by the terminal's presentation device(s). All these processes are synchronized according to the systems decoder model (SDM) using the synchronization information provided at the synchronization layer" [14496-1].

Fig. 5.1 gives an overview of the MPEG-4 layer model. The transmission / storage layer in the lower part of the figure relates to the transmission infrastructure of the ancillary layers. MPEG-4 data can be transmitted across a wealth of transmission paths, including MPEG-2 transport streams, RTP / UDP over IP, or even stored as files in the MPEG-4 file format .mp4. MPEG-4 provides a flexible multiplexing tool (FlexMux) in the delivery layer to pass on data correctly and in time to the sync layer [per04].

In the sync layer, all types of elementary streams (ES) are synchronized. In the compression layer, the ES are supplied to the respective decoders that process the data. The header of each ES contains a unique object descriptor (OD) that specifies the content of the data packets. This way, information about the receiver of the data packets is contained in the ES itself.

An MPEG-4 scene must provide at least one stream that contains the scene description, i.e. the spatial and temporal relation between the objects (see section 5.2). The scene description in turn can access the ODs in order to refer to these objects [per04]. Other ES usually present can contain e.g. control data, audio data, or video data. They form the so-called media stream that encompasses the whole MPEG-4 scene.

The required architecture results in a rich, expandable set of features offered at the system level. This creates the problem of an almost infinite wealth of abilities to be implemented for a "terminal", when offering all of the functionalities defined in the full MPEG-4 standard. Because it is neither feasible nor sensible to have a terminal that encompasses all MPEG-4 functionality, the MPEG has come up with a set of interoperability or conformance points. These are aimed to make sure that MPEG-4 products work with other MPEG-4 products from other vendors. They specify items such as codecs, bitrates, image sizes, number of objects, etc. that the terminal must be able to handle. In MPEG-4 these conformance points are organized in so-called profiles and levels.

The I3D does not specifically conform to certain profiles or levels, because it was never meant to be a commercial product. Instead, its development has been started to create a tool that allows the reproduction of interactive audiovisual content for scientific purposes. As such it has evolved significantly, but without the need for comparison with other MPEG-4 players. Nowadays it is mainly used for performing subjective assessments of perceived quality of interactive audiovisual content, and naturally its development path has been dictated by the necessities of these assessments.

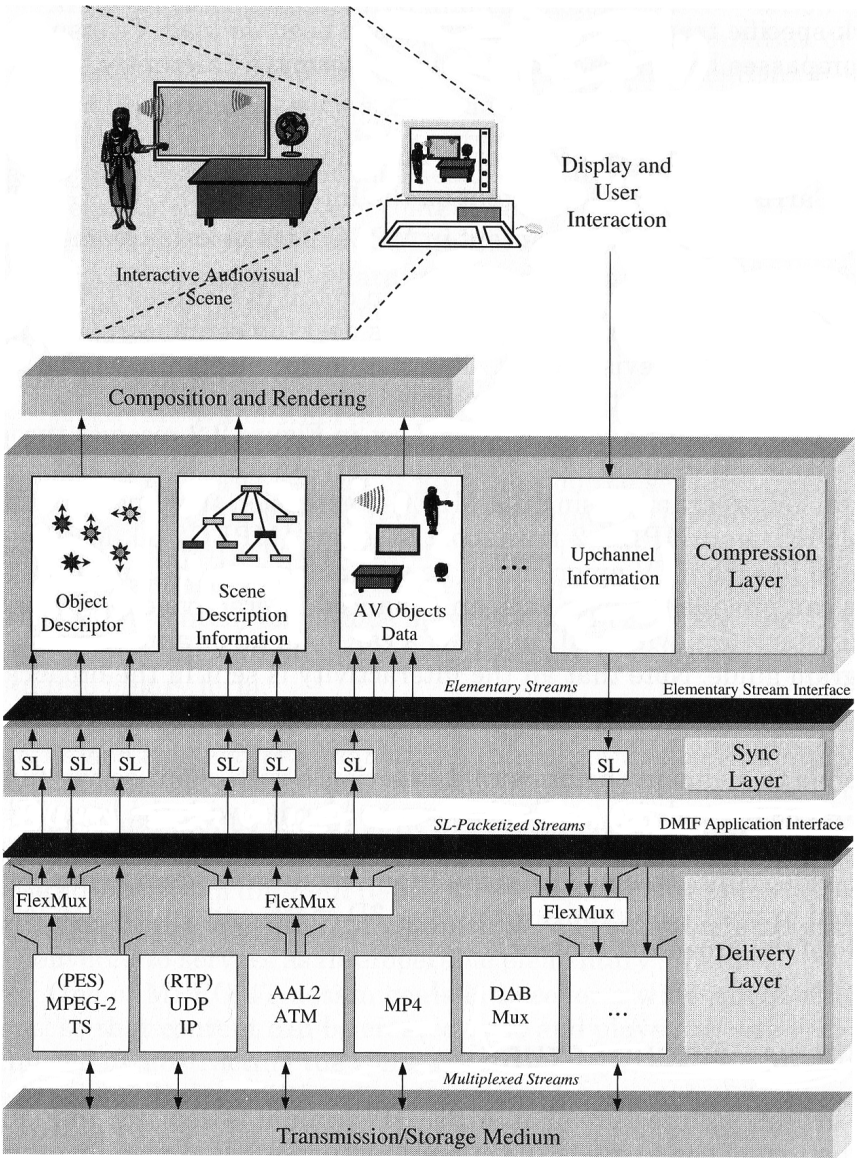


Fig. 5.1: The MPEG-4 layer model, *from [per04].*

5.2 MPEG-4 Audio and Scene Description

> "ISO/IEC 14496-3 (MPEG-4 Audio) is a new kind of audio standard that integrates many different types of audio coding: natural sound with synthetic sound, low bitrate delivery with high-quality delivery, speech with music, complex soundtracks with simple ones, and traditional content with interactive and virtual-reality content. By standardizing individually sophisticated coding tools as well as a novel, flexible framework for audio synchronization, mixing, and downloaded post-production, the developers of the MPEG-4 Audio standard have created new technology for a new, interactive world of digital audio" [14496-3].

Rather than defining the complete audio functionality of MPEG-4, *MPEG-4 Audio* merely defines the coding of the audio signals themselves. Therefore, in this thesis MPEG-4 Audio (ISO/IEC 14496 Part 3) will always be related to ISO/IEC 14496 Part 11, MPEG-4

Scene Description[1] and Application Engine[2] [14496-11]. It is the scene description that is used to define and dynamically change the auditory and visual appearance of virtual 3D spaces. Although *MPEG-4 Audio* is used for the coding of audio material, it is ultimately *MPEG-4 Scene Description and Application Engine* that controls the overall quality impression of whole scenes.

The MPEG-4 standard itself does not define the way in which the room acoustic approximation is computed (except for the so-called *Perceptual Approach*, see section 5.6). It is therefore the responsibility of the scene designer to provide information on how audio shall be rendered, and of the rendering software designer to implement a computation method that interprets the parameters defined in MPEG-4 AudioBIFS accordingly and in real-time.

The scene description format BIFS consists of a hierarchical assembly of nodes, incorporating but outperforming in number the nodes already defined in the VRML 2.0 standard [14772-1]. Scheirer et al. give a good summary of the scene description format: BIFS enables the transmission and rendering "of audiovisual scenes composited from several component pieces of content such as video clips, computer graphics, recorded sound, and parametric sound synthesis" [sch99]. AudioBIFS as a sub-part of BIFS "provides a unified framework for sound scenes that use streaming audio, interactive and terminal-adaptive presentation, three-dimensional (3-D) spatialization, and/or dynamic download of custom signal-processing effects" [sch99]. These custom signal-processing effects as well as the parametric sound synthesis are supposed to be realized by means of the *Structured Audio Orchestra Language* (SAOL), a newly designed computer programming language for audio processing purposes. SAOL is similar to other well-known computer music programming languages like CSound [sch99a], but unfortunately there is no MPEG-4 player actually capable of rendering SAOL scenes of more than low complexity. This is mainly related to the fact that SAOL is an interpreter language, that is it needs to be interpreted and translated into machine code in real-time, which makes it prohibitively slow. A number of decoders for *Structured Audio* (SA) have been presented that translate SAOL programs e.g. into C programs (sfront, see [laz01]) or Java programs [sux04], but they are either thought to be used as stand-alone applications or they have vanished because of lack of interest or technical issues. All other features mentioned in Scheirer's summary have been proven to be actually working and feasible as defined in the MPEG-4 standard, in part by the work described in this thesis.

5.2.1 BIFS Nodes

A BIFS scene is, as already stated, a composition of hierarchically arranged nodes. A node consists of numerous fields which are the attributes and the interface of this node. A field of a node can be assigned to one of the field types in table 5.1.

The data types of node fields can be Single-value Fields (SF) or Multiple-value Fields (MF). Fields that can contain only one value are Single-value fields (e.g. SFFloat). A single value may be either of a simple data type, like a boolean or a float value, or it may be of a more complex data type, like a more-dimensional vector. In case of complex data types, the single value is composed of several numbers, which are separated by a whitespace. Fields of the data type MF (e.g. MFFloat) can contain an array of single fields. The most

[1] also called BIFS, BInary Format for Scenes
[2] MPEG-J, Java API for runtime control of audiovisual objects

Field Type	Description
eventIn	event received by the node; other nodes can set this field
eventOut	event sent by the node; other nodes can receive values from this field
exposedField	public member of the node; other nodes are free to set these fields and also to receive them
field	private member of the node; no access from other nodes

Table 5.1: Field types of BIFS nodes in MPEG-4. eventIn/-Outs and exposedFields are connected via Routes as described in section 5.3.2.

common data types are presented in table 5.2. A number of other data types exist that are not mentioned here.

Data Type	Description	Example
SFBool	Boolean value	TRUE, FALSE
SFFloat	32-bit floating-point values	3.141
SFInt32	32-bit signed integer values	3
SFNode	VRML node	Sphere
SFString	sequence of characters	"Example"

Table 5.2: Data types of fields in MPEG-4 nodes.

Here, only an overview of the most important BIFS scene description nodes in the context of subjective assessments of interactive audiovisual scenes can be given. For the full MPEG-4 scene description node reference see [14496-11.1].

Group A Group node is placed first in the scene hierarchy and is the parent node for all other nodes. It possesses a field *children* to which all other nodes that describe objects in the scene are subordinated.

Transform The Transform node creates a 3D coordinate system for its children and can apply three types of transformations to these: translation, rotation, and scale. It is also a parent for subordinated nodes and inherits its properties to these.

Shape The Shape node defines visual objects and includes both, the geometry of the object and its visual properties. These properties are e.g. color or texture mapped onto the object.

Geometry The object's geometrical shape is defined in the Geometry node. MPEG-4 provides four primitives: cube, sphere, cylinder, and cone. More complex shapes can be defined by creating Indexed Face Sets that contain the coordinates of the surface points. BIFS allows to use prototyping, i.e. the multiple use of objects and their properties once defined in the scene description.

ViewPoint ViewPoints define certain locations in the scene from which a viewer can watch. ViewPoints can be switched or interpolated between, such that a viewer can change his location 'jumping' or 'walking'. When enabled, a viewer can also move freely in the virtual scene.

5.2.2 AudioBIFS Nodes

Some of the audio related nodes provided by MPEG-4 AudiBIFS are briefly described in this section. The implementation of audio functionality discussed in the following sections 5.4 to 5.9 makes use of scene descriptions containing these nodes. Again, for a complete reference turn to [14496-11.1].

AcousticScene The AcousticScene node determines the spatial extent to which the acoustic properties defined in the AudioBIFS nodes of the scene are valid.

AudioSource The AudioSource node is used to incorporate an audio stream into a scene. This node may be used as the source for a sound node, e.g. the DirectiveSound node defined below. The *url* field determines which ES is attached to the node. Start- and stop-time for playing can be set, as well as pitch and speed.

DirectiveSound This node determines if a sound source emits omni-directionally. Alternatively, it defines its radiation pattern. It further enables a distance dependent attenuation and air absorption modeling, as well as incorporating the propagation delay between the source and the listener. Two room acoustic rendering methods can be selected: either the so-called *Perceptual Approach* or the *Physical Approach*, see section 5.4.

AcousticMaterial The AcousticMaterial node contains all fields of the Material node plus those defining the sound reflection and transmission properties of objects' surfaces. The room acoustic computation in the MPEG-4 Audio *Physical Approach* is based on these properties.

PerceptualParameters The PerceptualParameters node contains the parameters necessary for the generic reverberation algorithm in the *Perceptual Approach*, see 5.6.

ListeningPoint This node specifies the reference position and orientation for spatial audio presentation. Usually, the ListeningPoint (the listening position of the user) is slaved to the active ViewPoint, see above.

5.3 Manipulation and Control of Audiovisual Scenes

As described above, the BIFS scene description concept realized in MPEG-4 allows the user to navigate through 3dimensional virtual worlds, to interact with objects, and thus in turn to influence these worlds to a degree previously defined by the scene author. For this, so-called sensors are used.

5.3.1 Sensors

Sensors are nodes that can be triggered by events in order to generate other events. These in turn can be used to control the course of events of a scene. Two categories of sensors exist. The first category of sensors reacts to user input. These can be used to provide an interactive user interface. An example for a sensor of this category is the *TouchSensor*. It verifies whether the user hovers over an object with the mouse or if he has clicked on the

object. Another example is the *ProximitySensor* which emits an isActive event whenever the user is located in a cuboidal region around the sensor's mid point. As long as the user moves inside this region, the *ProximitySensor* permanently emits so-called eventOuts, i.e. notifications that the user's position and / or orientation has changed. *PlaneSensors* and *SphereSensors* can be used to project a dragging motion of the mouse onto a translational or rotational movement of the local coordinate system, respectively. By this, they provide navigation possibilities to the user of a scene.

The second category contains sensors that are triggered by events that cannot be controlled by a user. The *TimeSensor* is an example for this, as it is evoked by the time passing and may generate single or repeated events at pre-defined points in time.

5.3.2 Routes

The events generated by sensor nodes can be passed on to other nodes. For this, so-called Routes are used. As an example, a ProximitySensor may trigger the playback of a video object, in which case the route command would look like described in listing 5.1.

```
1 | ROUTE ProximitySensorObject.enterTime TO VideoObject.startTime
```

Listing 5.1: An example for a ProximitySensor triggering the playback of a video object with the help of a Route.

5.3.3 Interpolators

Routes can also be used to realize animations, e.g. by routing the output of a TimeSensor node to a so-called Interpolator node that in turn is routed to an object. MPEG-4 provides a number of Interpolators that can be used in the I3D. These are the *ColorInterpolator* (interpolates between RGB color triples), *CoordinateInterpolator* (interpolates between points or vectors in 3D space), *NormalInterpolator* (interpolates between normals of a plane), *OrientationInterpolator* (may be used to produce rotational movement), *PositionInterpolator* (interpolates between locations in 3D space, may be used to continuously relocate objects), and *ScalarInterpolator* (interpolates between floating point values).

5.3.4 Valuators

The Valuator node is a very flexible node that can act as a type casting method. It enables conversion of data types. This is especially important for the use of Routes, where eventOut and eventIn data has to be of the same type. A Valuator receives an event of any data type and returns an eventOut of different type. Additionally, an input value can be multiplied by a factor, and an offset can be added to the output value. Therefore, the output value is dependent on the input value according to eq. 5.1:

$$output_i = factor_i \times input_i + offset_i \tag{5.1}$$

Listing 5.2 shows an excerpt of the node definition of the MPEG-4 Valuator node. If sum is TRUE, then a multiple-component input value (e.g. MFFloat) is converted to a single component value by adding all components. An eventIn can have up to four components (e.g. when it is of type vector).

```
1  Valuator {
2    eventIn       SFFloat    inSFFloat
3    ...
4    eventOut      SFFloat    outSFFloat
5    ...
6    exposedField  SFFloat    factor1       1.0
7    ...
8    exposedField  SFFloat    offset1       0.0
9    ...
10   exposedField  SFBool     sum           FALSE
11 }
```

Listing 5.2: Excerpt of the node definition of the MPEG-4 Valuator node.

5.3.5 ECMA Script

For more complex functions, BIFS introduces the Script node that is based on the ECMAScript standardized in ISO/IEC 16262 [16262]. Similar to ordinary BIFS nodes, Script is a node that possesses fields, values, and events. A function of a Script can be accessed via a Route when it has been declared as eventIn in the beginning. Variables that should be passed on need to be declared as eventOuts. Global variables must be declared as fields, followed by a type definition and an initial value, if applicable.

The main use of the Script node is to define complex, mathematical functions. By using if-instructions and loops, objects can be manipulated efficiently and with great flexibility. Only because of the Script functionality provided by BIFS, MPEG-4 scenes can be used as the basis for subjective assessments as described in chapter 8.

5.3.6 BIFS-Commands and BIFS-Anims

An important feature for complex modifications of the current scene are the so-called BIFS-Commands. With these, nodes, fields, values, and routes can be inserted, deleted or replaced dynamically. This allows to modify the scene itself and its behavior in dependence on e.g. the user's input. In the context of subjective assessments, BIFS-Commands could be used to modify the items (e.g. different degrees of complexity of a room acoustic simulation method) between trials.

When a continuous change in the scene's appearance is desired, this can be achieved with BIFS-Anim streams (animation streams). A BIFS-Anim stream is an ES containing modifications of e.g. location, rotation, scaling, color, etc. of objects present in the scene.

It is important to note that both BIFS-Commands and BIFS-Anims are extrinsic functionalities, i.e. the changes in the scene have an external source. This is opposite to the other scene modification possibilities (that also exist in VRML) mentioned above. These are intrinsic, i.e. the nodes responsible for the modification of the scene (like TimeSensor and Interpolator nodes) must be present in the scene description before the actual animation or modification starts [per04].

5.3.7 Conditionals

The Conditional node can be used to execute BIFS-Commands. For this, it needs to be activated by an event. Consequently, with the help of a Conditional node, events can update, erase or insert new nodes. Conditionals can be chained, and the status of the BIFS-Command (currently active or already executed) is indicated in the *isActive* field.

5.3.8 User Input

A number of input devices can be used directly to interact with audiovisual scenes in the I3D. Apart from the computer mouse and keyboard, currently these are the 3DConnexion SpaceNavigator [w-3dc], the Nintendo Wii Remote control [w-wii], and the Joytech Dancing Mate [w-joy]. Furthermore, also the Input Device described in section 6.3 (originally designed for subjective assessments of perceived quality) can be used to interact with scenes. Because the Input Device is MIDI-based[3], all other controllers capable of generating MIDI messages (piano keyboards, touch pads, modulation wheels, benders, joysticks, sliders, breath controllers, data gloves, etc.) can also be used.

5.4 Room Acoustic Rendering in MPEG-4

MPEG-4 offers two basic methods to provide a scene with room acoustic information and to render this information audible. The first method relates the acoustical characteristics of the room to its visual appearance - the so-called *Physical Approach*. The geometry of the room is the basis for the acoustic rendering, and any geometry-based method might be used for the computation of reverberation (see section 3.2). How this method is implemented in the I3D player is described in detail in section 5.7.

The second method is the so-called *Perceptual Approach*, a generic reverberation algorithm originally created by IRCAM and France Telecom, see section 3.3. Its implementation in the I3D player is described in section 5.6.

The two approaches differ greatly in terms of complexity and reference to reality. The geometry-based *Physical Approach* promises an early reflections and reverberation computation that relates at least to a certain degree to the virtual room presented visually. The *Perceptual Approach* is usually less demanding computationally-wise, but cannot e.g. account for changes in the early reflection pattern when the user moves through the room. It has been shown that especially the patterns of early reflections are of utmost importance for the perceived quality of artificial reverberation (e.g. [beg96, gri99]), for the perceived shape of a room (e.g. [kut91, suz01]), and for speech intelligibility (e.g. [bra03]) in the audio-only, unimodal case. Whether they are equally relevant for the overall quality perception in the audiovisual, bimodal case remains to be examined.

5.5 TANGA Modular Real-time Audio Rendering Engine

TANGA is a tongue-in-cheek acronym for *The Advanced Next Generation Audio*. It is an object-oriented, platform independent, and modular software framework used to model real-time audio signal processes in C++. TANGA is designed to allow an abstract representation of these processes. It encapsulates the underlying signal processing, thus providing a uniform, relatively simple to use interface to its functionality. A major incentive for using a software framework usually consists in the straightforward reusability of software components. This concept is consistently applied in the TANGA.

By means of using interfaces, additional functionality can be integrated, and new signal processing algorithms can be realized easily. In the TANGA system, the processing of audio signals is done through so-called TANGA *Components*. Each TANGA Component constitutes a signal processing unit with a given number of input and output channels. The

[3]For a short introduction to MIDI refer to section 6.1.1.

audio signal arriving at the input is transformed by the signal processing logic implemented in the Component. It is then available (e.g. for further processing in other Components) at the output of the Component. So-called ComponentConnectors provide signal connections between Components. MessageDispatchers provide control message connections between Components. The TANGA System can be expanded to virtually any audio functionality by means of writing a new 'plug-in' (read: Component). Because of this, TANGA is a most powerful and flexible software system.

Fig. 5.2 shows a schematic view of the algorithm host model as used in the TANGA system. In the TANGA system, the hardware (sound card) related parts are strictly separated from those related to the functionality (signal processing algorithm). Therefore it is divided into four entities: The Hardware, the Host, the Engine and the Signal Processing modules.

5.5.1 TANGA Hardware Module

The hardware module uses the sound card drivers registered in the system for audio input and output to any multichannel sound card. As it is linked to the host, the host has access to the hardware via the drivers. Functionality such as opening a driver and obtaining information on the current hardware setup is provided. The TANGA host module supports various drivers and is cross platform compatible. The performance of a single sound card can vary for different drivers, depending on the quality of the driver implementation. The drivers supported are ASIO[4], WMM[5] and DirectSound[6] on Windows and ALSA[7] on Linux systems.

Especially the ASIO API (and thus the drivers based on that API) is very suitable for real-time audio applications, because it provides the necessary functions and callback mechanisms to handle two-layer buffers: one for the audio input, and one for the audio output. Therefore, while the TANGA is writing to the output buffer 0, the sound card driver is writing to the input buffer 1. After that, beginning with the following frame, the buffers are switched to output buffer 1 and input buffer 0, respectively [w-asi]. The length of the buffers can be as low as 64 samples on some sound cards[8].

Upon the first start of the I3D, the hardware is instructed to describe the capabilities of the sound card (number of input and output channels available, drivers available in the system, and available sample rates). The user must choose from and set these parameters, which are then registered with the system.

5.5.2 TANGA Host Module

The host module implements the bridge between the hardware and the TANGA Engine. It consists of the PortAudio Application Programming Interface (API) [ben01]. Commu-

[4] Audio Streaming Input Output, an API defined and proposed by Steinberg Media Technologies AG, see [w-asi]. Letz [let01] describes a port of the PortAudio API using the ASIO API on Macintosh and Windows systems.

[5] Windows MultiMedia

[6] DirectSound is part of the Microsoft multimedia framework DirectX. For a current description of the API see [w-msd].

[7] Advanced Linux Sound Architecture, see [w-als].

[8] The buffer sizes depend on the efficiency of the drivers of the sound card. A 64 samples buffer at a sampling rate of $48kHz$ is equivalent to a latency of around $1.33ms$. Usually, larger buffer sizes are chosen to decrease the system load caused by the repeated buffer accesses.

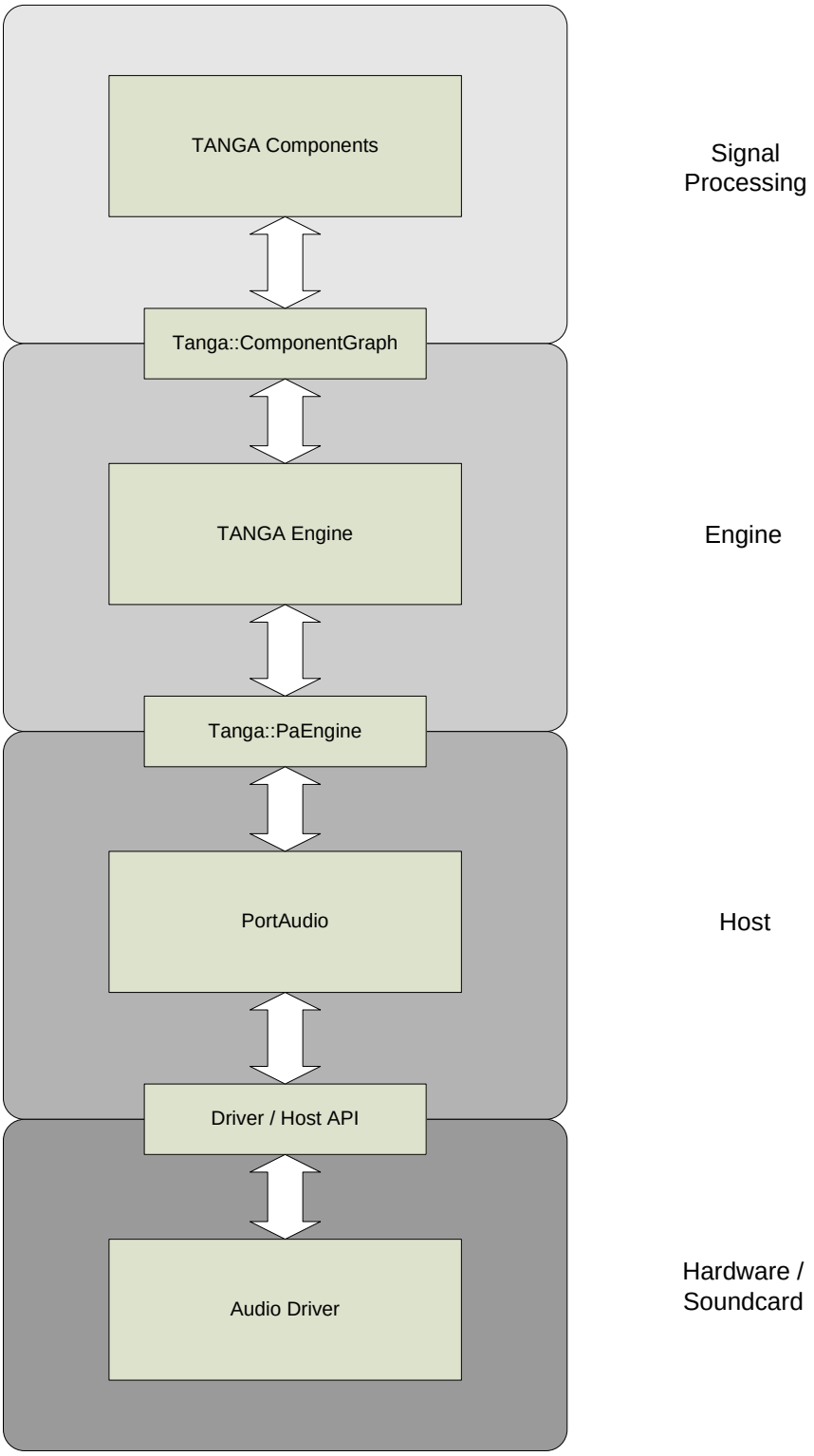


Fig. 5.2: The algorithm host model as used in the TANGA system.

nication between hardware and host is realized via the driver.

One of the most important requirements for an audio API to be used in the TANGA System is that it should provide a DAC[9] output time stamp. The output time stamp identifies the time when the buffered samples are actually played at the audio output of the sound card. Such a feature is essential for synchronization purposes. PortAudio was chosen because it provides such a time stamp and has in general a very good support for real-time operations. Whereas PortAudio also provides audio streams in blocking read / write mode, this feature is not useful for the TANGA System. TANGA relies on the non-blocking audio streams which use a callback method for filling the output buffers. The callback function invoked by PortAudio is used to control the TANGA Engine. PortAudio ensures that this function is always called in time such that the hardware output buffers are filled as needed by the sound card, to continuously output the correct audio signal.

Before signal processing starts in the signal processing module, the hardware is instructed by the host to open the driver specified in the registry. Then the callback function is activated and called whenever the output buffer of the sound card needs to be filled. The callback function in turn calls the perform method of the Tanga::ComponentGraph class to do the actual audio processing, see section 5.5.4.

5.5.3 TANGA Engine Module

The engine module is responsible for controlling the signal processing module. At the same time it has to pass on the computed audio samples to the host module for audio output on the sound card or to a sound file. Thus the engine module has to cope with different use cases, a fact that suggests implementing the engine as an abstract interface class.

Fig. 5.3 shows the Unified Modeling Language (UML) class diagram of the classes Tanga::Engine and Tanga::PaEngine. Tanga::PaEngine is an implementation of the Tanga::Engine interface class. It uses the API specified in the host module, here the PortAudio API. Alternative APIs can easily be used when Tanga::PaEngine is modified accordingly. Tanga::PaEngine has a member of the class Tanga::ComponentGraph (m_componentGraph), in which a ComponentConnector registers the signal paths that exist between signal processing Components (see section 5.5.4). m_outputBuffers contains the buffers that the final component in the signal processing chain is writing the samples into.

5.5.4 TANGA Signal Processing Module

The signal processing module contains the actual audio processing functionality of the TANGA system. It is connected to the engine module via the Tanga::ComponentGraph class. Tanga::ComponentGraph is one of the most important elements of the TANGA framework. It abstracts the connections between Components that make up the Signal Flow Chart (SFC, see fig. 5.8) of the signal processing algorithm to form a directed graph, the so-called Component Graph. The Component Graph consists of nodes (the Components) and edges (the signal connections between the Components). The edges of the Component Graph are directed against the signal flow direction, and only ever one edge exists between two nodes, independently from the actual number of signal connections (number of channels) between the respective Components.

[9] Digital to Analog Converter

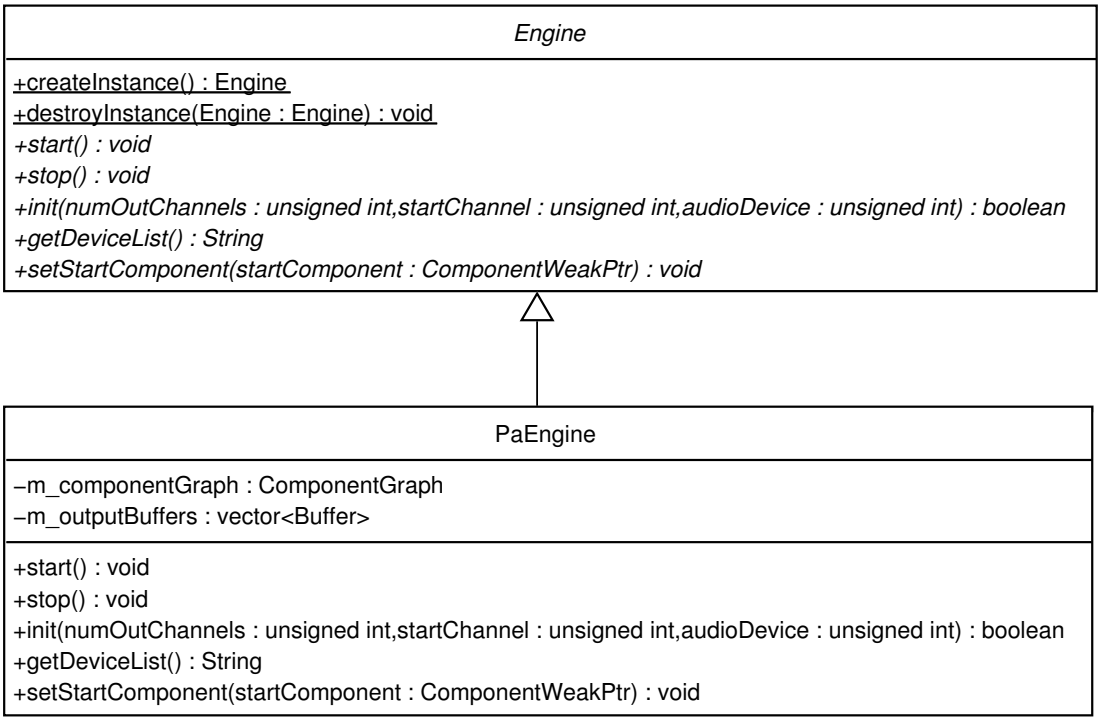


Fig. 5.3: UML class diagrams of abstract interface class Tanga::Engine and implemented class Tanga::PaEngine. *From [par07].*

5.5.5 TANGA Class Overview

Signal processing module and engine module constitute the heart of the TANGA framework. They represent the literal core of the real-time audio processing system. The framework's classes are organized in three packets: TangaBase, TangaComponents, and TangaUtils. Fig. 5.4 shows an overview of the most important classes contained in TangaBase and TangaComponents, and their dependencies.

As can be seen, the TangaBase packet contains all classes related to the organization of a signal process. The relation of Tanga::Engine, Tanga::PaEngine and Tanga::ComponentGraph classes has already been described in section 5.5.3. The Tanga::ComponentConnector class provides methods for creating and destroying signal connections between Components. Whenever such a connection is created, it is registered as an edge in the Component Graph.

The Tanga::Buffer class is a data class that provides temporary storage for the audio processing by encapsulating a field of audio samples. In principle, this field can be arbitrarily long, its size correlates with the block length used for the processing. TANGA is usually set to work with a block length of 1024 samples. Each instance of the Tanga::Buffer class correlates to one audio channel. Therefore, if two Components are connected using eight channels of audio, eight instances of Tanga::Buffer have to be created.

The Tanga::MessageDispatcher is very similar to the Tanga::ComponentConnector, only that it does not connect Components for transmission of audio signals but for control messages. Therefore, MessageDispatcher connections are not included in the Component Graph (they are not registered with the Tanga::ComponentGraph class). Whenever there is a need to continuously transmit control messages to a Component, the sending and receiving Components must be additionally connected using the Tanga::ComponentConnector

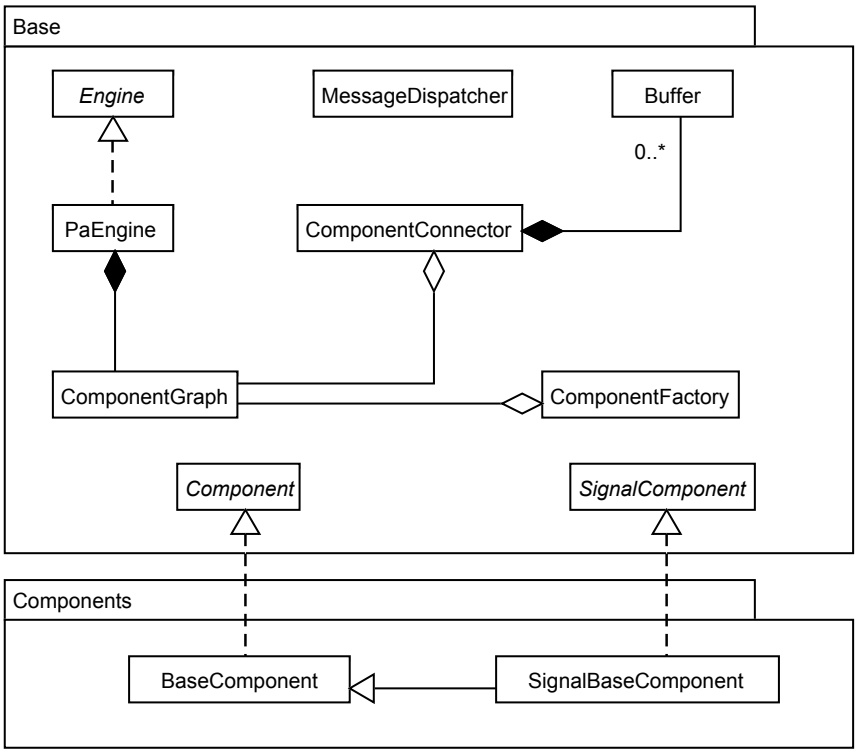


Fig. 5.4: An overview of the most important classes of the TANGA framework given in UML notation. *After [par07], modified.*

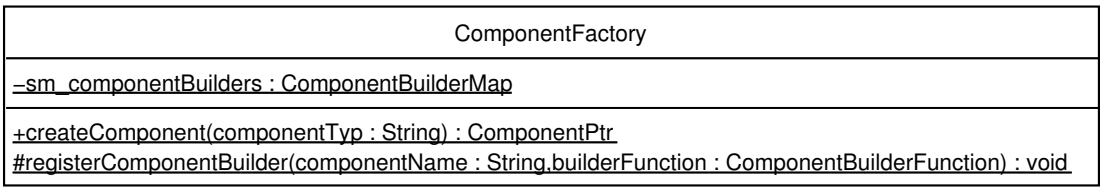


Fig. 5.5: UML class diagram of static class Tanga::ComponentFactory. *From [par07].*

class.

In the TANGA framework the so-called factory design pattern (also called factory method pattern) is used for creating Components. This is an object-oriented design pattern, and like other creational patterns, it deals with the problem of creating objects without specifying the exact class of object that will be created. The factory method design pattern handles this problem by defining a separate method for creating the objects. Subclasses can then override to specify the derived type of object that will be created. In order to create a Component, the class Tanga::ComponentFactory is used. It is the only instance in the TANGA framework that can create Components. It possesses only static methods and attributes, see fig. 5.5. Whenever Tanga::ComponentFactory receives a call to create a Component via its createComponent method, the string contained in the call is analyzed. This identifies the constructor to be executed. The created object is then returned as a pointer (ComponentPtr) based on a smart pointer class[10].

Tanga::Component and Tanga::SignalComponent are the base classes for all Components

[10] A smart pointer is an abstract data type that simulates a pointer while providing additional features, such as automatic garbage collection or bounds checking. Using smart pointers is an elegant way of avoiding memory leaks while retaining efficiency [mey05, w-col].

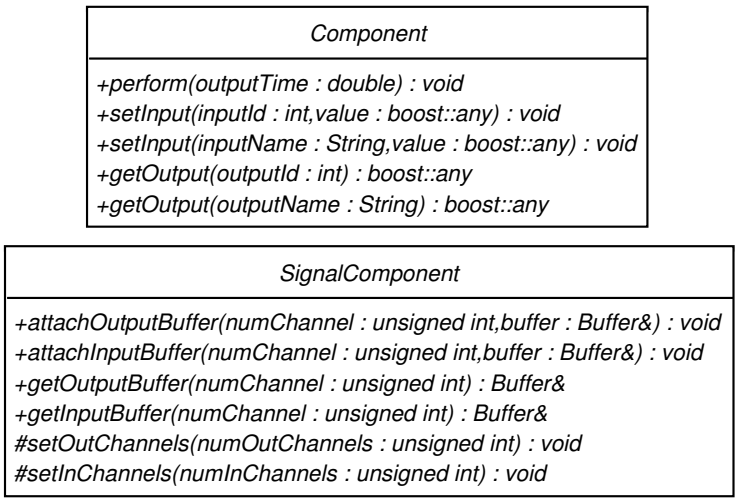


Fig. 5.6: UML class diagram of abstract classes Tanga::Component and Tanga::SignalComponent. *From [par07].*

in the TANGA framework. They are declared as abstract classes and only contain pure virtual methods. Therefore they act as so-called interface classes which hide the actual implementation details of the components from other parts of the system while providing access to all the necessary functionality. The Tanga::Component class defines the basic operations that are necessary to access the parameters of a Component. These parameters can be set or requested either via an ID or via the parameter's name contained in a string, see fig. 5.6. The most important method in Tanga::Component is the perform method which performs the actual control message or audio signal processing computation of a Component. The Tanga::SignalComponent class provides methods for assigning buffers to a Signal Component, and for reading and writing the buffers upon execution of the perform method (fig. 5.6).

The TangaComponents packet contains all Components available in the TANGA framework. Each Component is based on a template class Tanga::BaseComponent or Tanga::SignalBaseComponent. All Components created with the template Tanga::BaseComponent are so-called Message Components, as opposed to Signal Components that are created based on the two template classes Tanga::SignalBaseComponent and Tanga::BaseComponent (multiple inheritance, see fig. 5.4). Only Signal Components actually perform audio signal processing, whereas Message Components can only perform control message manipulations. Fig. 5.7 shows the UML class diagram of the template classes Tanga::BaseComponent and Tanga::SignalBaseComponent.

Finally, the packet TangaUtils contains two types of classes. One type are helper classes that can be used to compute control data for the parameters of the Components. An example for this type of class is Tanga::FrequencyGains which computes filter coefficients based on a given set of frequency / gain pairs as defined in the MPEG-4 AcousticMaterial and DirectiveSound nodes, see [14496-11.1]. The other type are classes that create groups of Components (sub-graphs of a Component Graph) to represent a signal process that is always set up the same way. These classes might be regarded as macro classes, because upon calling them a certain set of Components is created and connected in a pre-defined fashion. An example for this is the MPEG-4 DirectiveSound node that is implemented in the TANGA as a group of Components, see fig. 5.10. A Util class can also call another Util class, thus generating a large network of Components while keeping

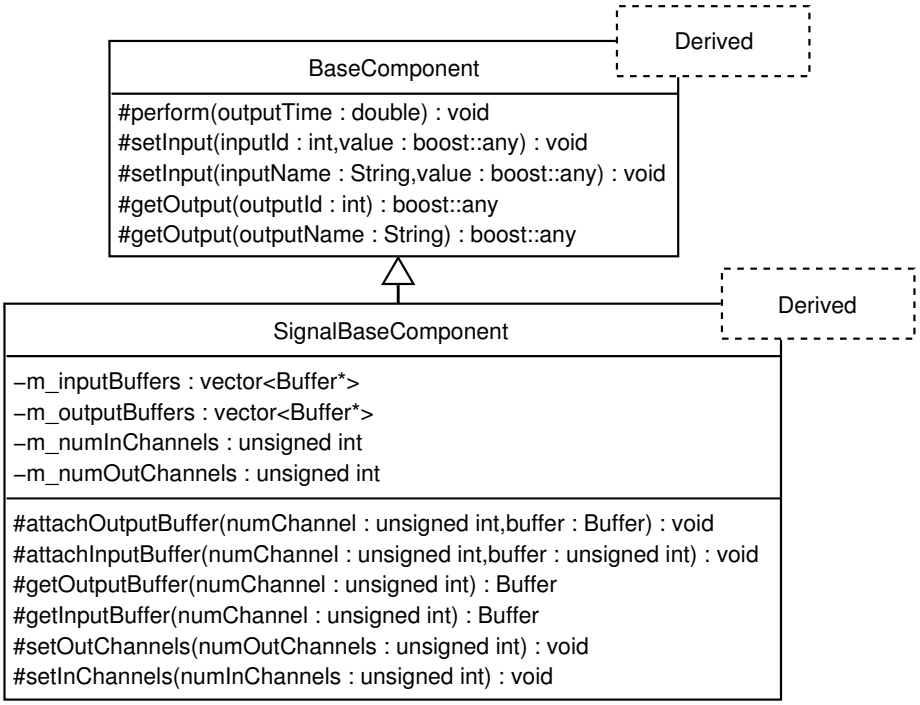


Fig. 5.7: UML class diagram of the template classes Tanga::BaseComponent and Tanga::SignalBaseComponent. *From [par07].*

the structure of this network simple and well defined. Section 5.6.3 describes in detail the Tanga::PerceptualSource Util class that creates a generic reverberation module by applying this methodology.

5.5.6 Signal Flow Chart and Component Graph Example

An example for a combination of Components accounting for a very simple Signal Flow Chart (SFC) is given in fig. 5.8. In this example, the correct panning of two sound sources (Components *Tanga::SoundFile*) for a four channel loudspeaker setup (positions defined in *Tanga::LSSetup*) is provided. Audio is sent to the loudspeakers from a *Tanga::Mix* Component in which the audio signals coming from the *Tanga::SoundFile* Components are mixed together. Because the panning needs to be updated continuously (the listener's location or azimuth may change anytime in interactive applications), *Tanga::Matrix* and *Tanga::Mix* Components are additionally connected by a ComponentConnector.

Fig. 5.9 shows the corresponding Component Graph. As can be seen in the example, only Signal Components (those Components that actually perform audio signal processing) are part of the Component Graph. Here, these are the two *Tanga::SoundFile* Components and the *Tanga::Mix* Component. The other Components are not involved in the actual audio signal processing.

5.5.7 Currently Implemented TANGA Components

Of the large number of audio nodes defined in the MPEG-4 standard, only a subset has been implemented in the TANGA system yet. However, it is important to note that MPEG-4 audio nodes usually have no direct correspondence to a TANGA Component. This means that often a number of TANGA Components is needed to implement one MPEG-4 audio

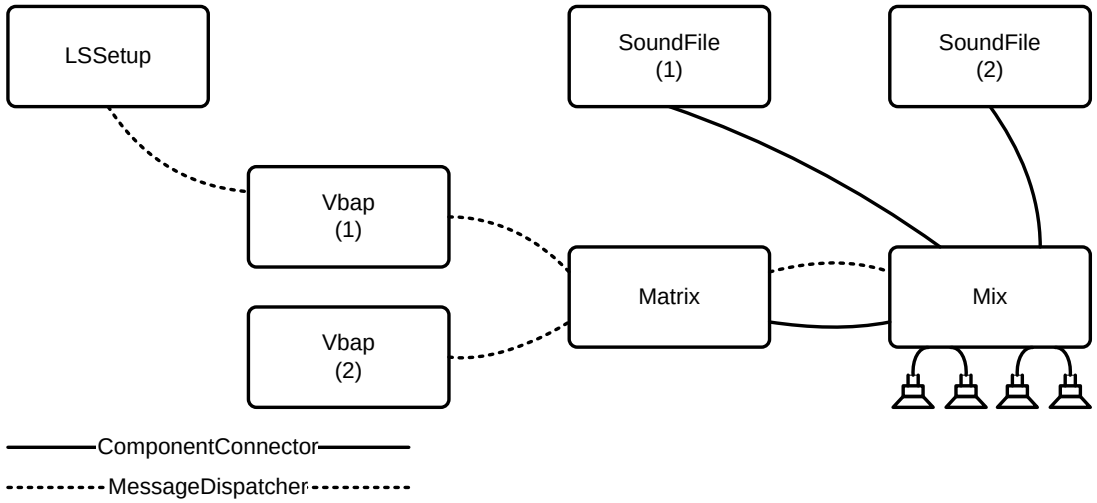


Fig. 5.8: A simple example Signal Flow Chart (SFC) for two audio signal sources being panned using the VBAP and Matrix Components.

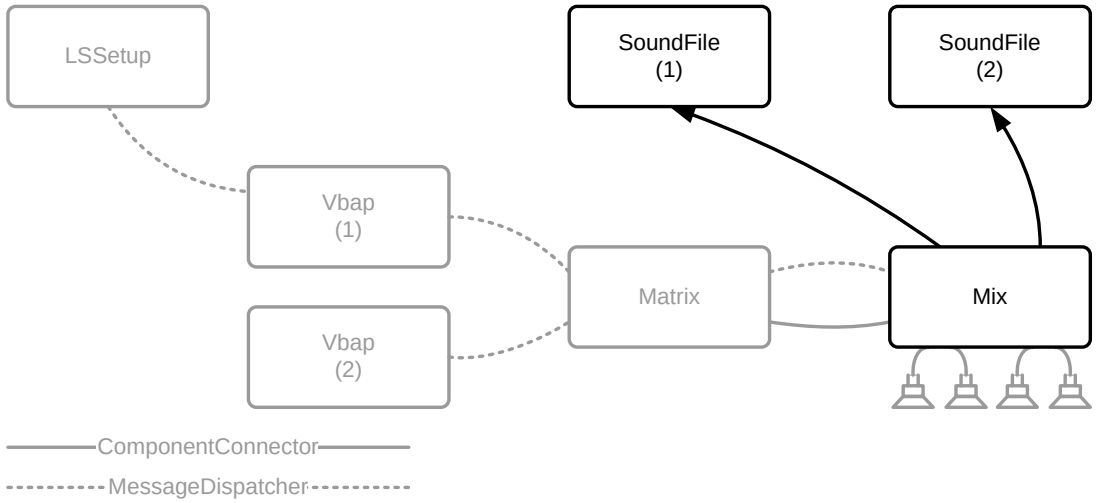


Fig. 5.9: A simple example Component Graph (solid black) corresponding to the SFC in fig. 5.8. Vertices and edges of the SFC that are not part of the Component Graph are shown in gray.

node and its functionality. A good example for this is the DirectiveSound node that can be used to obtain sound source directivity. For omni-directional sound sources the *Tanga::SoundFile* Component can be connected directly to the *Tanga::Mix* Component, see fig. 5.8. To implement a radiation pattern as specified in the MPEG-4 standard, one can simply add two *Tanga::Filter* Components along with a *Tanga::FilterBlender* and a *Tanga::Sum* Component as shown in fig. 5.10. One way of defining an MPEG-4 sound source directivity pattern is to indicate so-called reference angles along with a corresponding set of filter coefficients in the fields of the DirectiveSound node. Each reference angle then defines the exact filtering for the sound source when the listener hears the source from the reference angle. The *Tanga::FilterBlender* Component computes gain factors for both *Tanga::Filter* Components in order to interpolate between the sound characteristics given for the reference angles. E.g., when the listener is located in the middle between two such reference angles, then the gain factors for the *Tanga::Filter*

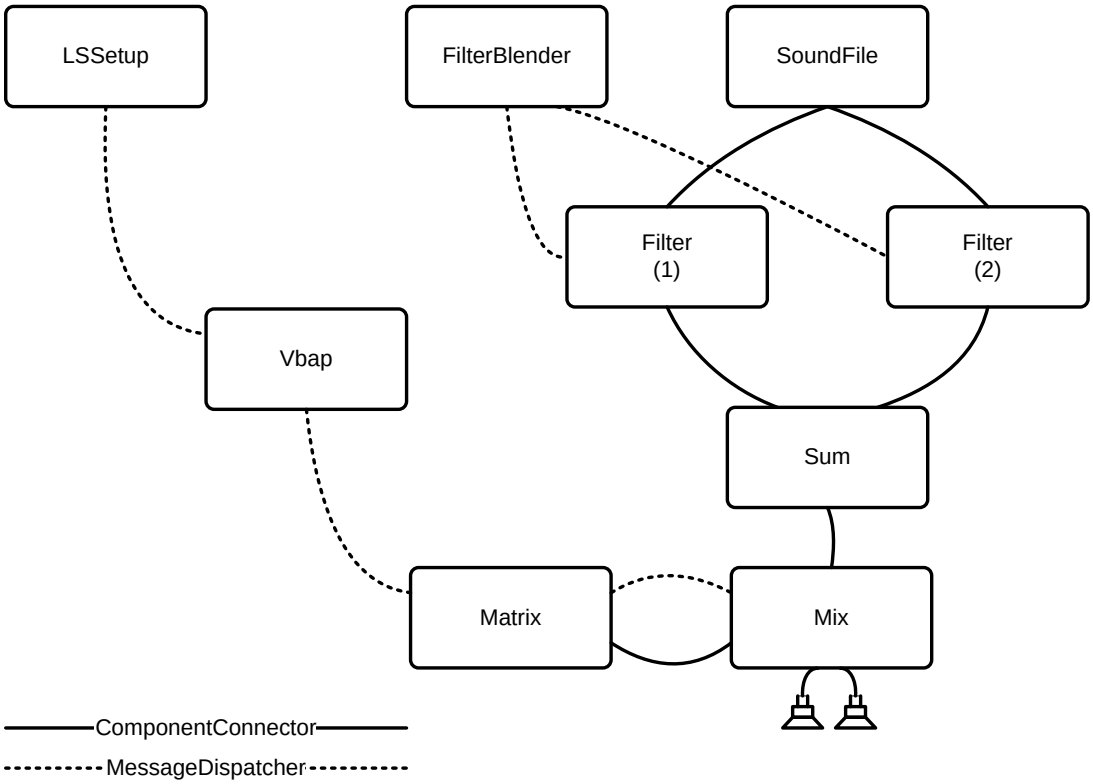


Fig. 5.10: MPEG-4 audio nodes do not correspond directly to TANGA Components; instead, a node's functionality can be composed by connecting TANGA Components accordingly. Here, the SFC implementing the DirectiveSound node is shown. *Tanga::FilterBlender*, *Tanga::Filter* and *Tanga::Sum* Components can be created automatically by calling a corresponding TangaUtil class, because these Components always have the same structure for the DirectiveSound node.

Components are both set to 0.5.

Table 5.3 lists the currently implemented TANGA Components, their functionality, and the parameters that can be controlled.

5.5.8 Scene Description to Signal Flow Chart: The TANGA Mediator

Unlike other audio processing applications, TANGA is not fed with audio and control data directly. Instead, TANGA is part of the I3D MPEG-4 player. The I3D reads, decodes and interprets a scene description including references to the audio data and its virtual environment[11]. Apart from audio related information, the scene description also contains parts related to non-audio objects. Therefore the information relevant for the audio processing needs to be identified and forwarded to the TANGA engine.

This is done by the I3DTangaMediator class. The I3DTangaMediator acts as an intermediary between the I3D MPEG-4 player and the TANGA Engine. It is used by the I3DUpdateVisitor[12] to collect information about the audio-related nodes in the MPEG-4

[11] The term 'virtual environment' embraces all factors that influence the audio rendering from a signal processing perspective, although the processing instructions may not be given directly. They may need to be derived from the 'virtual environment' / the scene description.

[12] Visitors allow the user to insert their own operations at various steps within a graph algorithm [w-boo]. In the I3D, the visitor concept provides means for traversing the whole scene graph starting at a given

Name	Type	Parameters	Functionality
Decorrelator	signal	OutChannels ScaleFactorRange	element creating decorrelated signals from one monaural source
Delay	signal	Delay DelayInMs MaxDelay	simple delay element with one in- and output each
DelayLine	signal	Delays DelayValues NumOutChannels MaxDelay	delay element with one input and an arbitrary number of outputs that can be assigned different delay times
DiffuseReverb	signal	ReverbTime ReverbGain Reflection_coeffs	element generating reverberation based on nested all-pass filters
Filter	signal	Gain Coefficients	FIR/IIR filter of arbitrary order with one input and one output
Mix	signal	OutChannels InChannels Channels MixMatrix	signal mixing element with arbitrary numbers of in- and outputs based on a mixing matrix
SoundFile	signal	Filename	sound source that reads samples from a .wav file
Sum	signal	NumChannels NumInputs Gain	simple signal summation element with arbitrary number of outputs, number of inputs is always a multiple of number of outputs
FilterBlender	message	FilterBlendCoeffs FilterBlend	control element to interpolate between two Filter Components
LSSetup	message	Directions	control element that provides the angular loudspeaker positions in the reproduction room
Matrix	message	Dimension Row Column Value	control element that embraces the signal distribution of all VBAP Components to pass it on to a Mix Component
Obstruction	message	ListenerLocation SourceLocation	control element for a Filter Component based on the calculation of visibility of a sound source (acoustic occlusion)
VBAP	message	LoudspeakerSetup AzimuthAndElevation RowNumber ListenerMatrix	panning element, computes a signal distribution on the outputs of a Mix Component based on azimuth and elevation of a sound source and the listener's position

Table 5.3: Currently implemented Components in the TANGA-System.

scene. Once the I3DUpdateVisitor has completed the scene traversal the mediator will process that information and pass it on to the TANGA engine.

More precisely: the scene decoder of the I3D reads the .mp4 file and instantiates the objects described in the file. These are stored in the computer's RAM. The I3DUpdateVisitor is then responsible for keeping the image in memory up-to-date with all the changes that might occur during the rendering of the scene (interaction, time sensor, ...). When, for example, an I3DDirectiveSound node is visited, the node will be registered with the I3DTangaMediator. Then the mediator reads the attributes from memory for the registered nodes and builds the audio rendering graph. The TANGA engine alludes to this rendering graph represented by the SFC.

5.6 Implementation of the MPEG-4 Audio Perceptual Approach

This section describes how the MPEG-4 Audio *Perceptual Approach* has been implemented in the *I3D* MPEG-4 player. Also, the basic operating mode of the *Perceptual Approach* is explained here.

5.6.1 MPEG-4 Node PerceptualParameters

The node PerceptualParameters defines the acoustic properties of a sound source located in a reverberant room, which is rendered via the *Perceptual Approach*. It contains 19 parameters, of which the first nine are so-called high-level parameters (see listing 5.3)[14496-1].

```
1  PerceptualParameters {
2    exposedField  SFFloat  sourcePresence       1.0
3    exposedField  SFFloat  sourceWarmth         1.0
4    exposedField  SFFloat  sourceBrilliance     1.0
5    exposedField  SFFloat  roomPresence         0.0
6    exposedField  SFFloat  runningReverberance  1.0
7    exposedField  SFFloat  envelopment          0.0
8    exposedField  SFTime   lateReverberance     1.0
9    exposedField  SFFloat  heavyness            1.0
10   exposedField  SFFloat  liveness             1.0
11 }
```

Listing 5.3: High-level parameter fields of the MPEG-4 PerceptualParameters node.

For the acoustic rendering, all high-level parameters are mapped under inclusion of other fields in the node PerceptualParameters to their physical counterpart. This process, often denoted as high/low-mapping, is necessary to control the reverberation algorithm used in the *Perceptual Approach*. The algorithm generates an impulse response that can be represented in four parts, as shown in fig. 5.11.

The impulse response's model consists of direct sound ($R0$), directional early reflections ($R1$), diffuse early reflections ($R2$) and diffuse late reverberation ($R3$). $R0$, $R1$, $R2$, and $R3$ are energies in three frequency bands (*low*, *mid*, *high*), which are arranged in the time limits $t1$, $t2$, and $t3$, also defined in the PerceptualParameters node.

node and invoking methods upon a node when entering or leaving it. The I3DUpdateVisitor checks for updated fields in the nodes that it visits, such that changes in the scene description are detected.

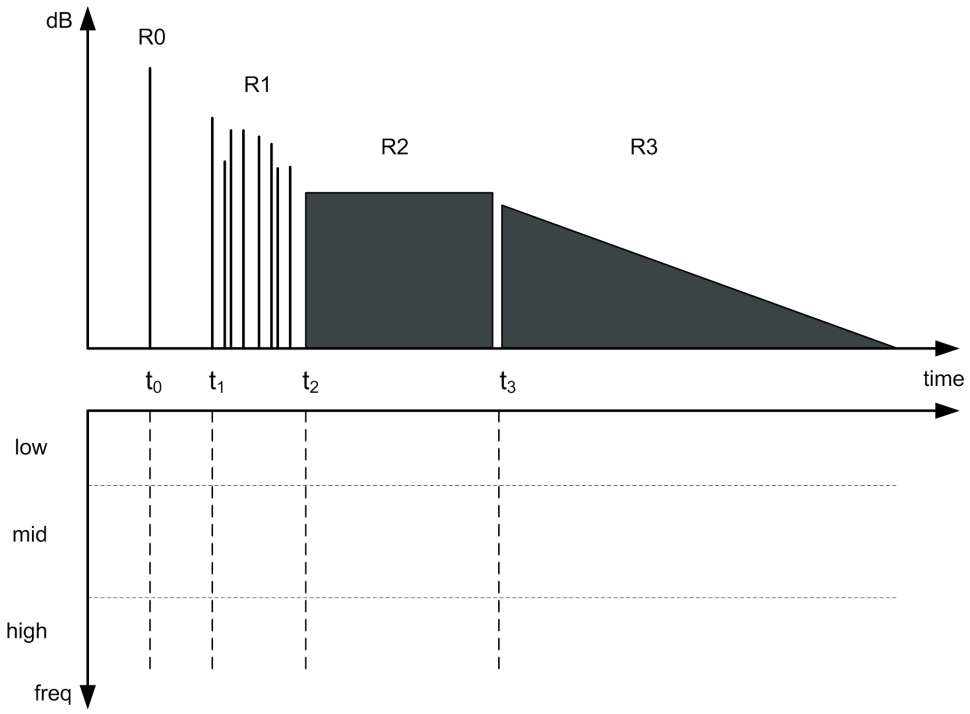


Fig. 5.11: Model of the four-part impulse response generated in the Perceptual Approach reverberation algorithm.

5.6.2 MPEG-4 Audio Implementation Example

In the second version of MPEG-4 Advanced Audio BIFS (AABIFS), an implementation example for the *Perceptual Approach* submitted by France Telecom and IRCAM was included in the standard [14496-11]. This example is subdivided into two modules: a so-called pan-module, and a room-module.

The pan-module is responsible for the panning. It defines seven signals, which have the panning positions shown in fig. 5.12. S is the location of the direct sound source with relative angular position (azimuth) C. Its panning position depends on the virtual location of the source and on the virtual position of the listener in the 3D scene. The signals L and R contain the early reflections and are always panned to $\pm 30°$ from the direct sound. $S1$, $S2$, $S3$, and $S4$ are diffuse reverberation signals with invariant absolute panning positions at $45°$, $-45°$, $135°$ and $-135°$.

Based on these definitions and the four-part impulse response of the *Perceptual Approach*, the room-module specifies a signal processing structure which generates the afore mentioned seven signals using eight internal reverberation channels (fig. 5.13).

The structure consists of the blocks *Direct*, *Early*, *Cluster* and *Reverb*. Each of these creates one part of the impulse response. *Direct* produces the direct sound with the energy $R0$. The early reflections with energy $R1$ are generated in the *Early* block by means of eight delays. In *Cluster* the signals from *Early* are mixed within a unitary Hadamard matrix. In the following they are delayed and weighted with $R2$. *Reverb* creates the late reverberation based on a Feedback Delay Network, again using a unitary Hadamard matrix. For a frequency-dependent reverberation time there are filters in each *Reverb* channel. The signals $S1$, $S2$, $S3$, and $S4$ result from the combined outputs of *Cluster* and *Reverb*.

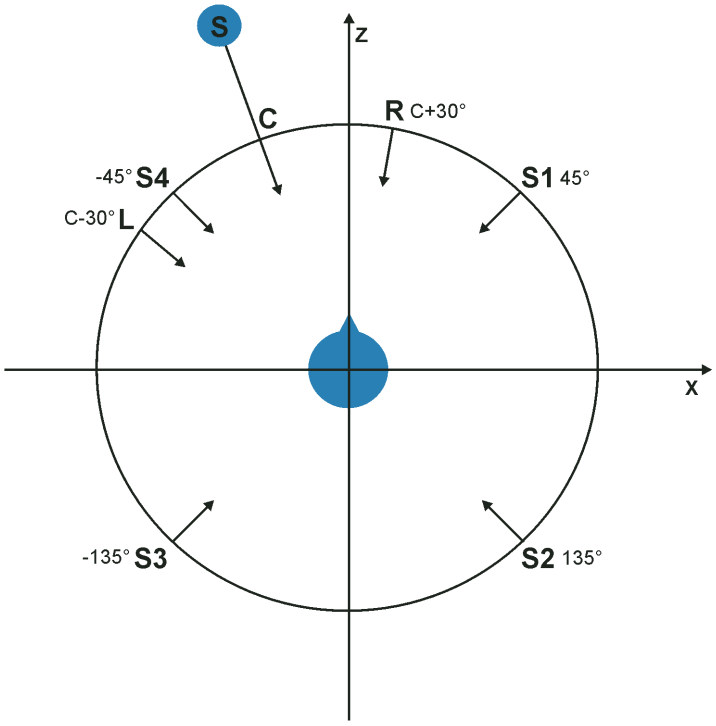


Fig. 5.12: The pan-module of the MPEG-4 AABIFS implementation example of France Telecom and IRCAM.

5.6.3 Perceptual Approach in the TANGA Audio Engine

Fig. 5.14 shows the SFC of the TANGA Util *Tanga::PerceptualSource*. It is responsible for the realization of the *Perceptual Approach* in the TANGA system in accordance to the implementation example of IRCAM and France Telecom. It can be seen that the Utils *Tanga::EarlyReflections*, *Tanga::EarlyReverb* and *Tanga::LateReverb* correspond to the Early, Cluster and Reverb blocks of the MPEG-4 Audio implementation example, respectively. As opposed to the IRCAM and France Telecom reference implementation, either two or four signals can be mixed in *Tanga::EarlyReflections*. This change has been implemented to evaluate the tonal influence of additional directional early reflection channels with different virtual positions in future tests. In the actual implementation the two additional signals, denoted L_{add} and R_{add}, have the same panning positions as L and R, respectively.

The Components *Tanga::Vbap*, *Tanga::Matrix* and *Tanga::Mix(Outsum)* in the SFC take over the panning of the direct sound. Different from the original implementation example by France Telecom and IRCAM, the implementation in the TANGA engine is scalable in order to allow further investigation of the *Perceptual Approach*. It is possible to vary the number of internal reverberation channels, also denoted as *workchannels* here. An increase in the workchannel count theoretically increases the echo density generated by the algorithm - unfortunately it also increases the computing power necessary for the computation. The internal channels can be scaled with multiples of four: e.g. 4, 8, 12, etc. workchannels can be used. Section 8.8 presents a reverberation quality assessment comparing different workchannel counts for the use in interactive audiovisual applications.

The *Tanga::EarlyReflections* Util groups the functionality necessary to create the directional early reflections ($R1$) part of the RIR. *Tanga::Filter(OmniDirectionalFilter)* ac-

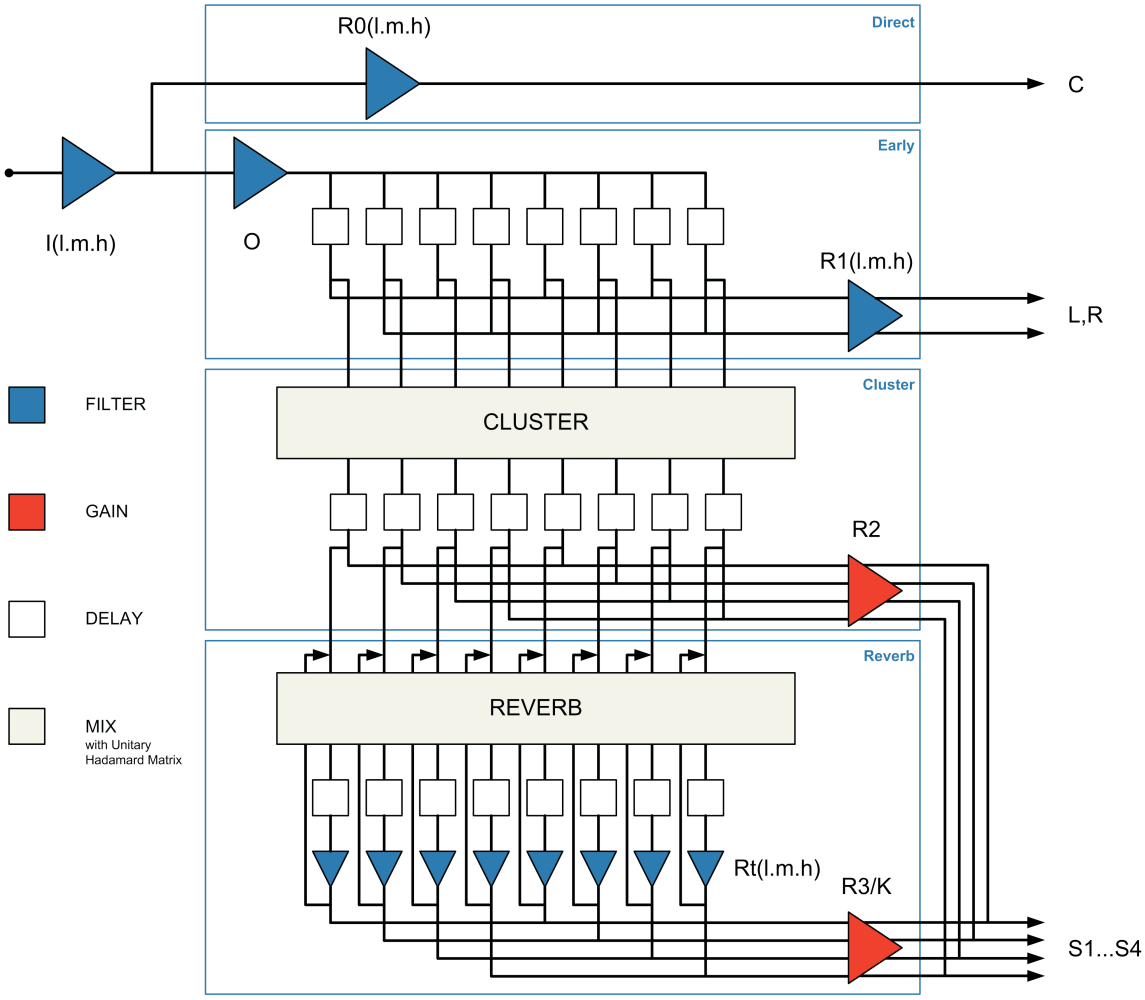


Fig. 5.13: The room-module of the MPEG-4 AABIFS implementation example of France Telecom and IRCAM.

counts for the diffuse-field spectrum of the source (which is omni-directive), whereas *Tanga::DelayLine* creates the time-delayed taps for the reflections. These are simultaneously passed on to the *Tanga::EarlyReverb* Util and to the Tanga::Mix(R1SumMix) Component that creates the four directional early reflection channels L to R_{add}. Fig. 5.15 also shows that these are filtered according to the $R1$ *Perceptual Approach* parameter (level of early reflections at the frequency ranges low, mid and high, see fig. 5.11) with the four *Tanga::Filter(R1FilterL...R1FilterRadd)* Components.

Both *Tanga::EarlyReverb* and *Tanga::LateReverb* Utils are also composed of various TANGA Components, but the details are not presented here.

5.7 Implementation of the MPEG-4 Physical Approach

The MPEG-4 Audio *Physical Approach* implemented in the TANGA consists of a simplified Image Source Method (ISM). The geometrical description of the room is taken directly from the scene description via I3DUpdateVisitor and TangaMediator classes. It is important to note that only the objects having an AcousticMaterial node attached to

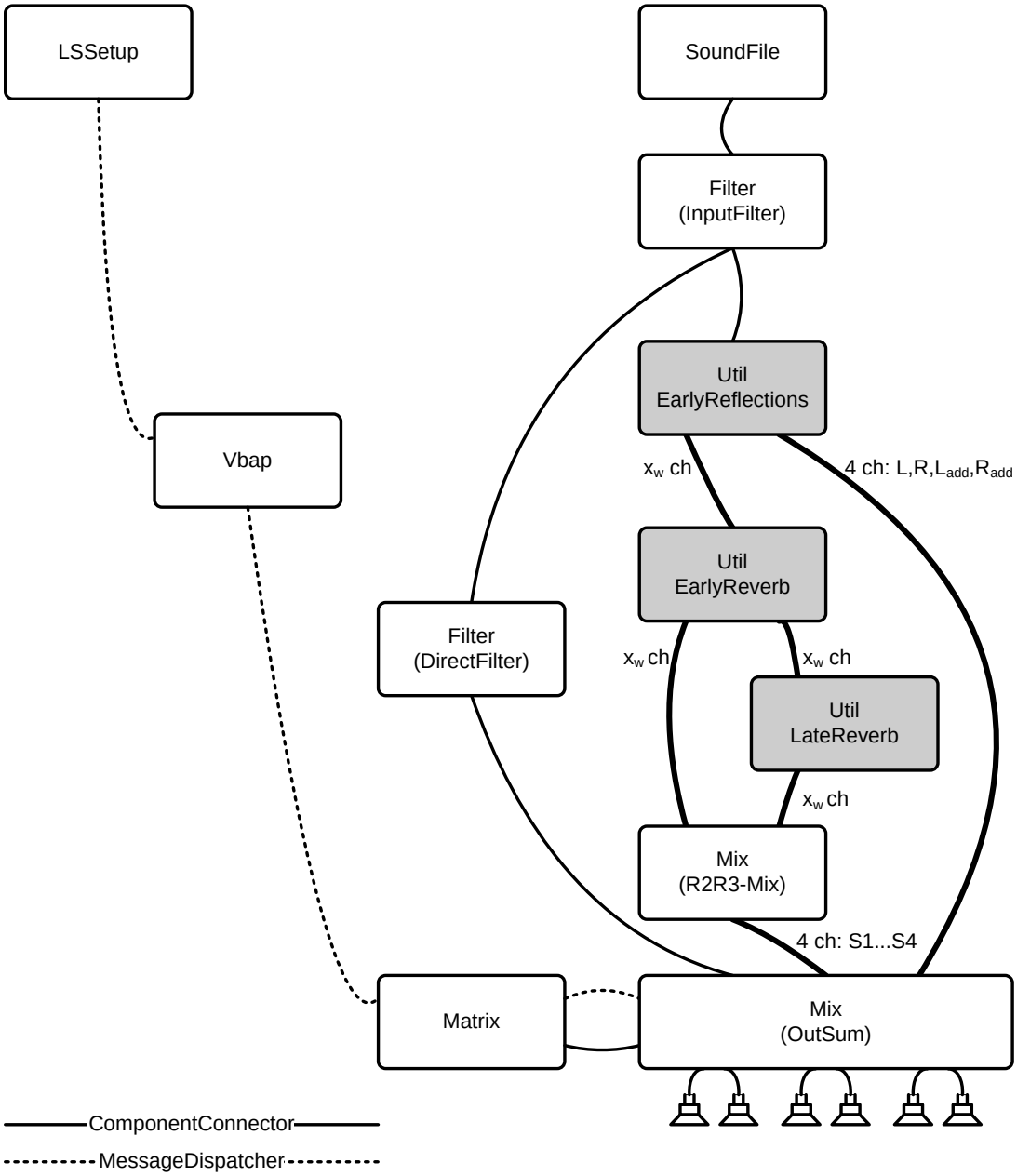


Fig. 5.14: The SFC of the *Tanga::PerceptualSource* Util. It contains the Utils *Tanga::EarlyReflections*, *Tanga::EarlyReverb* and *Tanga::LateReverb*, marked in gray. Bold ComponentConnectors indicate that more than one channel of audio is passed between Components, with x_w denominating the number of 'workchannels' currently defined in the TANGA. Good results have been achieved with a setting of $x_w = 8$.

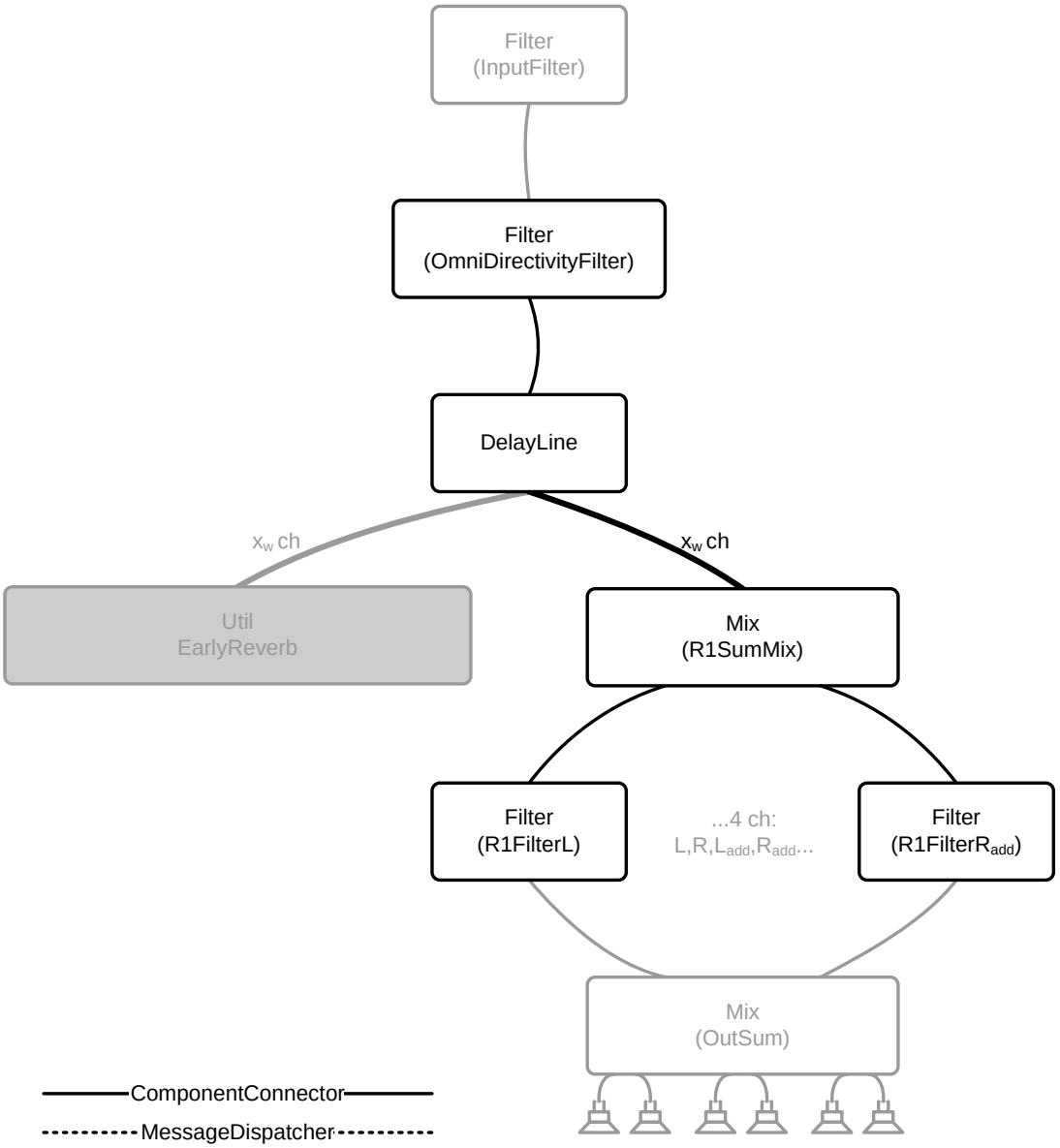


Fig. 5.15: The SFC of the *Tanga::EarlyReflections* Util. Components that the *Tanga::EarlyReflections* Util connects to are shown for reference and are grayed out.

them are contributing surfaces, that means surfaces on which sound sources are mirrored. All other objects are not taken into consideration for the room acoustic computation. The AcousticMaterial node is an extension of the Material node. It contains all fields of the original Material node (describing the visual characteristics of an object's surface), plus fields describing the acoustic transmission and reflection characteristics of the surface [14496-11].

Therefore, the complexity of the simulation process can be controlled by the scene author. To reduce the problem of exponential growth in the number of image sources, some considerations are discussed shortly.

- The reproduction setup consists of a circular loudspeaker array. As elevation information cannot be reproduced by such a setup, only image sources close to the horizontal plane need to be considered. Therefore, the process of computing image

sources should be reduced to two-dimensionality.

- The number of image sources to be computed can be controlled directly. For rendering on a fast computer a higher number of image sources might be desirable, whereas less powerful computers can only render up to a certain number of image sources before audible artifacts appear.

- The computationally most expensive part of the traditional ISM is the occlusion culling algorithm (also called "visibility check"). It needs to be performed whenever a room is of concave shape, i.e. not all points of every wall are visible from every location inside the room. For convex rooms the visibility check can be skipped because full visibility is always given. For rectangular rooms, a subset of convex rooms, a large number of higher order image sources coincide locally, thus decelerating the exponential growth of complexity.

- It is reasonable not to attach AcousticMaterial nodes to objects smaller than a certain size. This is also common practice in high-quality, off-line room simulation applications and does not represent a perceivable limitation of fidelity in itself. It is rather a question of threshold of object size - for performance reasons, in the audio-visual scenes used in the course of this work, none of the objects inside of a room are considered. Only the outer, delimiting walls of a room are given AcousticMaterial properties.

- The AcousticMaterial node fields for transmission and reflection properties can either be filled with sets of filter coefficients, or with so called frequency-gain pairs that define a gain factor between 0.0 (no reflection / transmission) and 1.0 (all energy is reflected / transmitted) for a given value of frequency. TANGA automatically transforms these frequency / gain pairs to filter coefficients for the use in a *Tanga::Filter* Component. A lower number of filter coefficients (and thus of frequency / gain pairs specified in the scene description) reduces the computational load of the respective *Tanga::Filter* Component.

Whereas the first two measures are part of the algorithmic realization of the simulation (and therefore valid independently from the scene description itself), the last three suggestions need to be considered during the authoring process of the scene. A method potentially suitable for dynamically reducing the room geometry's complexity (and thus the number of walls to be considered) as part of the audio rendering process itself is presented in section 5.10.

5.7.1 Physical Approach in the TANGA Audio Engine

The implementation of the MPEG-4 Audio *Physical Approach* follows the paradigm of separating the calculation of early reflections from the diffuse reverberation part. This can be seen in fig. 5.16. It shows the SFC of the *Physical Approach* implementation for an eight loudspeaker setup. The *Tanga::SoundFile* Component provides the sound source, in this example an omni-directional one. The direct sound path is panned via *Tanga::Vbap* in the *Tanga::Mix(EarlyMix)* Component.

The early reflections are generated as image sources in a set of *Tanga::Filter* and *Tanga::Delay* Components (ImageSource1 to ImageSourceX) with corresponding

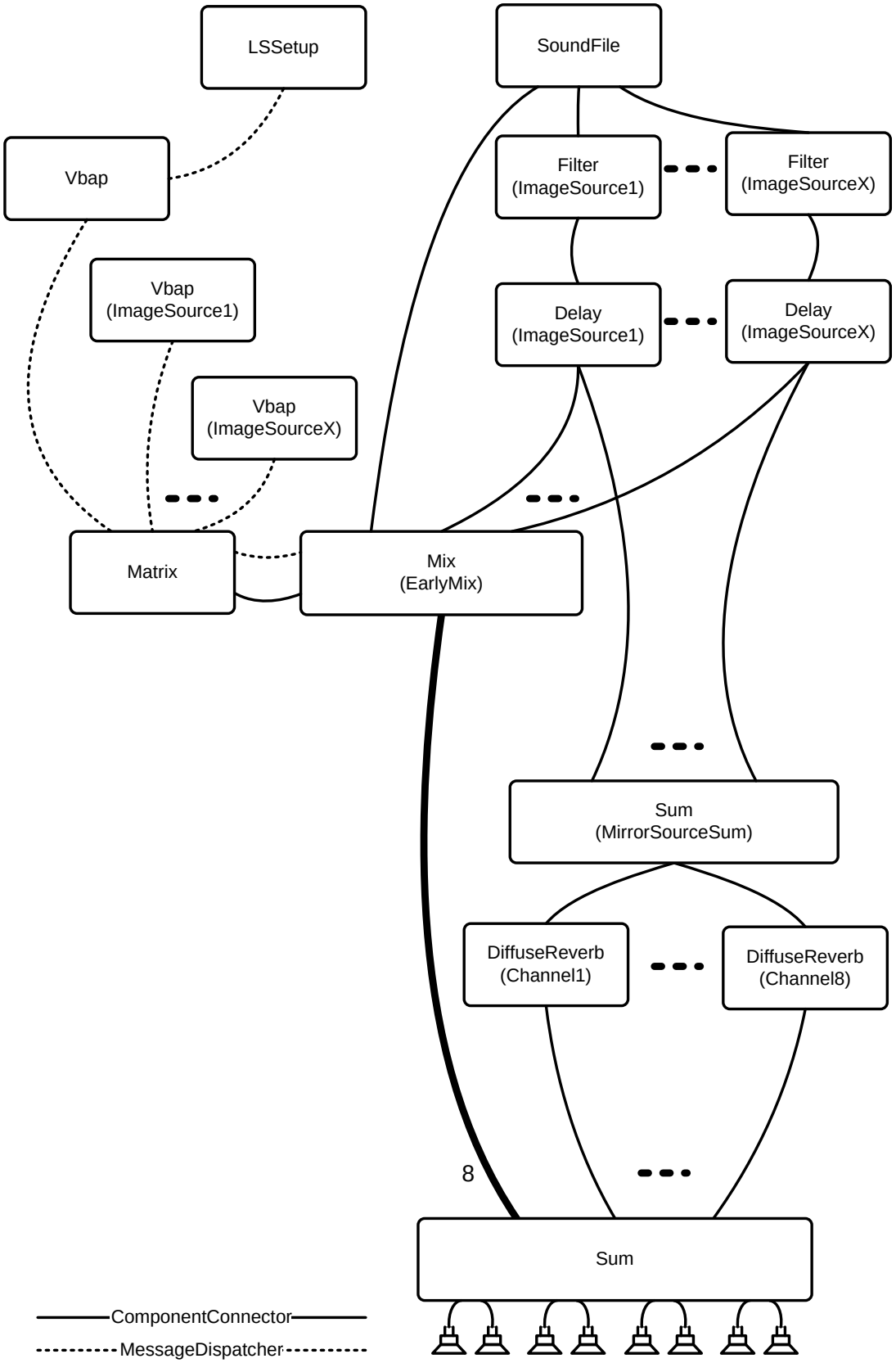


Fig. 5.16: The SFC of the MPEG-4 Audio *Physical Approach* implementation in the TANGA engine.

Tanga::Vbap Components for the directional panning of each image source. This is done by the *Tanga::SoundForest* Util. It provides the mirror() method with which an original sound source is mirrored at surfaces present in the scene description that have an AcousticMaterial node attached. From each original source a hierarchical tree structure is created that contains all image sources derived from the original source as children. Each source defined as following the *Physical Approach* in the scene description is the root of one such tree. Hence, when more than one such source is present in the scene description, a "Sound Forest" is created.

The *Tanga::Filter(ImageSource)* Components perform the filtering according to the reflection coefficients defined in the reflection properties fields of the corresponding AcousticMaterial node in the scene description. The *Tanga::Delay(ImageSource)* Components are responsible for time delaying the image sources according to the distance to the receiver. This calculation is dependent on the speed of sound defined in the scene.

Air absorption can be simulated by inserting a *Tanga::Filter* Component into the corresponding sound path in the SFC. Fig. 5.16 does not show the *Tanga::Filter* Components responsible for the air absorption calculation. Depending on the computational complexity allowed for the audio rendering, not only the direct sound, but also image sources could be filtered accordingly. The filter parameters can be controlled dynamically by the distance between locations of (image) sound source and virtual listener.

Late reverberation is created by the *Tanga::Reverberator* Util. The algorithm used for the creation of reverberated signals follows Gardner's approach of using nested all pass filters, see section 3.2.5. In order to achieve uncorrelated diffuse reverberation signals at the loudspeakers, one *Tanga::DiffuseReverb* Component is created per output channel by the *Tanga::Reverberator* Util: each of these Components is instantiated with a slightly different delay time between the nested all pass filters. This variation is created at random. Fig. 5.16 shows that the *Tanga::DiffuseReverb* Components are fed with the early reflections signals that are summed in the *Tanga::Sum(MirrorSourceSum)* Component. This adds to the reverberation density of the output signals.

The *Tanga::DiffuseReverb* Component differs from other Components in that it comprises a number of elements that according to the TANGA concept normally would have been split into separate Components (and then joined again by a TANGA Util). These elements are mainly the delay and filter classes, which are already available as Components in the TANGA. Yet, the code creating the diffuse reverberation according to Gardner's nested all pass filter network was already available at the IMT[13], and to simplify matters that code was integrated monolithically to form the basis of the *Tanga::DiffuseReverb* Component.

5.7.2 Reduction of Image Source Count

Computational complexity of the image source method grows with $O(n^r)$ (for n surfaces and r reflections). Apart from limiting the order of the image sources to be computed, the number of surfaces in the model should be kept as low as possible. The implementation performs a simple vicinity check before adding image sources to the Component Graph. As soon as image sources coincide or nearly coincide locally, the new source is discarded, whereas the already existing source is kept. This way, the number of image sources that do

[13]The code is written in C language and based on the MATLAB implementations of Gardner's Real Time Multichannel Room Simulator by Holzem and Beltrán, see [holz99, bel02].

not contribute salient peaks to the early reflections part of the RIR is reduced. Depending on the size of the room, a good vicinity distance has been found to be below $5m$ for very large halls and significantly less for smaller rooms. Because the algorithm starts creating sources of first order and then adds higher order sources order by order, sources with less reflections (statistically higher energy contributions) are preferred. Alternatively, the energies and arrival times of neighboring image sources could be averaged. The vicinity check is performed on every newly created image source.

At least some of the image sources may be located outside the horizontal plane, even if ceiling or floor do not have any AcousticMaterial properties attached. This is the case whenever sound is reflected on non-vertical (AcousticMaterial) object surfaces. When using a circular array of loudspeakers for the reproduction, sources which are quite close to the horizontal plane can be approximated, but elevated sources are fairly difficult to reproduce with this loudspeaker setup. Therefore, the amplitude of the image sources is scaled by the cosine of the elevation angle as suggested by Gardner [gar92], which results in ignoring sources far off the horizontal plane. Sources close to the horizontal plane remain unaffected.

5.7.3 Stop Criteria for Acoustic Simulation

The room acoustic simulation based on the ISM is growing exponentially in complexity with the length of the impulse response generated by the early reflection calculation. It is therefore important to have means of flexibly controlling the number of early reflections that are computed. For this, two different stop criteria have been implemented in the TANGA engine's simplified image source method.

The first stop criterion determines the order up to which the image sources are computed. It provides an easy to handle parameter that is set globally upon initialization of the I3D player. It is independent of the scene description. The total number of image sources actually computed depends on the geometry of the room: a stop criterion of second order will render 12 image sources in the two-dimensional case for the rectangular room, whereas in a room with eight convex walls a total of up to 64 image sources (also second order, 2D) are possible[14].

The second stop criterion specifies the absolute number of image sources to be computed, independently from their order. Sources are created starting with the first order, then second, then third and so on. As soon as the specified total number is reached, no more sources are created.

5.7.4 Assembly of Early Reflections and Diffuse Reverb

The *Tanga::Sum* Component is used to adjust the gain of the reverberated signals to match the level of the early reflections. There is no need for panning of the uncorrelated signals coming from the *Tanga::DiffuseReverb* Components, as each of them is created for reproduction on a separate loudspeaker.

The assembly of early reflections and diffuse reverberation is done using Gardner's approach described in section 3.2.6. The reverberation time T_{60} is taken directly from the scene description itself (*reverbTime* field of the AcousticScene node).

[14]The total number depends on the exact geometry of the room, because image sources of higher order may coincide locally. This is the case in the rectangular room.

5.8 Acoustic Obstruction

The ability to move freely in scenes rendered by the I3D can lead to situations where sound sources are hidden behind walls or other large objects. Fig. 5.17 shows a simplified

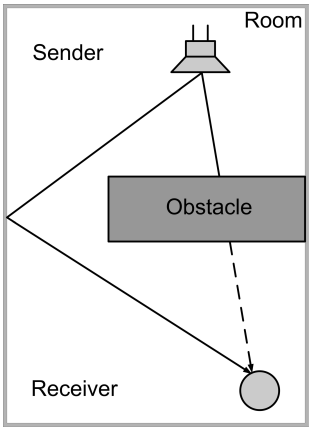


Fig. 5.17: Draft showing an exemplary situation for acoustic obstruction. The direct path is muffled due to diffraction and/or transmission and the reflected path remains unaffected.

situation in a room in which an obstacle is located between sound source and listener. The direct sound path from the obstructed source is interrupted. The direct sound can reach the listener only via diffraction around the wall or transmission through the obstacle. The reflected path of the sound source remains unaffected.

The Interactive Audio Special Interest Group (IASIG) gives a definition of acoustic obstruction and its frequency dependent effects on audio signals in [ias99]. Notably, the IASIG differentiates between *obstruction* and *occlusion* in an acoustic environment. If an acoustic environment is separated into two sub-environments by a partition in such a way that listener and sound source are located in different sub-environments, then transmitted, diffracted and reflected sounds are affected. This effect is called occlusion. When only the former direct sound path is affected, but (part of) the reflected sound paths remain unchanged, the effect is called obstruction.

In the I3D MPEG-4 player, the effect of obstruction of sound sources has been implemented. Various useful methods for the detection of obstruction along with their respective advantages and disadvantages have been discussed by the author in [rei03].

Despite the inaccuracies associated with Bounding Spheres[15] for non-spherical objects as shown in fig. 5.18, the intersection test with Bounding Spheres for detecting obstructions has been chosen for an implementation in the I3D. Intersection allows for a very fast detection of obstruction and is therefore most interesting for real-time interactive applications.

All potential obstacles in the BIFS-scene are polygonal *Shapes*, so called *IndexedFaceSets*. They need to have an *AcousticMaterial* node attached to be included in the computation of obstruction. *IndexedFaceSets* can be very complex polygonal structures consisting of many surfaces, each of which would have to be checked for obstruction. In the implementation of obstruction used in the I3D, for each acoustically relevant *IndexedFaceSet* a Bounding Sphere is computed. This replaces the complex object for the obstruction calculation. This

[15] A Bounding Sphere or BSphere is the smallest possible sphere that encloses an object completely.

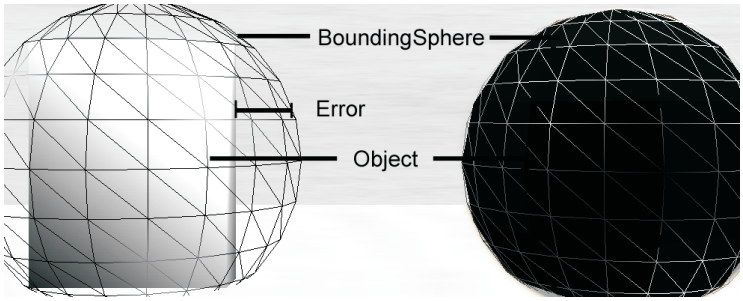


Fig. 5.18: Inaccuracy when using Bounding Spheres for non-spherical objects. The error introduced is small when the three dimensions of the obstructing object are relatively close to each other or the object is of convex shape (right). Problems arise when one dimension is significantly larger or smaller than the other two (left).

way, the detection of obstruction is a computationally light-weight problem easy to solve under real-time conditions.

Fig. 5.19 shows the intersection test based on [hai89]. Each time the position of listener or sound source is updated, a ray is emitted from the listener to the sound source. For each potential obstacle with an *AcousticMaterial* node attached, an intersection test with the ray is performed. Whenever the intersection test results positively, the sound source is muffled by adding a Filter Component to the Signal Flow Chart. The parameters of the Filter Component are determined from the *AcousticMaterial* node characteristics of the obstructing object. For a detailed description of the intersection test see [rei03] and [hai89].

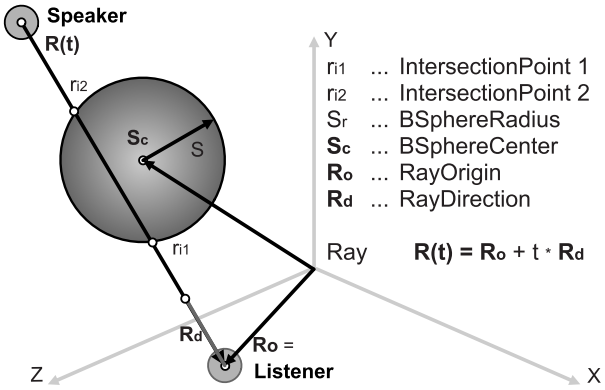


Fig. 5.19: Intersection test with a Bounding Sphere, *after [hai89]*.

The algorithm using Bounding Spheres for the detection of obstruction is fast but rather coarse. This means that the algorithm frequently detects obstruction, although the obstructing object does not cover the sound source at all. Typically this happens with objects of significant height but negligible width, like pillars, vases, etc. In fact, informal subjective assessments have shown that it is rather irritating and perceived as unnatural, when the onset of acoustical obstruction starts before the sound source is actually located (at least in part) behind the obstructing object. A general ad-hoc solution to this problem is to reduce the radius of the Bounding Sphere by a factor of between 0.5 and 0.8. Because the avatar size[16] is usually not changed in the course of an application session, potential

[16]In MPEG-4 the avatar size is defined as the distance that the virtual character's head exploring a 3D

obstruction errors occurring below and above the object are not noticed. Of course, this is only a workaround. For an effective solution, different primitives (e.g. Bounding Cylinders) should be examined for computational complexity when intersecting with a ray.

5.9 Distributed Sound Sources

MPEG-4 Part 11, more precisely AudioBIFS v2 (also known as AABIFS), provides means to define and reproduce sound from virtual sound sources that can have a directional source pattern [14496-11]. Such audio sources are always perceived as point sources. Their location is defined as one point in 3D space. Whereas this is sufficient for most audio only applications, point sound sources in audiovisual scenes have been identified by the author to contrast greatly with our every day experience in real world situations. In these, the exact location or direction of incidence of e.g. a background sound often remains unclear, especially when approaching such a sound source. This is true for sound sources inside of larger vibrating objects (e.g. an engine in a car, household appliances like e.g. refrigerators, etc.), for wide opened windows to a noisy outside environment, and for musical sound bodies like choirs or groups of instruments of an orchestra, to name a few[17]. In interactive audiovisual applications it is therefore desirable to also have distributed sound sources available that can, not least because of easier handling, be controlled as one single sound source.

In MPEG-4 AudioBIFS, distributed sound sources are part of the AudioBIFS amendment adopted at the MPEG meeting in Redmond, USA, as a final draft amendment (FDAM) in July 2004 [N6591]. It is part of the international standard ISO/IEC 14496-11 since 2005. The corresponding conformance was adopted as FDAM at the MPEG meeting in Lausanne, Switzerland in July 2007 and will become an international standard later in 2007. Distributed sound sources are represented by the *WideSound* node. Based on the standard, a WideSound implementation for the I3D has been devised by Florian Voswinkel in a diploma thesis supervised by the author [vos06]. In the I3D's implementation, a distributed sound source is defined as a box of dimensions $width \times length \times height$ with a certain position and sense of direction in the virtual scene. Depending on the density parameter of the distributed source, it is represented by a number of uncorrelated point sound sources, see fig. 5.20. Their exact positions inside of the box are computed using a reverse median-cut algorithm [hec82].

In the I3D, distributed, box shaped sound sources can be used in virtual 3D scenes. They are automatically generated from monophonic sound files. To create highly decorrelated sound sources from one single monophonic sound source the *Tanga::Decorrelator* Component has been implemented. It is based on a nested all pass filter network.

The *Tanga::Decorrelator* Component performs the decorrelation processing of a monophonic sound source. It calls the helper class DecorProcessor and passes the number of decorrelated signals to be created, as well as a scale factor. DecorProcessor contains an implementation of the small room reverberator (a nested all pass filter network) by Gardner. The scale factor passed from the *Tanga::Decorrelator* Component determines the degree of correlation between the resulting signals by influencing the delay times within the all pass filter networks. As these may not be varied by more than $\pm 15\%$ without significantly

scene is located above the floor.

[17] In fact, in the natural environment these sources are usually composed of independent, discrete sound sources!

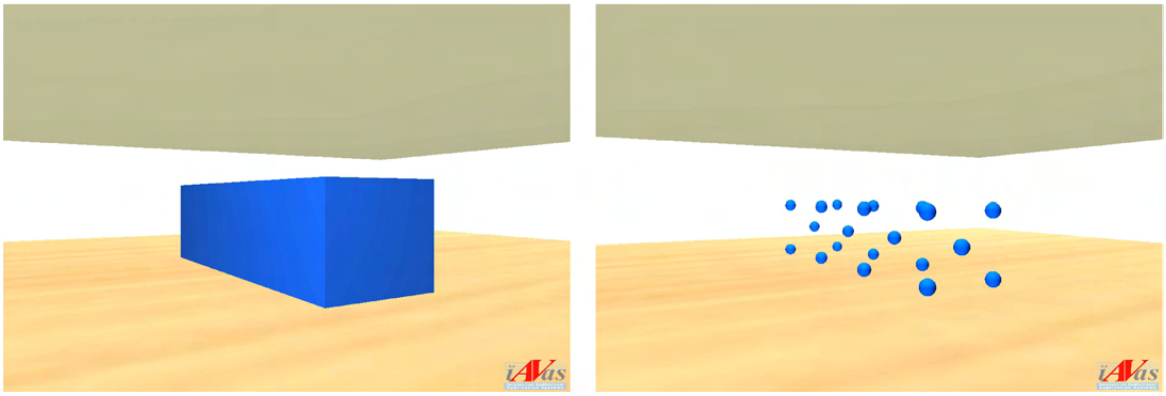


Fig. 5.20: A distributed sound source (left) and its implementation using uncorrelated point sound sources (right). *From [vos06].*

altering the tonal quality of the resulting audio signal[18], only a maximum of around ten decorrelated sources can be created with this method. Additional sources created within this range are not always sufficiently decorrelated from each other any more.

Another drawback of this method is that the filter network adds a slight reverberation to the dry monophonic sound used as input. Yet, informal listening tests have shown that this is acceptable if room acoustic rendering (be it *Physical* or *Perceptual Approach* based) is performed: the room's reverberation will mostly cover the reverberation added by the decorrelator.

More critical are the comb filter effects added by the very small delay differences in the nested all pass filter networks. Some sounds are commonly identified as being processed by a slight flanger or phaser effect when processed trough the WideSound node.

Both Kendall [ken95] and Potard [pot06] suggest a number of alternative methods for effectively decorrelating monophonic audio signals. The simplest one is to convolve an input signal with k different all pass FIR filters having random phase responses, resulting in k decorrelated output signals. Unfortunately, this method is reported by Potard to generally suffer from a tradeoff between temporal smearing effects which occur for filters longer than $10ms$, and insufficient decorrelation strength with too few filter taps. He suggests to use FIR decorrelation filters with 256 taps as "a good compromise" [pot06]. Potard also describes the use of IIR all pass filters for decorrelation purposes, which are computationally more efficient than the FIR variant. Common to both FIR and IIR decorrelation techniques is that the filter coefficients are fixed and therefore can only produce a relatively small number of decorrelated signals. Potard reports under 10 decorrelated signals, which is similar to what the method used in the TANGA engine described above can achieve, although at a lower computational cost.

Alternatively to fixed filter techniques, dynamic decorrelation techniques could be applied to create larger numbers of decorrelated sound sources. These techniques can consist either in a frequency varying decorrelation by applying sub-band decorrelators (which themselves consist of fixed FIR or IIR decorrelators as described above), or in a time varying decorrelation for which the decorrelation strength is modulated over time. For the IIR filter case this can be achieved by randomly varying the distances of the poles and

[18] Gardner recommends varying the delay times from ±2% to ±5% between (uncorrelated) loudspeaker channels for reverberation purposes [gar92]. A range of up to ±15% was found satisfying for decorrelation purposes in informal listening tests in the listening lab of the IMT in June 2006.

zeros from the unit circle and continuously updating the coefficients with the resulting new values. Kendall reports that "dynamic decorrelation imparts a quality of liveness to a sound field that is missing in the [fixed, author's note] FIR implementation" [ken95]. On the other hand, Potard notes that "dynamic decorrelation can lead to listening fatigue due to the constant changes of the decorrelation filter phase responses" [pot06].

Given these aspects, the decorrelation method applied in the TANGA engine should be revised in the future in order to produce more decorrelated sources, while at the same time making it computationally more efficient.

5.10 Dynamic Audio Scene Graph Simplification

As already seen, the complexity of the room acoustic simulation following the MPEG-4 Audio *Physical Approach* using the ISM can be controlled at different stages: at the authoring stage by reducing the number of AcousticMaterial nodes in the scene description, and at the rendering stage by determining the vicinity check distance and the stop criteria (maximum order and total number of image sources). All these methods require an a priori setting of parameters that cannot be changed during runtime. Whereas they render credible results for listener locations near the center of the virtual room, approaching a reflecting wall can make the simplifications become audible because the early reflections pattern does, under certain circumstances, not change the way it is expected to.

To overcome these restrictions, a dynamic simplification of the underlying Component Graph could be one possible solution. This way, a detailed room acoustic description using image sources could be achieved near the listener's location, whereas for objects farther away from the listener the description could be simplified. Such an approach was first suggested by Clark [cla76]. Fig. 5.21 shows the basic idea of this method: the level of detail (LOD) of the mesh describing the bunny object is increased when the object is approached. Transformed to the acoustical equivalent, this means that the acoustically relevant geometry of the room (and therefore the positions and number of image sources) depend on the listener's location / distance to the respective wall.

Fig. 5.21: Graphical example for the concept of dynamic level of detail (LOD). The LOD is increased for objects closer to the viewer, such that more details are visible (or audible in the equivalent acoustical case) when approaching the object (from left to right). *The Stanford Bunny model from the Stanford 3D Scanning Repository [w-sta].*

It is desirable to have a scene simplification strategy that provides more details for the parts of the scene that are closer to the listener and less details for the corners which are farther away. We generally perceive more detailed acoustic information from nearby objects, whereas most properties are more difficult to determine for sound sources which are farther away in the real world. This is even more evident if we look at early reflections represented by image sound sources of first or higher order, which have been reflected and thus inevitably diffused at most real walls. Therefore, a dynamically adapted level

of detail should be provided for the acoustic simulation of the scene. Objects which are located closer to the listener are rendered more precisely and the resulting image sources are rendered more accurately, too.

The scene simplification algorithm should therefore satisfy the following conditions:

1. To maintain the overall reverberation time T_{60}, the deviation from the original room volume should be as small as possible. This is relevant for a good agreement between visual and auditory spatial cues.

2. It should be possible to dynamically adapt the simplification process according to the listener position in such a way that for objects located closer to the listener, more details are retained.

3. The calculation should be computationally efficient enough to ensure real-time computability.

A well known candidate for the scene simplification task is geometry compression [dee95]. Geometry compression has been developed for computer graphics applications that deal with complex geometry models. The requirements set out for geometry compression were mainly to have a good compression for reducing storage space and enabling efficient transmission for distributed applications like collaborated modeling and multi-user games.

Whereas geometry compression can indeed be used for scene simplification, the emphasis of such algorithms is on good visual quality. Additionally, the computation is quite expensive, so even though this approach can be adapted to the needs of scene simplification, it does not entirely fulfill the conditions.

A better approach to scene simplification seems to be the use of quantization algorithms. There are many simple quantization techniques that could be used for simplification, but most of them do not take into account the special properties of objects in a virtual 3D environment. Here, an algorithm that "knows" about the proximity of objects in the 3D-space is needed, such that small features modeled by vertices situated closely together are simplified the most. The so-called octree quantization has been found to fit these demands best.

5.10.1 Octree Quantization

The concept of octree quantization was originally developed for color quantization [ger90] and has been adapted by Schmalstieg [sch96] for level of detail generation for VRML to an algorithm called LODESTAR.

The algorithm in its original form cannot be used for the problem here, since it has been developed for a static level of detail generation: From an original scene object a couple of simpler representations that have an ever smaller level of detail are computed. Depending on the distance of the viewer from the object, a different representation is chosen. This corresponds to the usual level of detail concept in VRML.

Here, an algorithm is needed that does not simplify a scene object as a whole. Rather, if there is for example one indexed face set used for modeling the walls of a room with a bay, different parts of the indexed face set need to be simplified to a different degree, depending on the position of the listener inside that room.

The construction of the octree is performed in a similar way to what is described by Schmalstieg in [sch96]. Additionally, in each branch node of the octree the midpoint of

up to eight underlying nodes is stored. This octree is then used for simplification in the following way: The octree is traversed from top to bottom. At each node the distance between the listener position and the point stored at that node is computed. If this distance is smaller than a given threshold, the algorithm proceeds with all the child nodes. Otherwise the stored midpoint is used as the representative for all the vertices stored in the subtree rooted at that node. The vertices are then replaced by the chosen midpoint which leads to the partly simplified representation of the scene's geometry. Of course, it has to be guaranteed that the resulting geometry is not corrupted during this simplification process.

The computation of the octree and the subsequent adaptive simplification of the scene objects according to the listener position can be done very efficiently, as long as the objects which have to be simplified are static. As soon as scene objects from which the octrees should be built are animated, the issue becomes more critical and further investigations have to be made. Here it is assumed that the listener is located in a static room and, if there are any animated objects in the scene, these are small enough not to be relevant for the audio rendering.

Fig. 5.22 shows a screenshot of a virtual jazz club scene. This is how the jazz club model

Fig. 5.22: Screenshot of a virtual jazz club scene. Located in the back of the room is a stage that forms an alcove in the wall.

is represented visually to the user. Fig. 5.23 shows the acoustically relevant objects of that scene from roughly the same perspective. This is the geometry that is used as a basis for the image source computation. The octree quantization has already been applied.

Fig. 5.24 shows the same part of the scene, but from a location farther away from the stage. As can be seen, the number of reflecting planes has been reduced significantly by applying the octree algorithm to the geometry of the alcove.

The octree quantization itself satisfies the three conditions established in section 5.10. The method was presented at the AES 116th Convention in Berlin, Germany in 2004 [dan04], and a demo application visualizing the results of the dynamic simplification was shown.

5.10.2 Audio Rendering of Simplified Scene Graphs

When moving through the virtual scene, the shape of the acoustically relevant objects changes rather often when the octree quantization is applied dynamically. Therefore, the simplified scene graph needs to be traversed repeatedly. After each traversal, the positions and also the total number of image sources need to be updated.

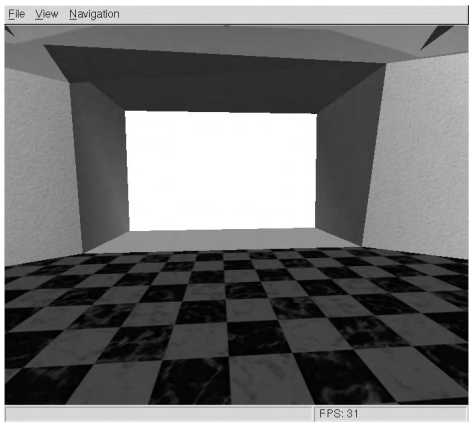


Fig. 5.23: The octree simplified jazz club scene from roughly the same perspective as in fig. 5.22. Note that the simplified scene is normally not visually rendered and is displayed here only to show the result of the algorithm.

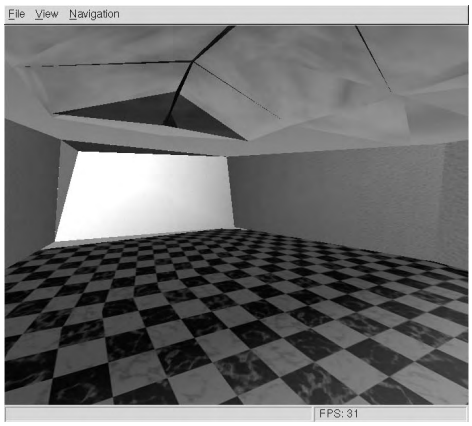


Fig. 5.24: The octree simplified jazz club scene from a location farther away from the alcove. Because the virtual distance to the alcove has been increased, the LOD of its geometry has been reduced by the algorithm.

Until now, there is no mechanism implemented in the TANGA that allows to update only part of the image sources organized in the Sound Forest. Therefore, the whole Sound Forest needs to be destroyed and built again, which is computationally very expensive. The example scene shown in fig. 5.22 with one sound source cannot be rendered audible on today's PCs without introducing audible artifacts. Therefore, the method has not been integrated into the I3D player yet.

For a future implementation, it might be possible to combine the octree algorithm with a Binary Space Partitioning (BSP) approach to limit the area in which the image sources are recomputed. Further investigations need to be performed before an integration is feasible.

5.11 Multi Core / Multi Thread Processing

Computing power available on consumer PCs has been continuously increasing as predicted by Moore's Law for the last decades. Whereas in the past this increase has expressed itself in ever higher CPU clock speeds, the last year has clearly shown that future increases in computing power will only be possible by integrating a higher number of processor cores into one CPU.

In the area of real-time audio rendering, this insight needs to be translated into a systematic approach for distributing and parallelizing the steps of a signal processing chain in order to benefit from future increases in computing power. Whereas this process is rather simple for traditional real-time algorithms and complexities, it is all but trivial for scene graph / object based applications.

Because the Component Graph contains all Components (vertices, nodes) and connections between Components (edges) that are relevant for the audio signal processing, the graph theory known from mathematics and computer science can be applied to it. This approach has the advantage of providing a generalized, universally valid solution for the problem of optimal performance of the graph. The solution is therefore independent from the processor's architecture or the implementation details of the audio signal processing Components.

Unlike in other implementations, the approach implemented in the TANGA can handle an arbitrary number of cores or CPUs for optimal distribution of Components to threads. It is therefore very well primed for future developments of multi core CPUs. In analogy to the TANGA Component Graph, the audio signal processing tree (SFC) is conceived as a graph upon which the optimization is performed. The result is an optimum distribution of computing processes to available cores or threads.

The task graph theory[19] primarily includes methods and approaches for the parallel execution of graphs. However, these procedures can not be simply transformed to the field of audio processing with several threads, because the graphs in the task graph theory are often subject to certain restrictions. One of these is commonly that no recursive structures are allowed. In audio signal processing, feeding back signals from output to input is a regularly used method to create IIR filters, echoes, reverberation and so on. Therefore, specialized solutions for the parallelization of audio processing need to be found. In the following, two such approaches will be described and compared to each other: the *Dynamic Component Parallel Rendering* (DCPR) and the *Dynamic Component Cluster Rendering* (DCCR). Both have been developed in the course of a diploma thesis by Andreas Partzsch performed under the supervision of the author [par07].

5.11.1 Dynamic Component Parallel Rendering

In the Dynamic Component Parallel Rendering (DCPR) the execution of a graph is realized along a special topological sorting (TS) of the Components. The topological sorting is a series of Components of a graph which takes into account the order in which the Components are executed. A graph usually has several possible sequences of execution which all comply to the criterion of topological sorting. For example, the topological sorting of Component Graph 3 in fig. 5.25 could be:

$$TS_1 = S_1, D_1, ..., D_6, F_1, ...F_6, M_1$$

or

$$TS_2 = S_1, D_1, F_1, D_2, F_2, ..., M_1$$

or...

[19] The task graph theory deals with the parallelization of tasks. For this it uses an abstract model - the so-called task graph - of a set of dependent tasks, which can be executed concurrently.

Yet, there are only few topological sortings *well suited* for the parallelization of a Component Graph. A good topological parallel sorting is generated with a modified *breadth first search* within a reverse edge Component Graph[20] starting from the root Component. In this graph traversal process, each Component may be visited several times. A Component will be pushed to the beginning of a list during its inspection in the traversal process. If the Component already exists in the list, the existing entry is deleted before adding the Component again. For Component Graph 3 in fig. 5.25, the method described delivers TS_1. It is obvious that $D_1, ..., D_6$ as well as $F_1, ..., F_6$ can be executed in parallel, because their input does not depend on each other.

For multi-threading purposes, the generated topological sorting might be written to a global list. Because the computation of digital audio signals usually is performed block-wise, every signal processing Component has an output buffer which is filled in each render pass. For every render pass, each so-called worker thread[21] takes a Component from the list and performs it. This is done until the list is empty. For a new render pass the list is refilled and the worker threads are started again. To avoid conflicts in the render sequence, the execution of a Component can only be started if the output buffer of the preceding Component(s) have been correctly filled. The worker thread itself verifies that all parent Components have been completely performed before starting. If rendering has not completed, the thread will wait until all parent Components have been performed.

The advantages of the Dynamic Component Parallel Rendering are its simplicity and the dynamic adjustment to the graph's structure and to different numbers of threads. Unfortunately, a major disadvantage of this approach is its fine granularity, which generates a lot of synchronization overhead. This is roughly the same for all Components, but for computationally less intensive Components the relative weight of the synchronization overhead is much higher compared to more complex signal processing Components.

A large Component Graph consisting of many simple signal processing Components is therefore bound to be computed rather slowly using the DCPR method.

5.11.2 Dynamic Component Cluster Rendering

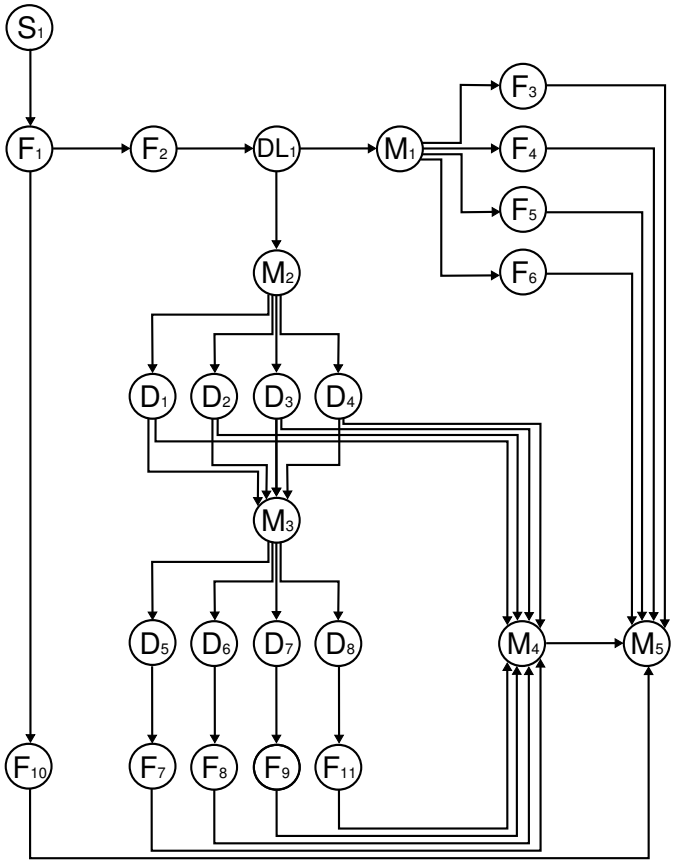
The Dynamic Component Cluster Rendering (DCCR) was developed to reduce the synchronization overhead compared to the Dynamic Component Parallel Rendering. The basic idea is to have worker threads executing whole sections of the Component Graph and not individual Components. Therefore a graph has to be divided into sub-graphs. For this, information about the computational complexity of a Component (referred to as *Processing Time Index* - PTI, see subsection 5.11.4) and the number of threads executing the graph is necessary. With this knowledge, an equal distribution of the workload between threads can be ensured, thus significantly reducing synchronization overhead.

Clustering of a graph is a so-called NP[22]-complete decision problem. This means that no deterministic algorithm is known which yields the exact right solution for the optimal clustering of a graph in polynomial time. To find a solution for this kind of problem, heuristics (also known as *rules of thumb*) must be used. These heuristics render satisfactory results in a very large number of cases. Yet, the solutions are not always optimal for all different input values possible (e.g. number of Components).

[20] In reverse edge Component Graphs the direction of all edges is reversed.

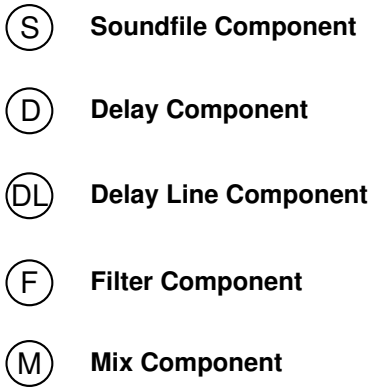
[21] The threads reserved for actually performing the computations defined in a signal processing Component are called 'worker threads'.

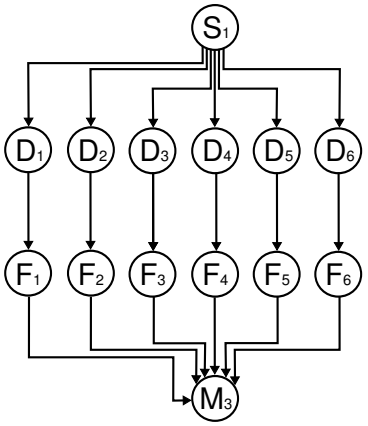
[22] Non-deterministic Polynomial time



Component Graph 1

Component Graph 2





Component Graph 3

Fig. 5.25: Examples for different structures of Component Graphs. Component Graph 3 can easily be parallelized, whereas Component Graph 2 has a strictly serial structure. Component Graph 1 is an example for a structure with serial and parallel elements.

A heuristic clustering method has been developed, the so-called *Color Clustering*. Color Clustering is an iterative clustering method with each iteration consisting of three steps:

1. Structural segmentation of a graph or sub-graph.
2. Optimization by clustering sub-graphs.
3. Estimation of the clustering efficiency based on the PTI. From this, the *Graph Processing Time Index* (GPTI) is derived.

In each iteration, segmentation of the graph or the largest sub-graph is effected until sub-(sub-)graphs can no longer be divided. This is the case when a sub-(sub-)graph contains only one Component or a sequence of Components. Component Graph 2 in fig. 5.25 is an example for a serial graph which can not be decomposed further. Then, an optimization by clustering Components to sub-graphs is performed. After that, the efficiency of the iteration is computed. Finally, after the last iteration, the clustering with the minimum GPTI is selected to perform the graph.

5.11.3 Color Clustering

The Color Clustering starts with the graph as a whole. The graph's efficiency without multi-threading is determined by computing the GPTI, which in this case is the sum of all PTIs. Then the first iteration is started with the structural segmentation. It is based upon sending colors into the graph.

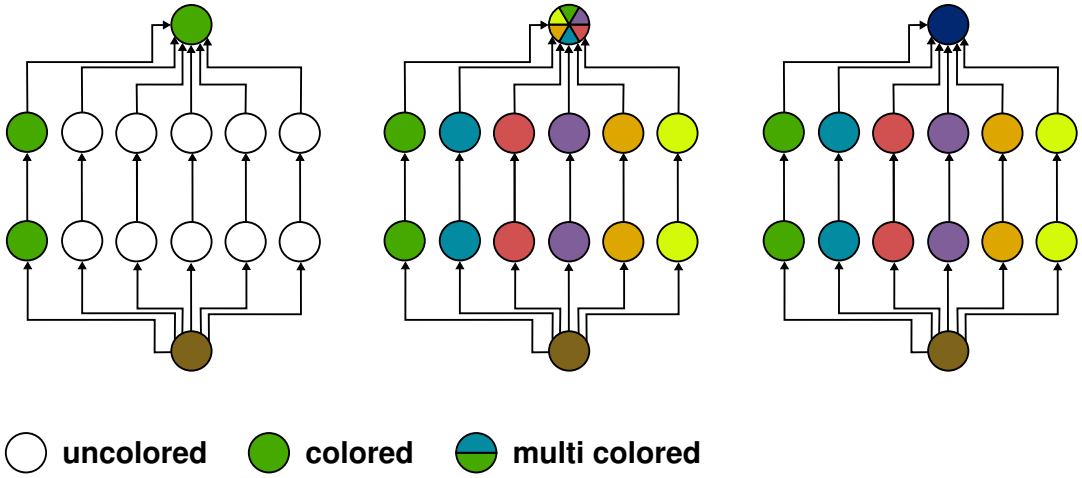


Fig. 5.26: The Color Clustering method.

Fig. 5.26 shows the process of injecting colors into a graph: starting at the root node[23], every single outgoing edge is traversed using a Breadth-First Search (BFS). For every single outgoing edge leaving the root node, a different color is used to mark the uncolored nodes that are traversed. Whenever the traversal process reaches a node that is already colored, the current color is added to that node, such that a multi-colored node emerges. After all edges of the root node have been used to inject different colors into the graph, the graph is completely colored - there are no more uncolored nodes.

[23]Note that the edges between Components / nodes are directed against the audio signal flow of the Component Graph.

Now, every multi-colored node is checked for other multi-colored nodes following. If there are no such multi-colored nodes following (i.e. all following nodes are single-colored), then all of these and the multi-colored node itself are re-colored using a new color. If there are no nodes following the multi-colored node, as is the case in fig. 5.26, then only the multi-colored node is re-colored.

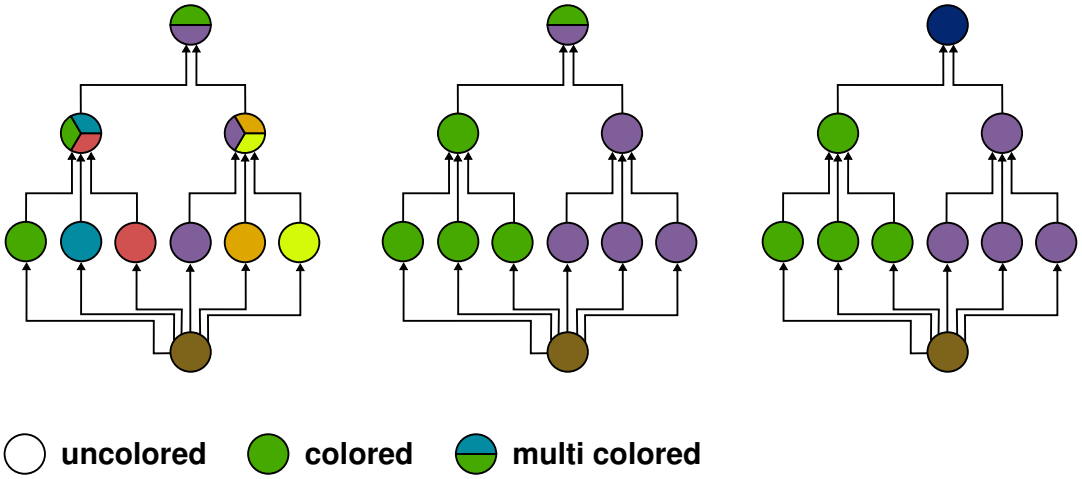


Fig. 5.27: An example for multi-colored nodes and the clustering process resulting in sub-graphs of equal color.

In case there are multi-colored nodes following other multi-colored nodes, see fig. 5.27, then the optimization by clustering sub-graphs is started (step two of the Color Clustering method). For this, the first multi-colored node as well as all preceding single-colored nodes are re-colored to a single color.

This process is repeated until all multi-colored nodes have been transformed into single-colored nodes. All nodes bearing the same color form one sub-graph. Because the Color Clustering method can result in a high granularity for certain graph structures, a number of specific adjustments are performed on the graph in order to keep sub-graphs above a minimum size.

Whenever the number of parallel sub-graphs is higher than the number of threads, the sub-graphs are clustered in such a way that the PTIs are distributed evenly between clusters, see fig. 5.28.

At the end of each iteration the clustering's efficiency is estimated by computing the GPTI. The GPTI includes the overhead arising from the parallelization. Then the next iteration follows. Fig. 5.28 shows the first segmentation and the first optimization of the Component Graph 3 in fig. 5.25 for the case of two threads.

The concept of DCCR is very different from the DCPR approach. In DCCR, a list is generated in which each element itself is a list of sub-graphs which can be executed simultaneously. A master thread is allocated to this list. The master thread assigns sub-graphs to the worker threads and starts them only if the list element contains more than one sub-graph. Whenever the list consists only of one element with a sub-graph, the master thread itself executes this sub-graph.

The Dynamic Component Cluster Rendering pledges better performance than the Dynamic Component Parallel Rendering, because it generates less synchronization overhead. However, the DCCR is more complicated than the DCPR, and an auxiliary variable - the Processing Time Index - is required.

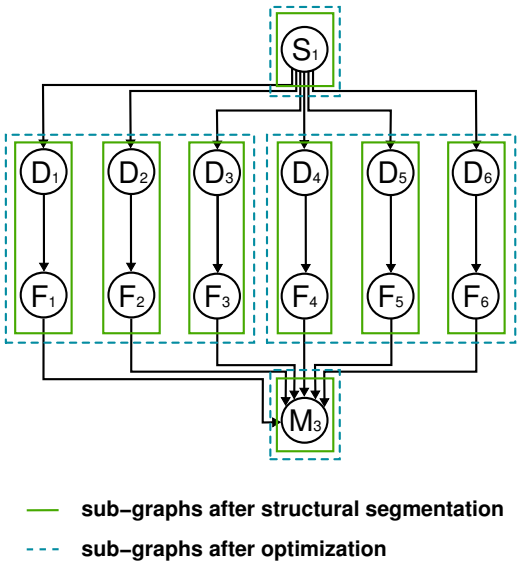


Fig. 5.28: Example for graph segmentation and optimization for two threads.

5.11.4 PTI - Processing Time Index

The Processing Time Index (PTI) specifies the computational complexity of a component. PTI is defined as a simple relation according to the following equation 5.2:

$$PTI = \frac{T_c}{T_{rc}} \tag{5.2}$$

T_c = execution time of the component
T_{rc} = execution time of the reference component

The idea behind the PTI is to provide a processor independent measure of the estimated computing power needed for the execution of a signal processing component. Because the PTI is defined as a ratio between component and reference component, it is assumed to be nearly the same across a large variety of different CISC (Complex Instruction Set Computer) computers / processors. Therefore the method introduced here is universally valid and does not depend on speed of the processor. The reference component that was chosen is a delay line with one input and one output. The Processing Time Index is a very important value for the estimation of the processing time for clustered sub-graphs (and therefore the quality of the clustering). The more precisely the PTI is determined, the more precise the quality analysis of the clustering is.

For measuring the PTI, tests have shown that performing ten repetitions of the measurement give reliable averaged results. Because the execution time - depending on the current state of the operating system and other applications running in the background - always slightly fluctuates, the values obtained should be averaged for a reliable value. PTIs on three different test systems have been compared:

1. AMD Athlon X2 4400+ @ 2.21GHz, 2GB RAM
2. Intel Pentium 4 HT @ 3.0GHz, 1GB RAM
3. Intel CoreDuo @ 1.86GHz, 2GB RAM

The PTI is different for every type of component and in many cases also depends on the parameters of the component (e.g. number of inputs or outputs). For example, a mix component with ten inputs and two outputs is more computationally complex than a mix component with two inputs and two outputs. Thus, the PTI has to be measured using the actual configuration of the component, as it is used in the component graph.

It appears that depending on the processor type, the PTI cannot be extrapolated for different configurations of a component. Fig. 5.29 shows the PTIs for an IIR filter component with different numbers of filter coefficients, measured on three different processors. As can be seen, the increase of the PTI over the number of filter coefficients is not constant for all of the tested systems (Intel Core Duo processor). Also, the increase in the PTI over complexity of the component is different on the three tested system.

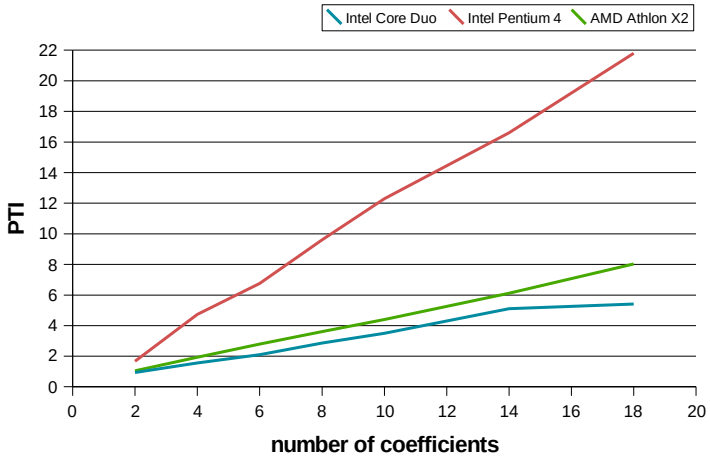


Fig. 5.29: PTI measurement for an IIR filter component.

For the Dynamic Component Cluster Rendering as applied in the TANGA engine, a PTI averaged over the three test systems for each Component is used. Potentially this method is slightly inaccurate but yet applicable.

To improve the accuracy of the PTI, it could be measured individually for all Components during the initialization phase of the audio engine, thus producing PTIs that correspond exactly to the type of processor actually used. Unfortunately, measuring the PTI of a Component is rather time consuming and needs to be done for different configurations of a Component. This needs some time and would delay the initialization of the audio rendering engine. In addition it could happen that another process accesses the CPU during the measurement, distorting the measured PTI. Yet, improving accuracy of the PTI is merely a question of weighing precision against effort - a question that must be answered individually.

5.11.5 Performance of the Methods

The two different approaches for parallelizing the audio processing described here (Dynamic Component Parallel Rendering and Dynamic Component Cluster Rendering) were compared on a test system (Intel Core Duo) using a number of different typical graphs. For this, the engine had to execute off-line audio rendering of a sound file with a duration of $60s$. Execution time was measured with audio rendering running in one or two threads. The test graphs consisted of the following five test cases:

Case 1: The Component Graph 1 in fig. 5.25. The Filter Components comprise three pairs of (IIR) filter coefficients each.

Case 2: The graph consists of two graphs of the case 1 type. Mix Components M_5 (see fig. 5.25) of both graphs are connected to an additional Mix Component. The Filter Components comprise three pairs of (IIR) filter coefficients each.

Case 3: The Component Graph 3 in fig. 5.25. The Filter Components comprise ten pairs of (IIR) filter coefficients each.

Case 4: The Component Graph 3 in fig. 5.25, but with 20 Filter Delay pairs instead of 6. The Filter Components comprise ten pairs of (IIR) filter coefficients each.

Case 5: The Component Graph 2 in fig. 5.25. The Filter Components comprise one pair of (IIR) filter coefficients each.

The performance values obtained verified the assumptions with respect to the advantages and disadvantages of the two methods. Generally, the Dynamic Component Cluster Rendering is faster than the Dynamic Component Parallel Rendering, regardless of the number of threads applied (see fig. 5.30). Primarily for graphs with many components, DCCR is significantly faster than DCPR, as can be seen in fig. 5.30 case 2 with two threads.

Test cases 3 and 4 contain Filter Components of relatively high computational complexity (ten pairs of IIR filter coefficients each). Therefore the difference in performance between DCPR and DCCR is not very large - synchronization overhead is relatively small with respect to the computational complexity of the Filter Component itself.

For test cases 1 and 2, computational complexity of the Filter Components is rather low (only three pairs of IIR filter coefficients each). This results in a significant performance difference between DCPR and DCCR, especially for test case 2 which can easily be parallelized. For test case 1 the performance gain of DCCR against DCPR is not so obvious, mainly because of the structure of the Component Graph 1 in fig. 5.25, which does not lend itself extremely well for parallelization.

By varying the number of filter coefficients used in the Filter Component, the influence of the synchronization overhead upon the overall performance can be observed very well.

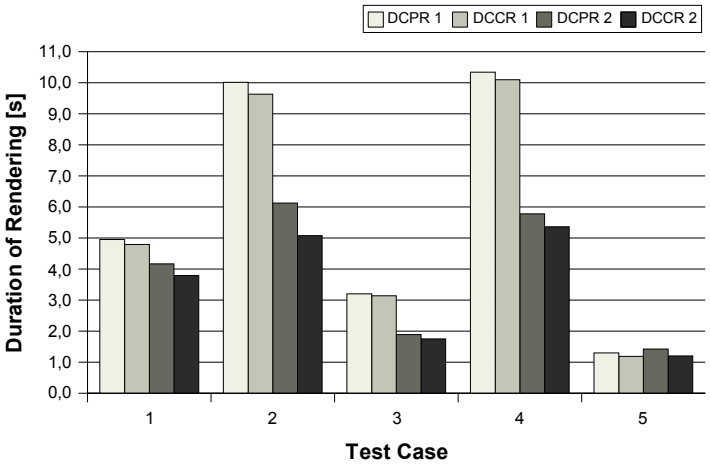


Fig. 5.30: Rendering duration of DCPR and DCCR for test cases 1-5, running in one (DCPR 1, DCCR 1) or two (DCPR 2, DCCR 2) threads on an Intel Core Duo processor.

The *speedup* of an optimization plays an important role in performance considerations. For optimizations regarding multi core processing, the speedup is defined in equ. 5.3:

$$S_n = \frac{T_1}{T_n} \tag{5.3}$$

S_n = speedup with n processors
T_1 = sequential execution time
T_n = parallel execution time with n processors

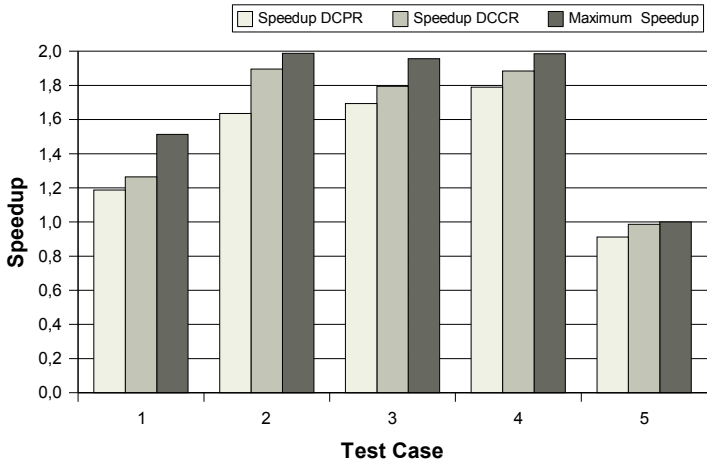


Fig. 5.31: Speedup S_2 of DCPR and DCCR for test cases 1-5.

The generalized maximum speedup for complex Component Graphs is difficult to determine, as the well introduced speedup formulas according to Amdahl or Gustafson predict the speedup only based on the amount of sequential structure in the graph. Amdahl [amd67] determines the maximum speedup S_{max} of a program running on an arbitrary amount of processors n against the ratio of the sequential portion in that program, see equ. 5.4:

$$S_{max} = \frac{1}{r_s + \frac{r_p}{n}} \tag{5.4}$$

S_{max} = theoretically achievable maximum speedup according to Amdahl
r_s = ratio of the sequential portion in one program
r_p = ratio of the parallel portion in one program, with $r_s + r_p = 1$
n = number of processors

Amdahl acts on the assumption that the computational task or complexity remains constant when the number of processors is increased, an assumption that is essentially correct and necessary for an understanding of parallelization. Thus, for programs with large sequential portions, parallelization can not significantly reduce the time necessary for execution[24].

[24] Gustafson has modified Amdahl's approach to include the fact that with a higher number of processors, also more complex tasks can be computed in the same execution time [gus88]. Both Amdahl's and Gustafson's law are widely used to predict speedups. Shi demonstrated the equivalence of the two laws [shi96].

The problem with Amdahl's law in the TANGA context is that Components themselves can be of varying complexity. To give an example, the Filter Component's computational complexity depends on the number of filter coefficients actually defined for that specific Component. The absolute overhead for a certain clustering remains constant, but there are occasions on which the gain in terms of GPTI of a certain clustering is outweighed by the overhead introduced by that clustering. As a consequence, the GPTI values computed in the Color Clustering method are not always representative of the actual speedup.

It is therefore necessary to consider the overhead introduced by the parallelization itself. The overhead is determined by comparing the theoretical speedup with the actually achieved speedup. Fig. 5.32 gives an example: on the left hand side, a simple graph consisting of four Components is shown. The GPTI of that graph for one processor is determined to be $GPTI_1 = 22$. For simplicity's sake let's assume that on a single processor the execution time has been measured to be $22s$. The GPTI for two processors is $GPTI_2 = 12$. Yet, when two processors are used, the execution time has been measured to be $14s$. This means that the time necessary to compute the overhead is $2s$ - the overhead is equal to 2 PTI. If the Components are replaced with other, more or less computationally complex Components, then the overhead remains constant whereas the GPTI changes.

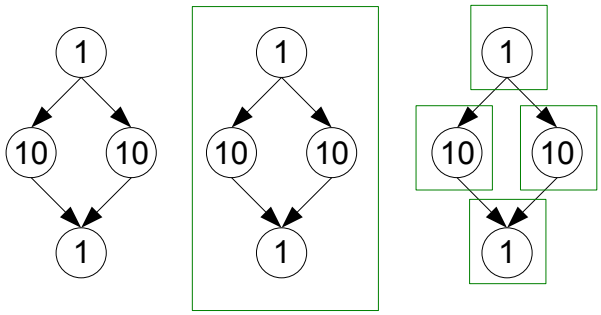


Fig. 5.32: Determination of overhead. *Left*: a simple graph example. *Center*: Clustering for processing in one thread. *Right*: Clustering for processing in two threads.

The theoretically possible maximum speedup can be calculated by applying the Color Clustering to the graph, assuming the overhead to be ideally small: it is assumed to be zero for the maximum speedup.

The theoretically possible maximum speedup has been compared to the actually achieved speedup of the DCPR and the DCCR for two processors (two threads). As can be seen in fig. 5.31, the DCCR is relatively close to the theoretically possible speedup values, but does not match with these. The amount of overhead depends on the structure of the graph and cannot be avoided. Although the DCCR may not necessarily yield the optimum clustering for all structures, the example test cases representing actual application scenarios show a satisfying result.

6 System for Subjective Audiovisual Assessments

This chapter presents a detailed description of the system created for performing bimodal subjective assessments in the auditory and visual domain. The assessment system has been created in the course of work for this thesis, as there were no other tools readily available offering the functionality necessary for performing audiovisual assessments in a fast and reliable way.

Section 6.1 gives an overview of the system's structure, with the three main elements *I3D MPEG-4 Player*, *Input Device* and *SALT logging tool*. It also documents the communication structure and protocol used to exchange messages between the elements.

Section 6.2 explains how the MIDI functionality integrates seamlessly into the I3D. As MPEG-4 itself does not provide an interface for input or output of MIDI data, a solution was created that extends the standard without provoking incompatibilities.

Section 6.3 describes the Input Device that has been designed for usage in audiovisual assessments. A short introduction explains why it is necessary to have such a device in audiovisual assessments at all. The micro-controller based hardware and the firmware are discussed, as well as usability issues which relate to the handling and user interface of the Input Device.

Finally, section 6.4 presents a tool used for logging all input from test subjects during the audiovisual assessments. This tool can also be used to export the collected data in a format suitable for further processing in statistical analysis software.

6.1 System Structure

An overview of the assessment system's structure is given in fig. 6.1. The central element is a PC running the IAVAS I3D MPEG-4 player, which renders the (interactive) audiovisual scenes to be assessed. The I3D can be (and actually is) used also in different contexts [rei07]. Only the features relevant for performing subjective assessments are described in this section. The I3D itself is discussed in more detail, with a focus on the I3D's real-time audio rendering engine TANGA, in chapter 5.

Connected to the I3D is an Input Device with which the test subjects can provide feedback about the perceived quality, and which can be used to control the course of an assessment session. The I3D analyzes the data coming from the Input Device, transforms or complements it, and passes it on to the SALT software running on another PC for logging and saving.

6.1.1 Communication Structure

The communication between the elements of the assessment system needs to be robust and easy to handle. Because only relatively small amounts of data need to be transferred, a serial data connection with a low channel capacity can be used. It should be available or easy to retrofit on modern PC architectures as well as on microcontroller-based devices.

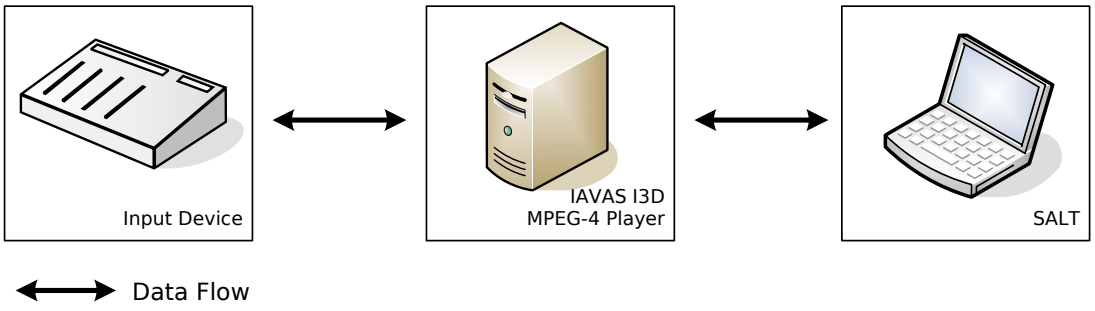


Fig. 6.1: The three elements of the assessment system: I3D MPEG-4 player as the central element, Input Device for input of user feedback and SALT software for logging the course of the assessment.

Therefore, the Musical Instrument Digital Interface (MIDI) was chosen[1]. MIDI is a serial interface with a flexible protocol allowing for custom data to be transmitted in so-called *System Exclusive* (SysEx) messages. It is easy to use and can be retrofit easily to any computer offering a USB interface. The necessary drivers are integrated into nearly all current operating systems. A number of libraries exist for different operating systems that allow the integration of MIDI functionality into own programs, see section 6.2.

MIDI specifies simplex (one-way) serial communication using current-loop signaling at $31.250 baud$, which is roughly a maximum of $3.125 bytes/second$. Since the majority of simple MIDI messages like note-on or note-off commands need three bytes, this equals to a maximum of a little over 1000 MIDI messages per second. It is important to understand that MIDI is a point-to-point communications link between two devices. MIDI is not a multi-drop *bus* like e.g. ethernet. If more than two receiving devices need to be connected to one source, they have to be daisy-chained together. If two devices need to communicate bi-directionally, they need to be connected by two MIDI cables.

Fig. 6.2 shows the MIDI interconnection scheme for the assessment system. The Input Device and the PC running the SALT logging software need to be equipped with one MIDI interface each, the I3D PC needs to be equipped with a double I/O MIDI interface such that two inputs and outputs each are available for the I3D.

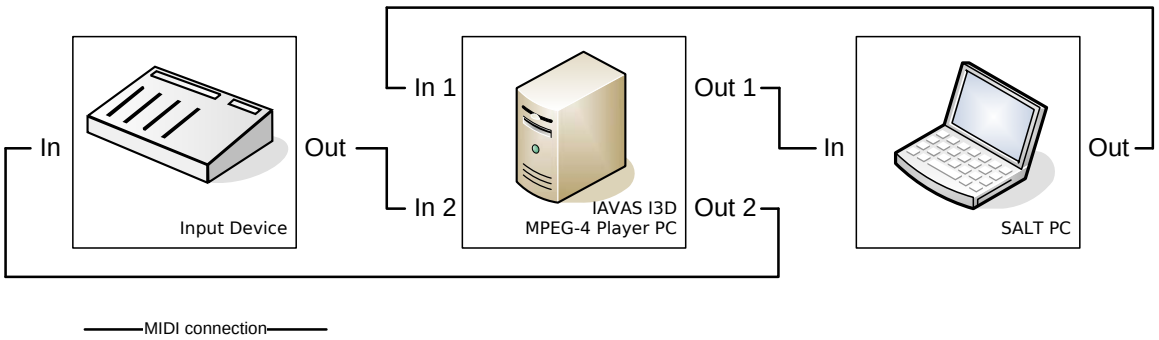


Fig. 6.2: The MIDI interconnection scheme allowing full communication between the three elements of the assessment system.

[1]MIDI was originally proposed as the 'Universal Synthesizer Interface' by Dave Smith in 1981 [smi81] and jointly standardized by the MIDI Manufacturers Association (MMA) and the Association of Musical Electronics Industry (AMEI) in 1983 [w-mma, w-amei].

6.1.2 Communication Protocol

For the MIDI communication between Input Device, I3D PC, and SALT PC, only three different MIDI message types are used. These are the standard MIDI note-on/note-off messages for button press events, standard MIDI Continuous Controller messages for fader movement events, and MIDI SysEx messages for more complex type of data, e.g. text messages to be displayed on the two LCDs of the Input Device.

MIDI Communication: Input Device $\Longleftrightarrow$ I3D

Buttons The buttons on the Input Device send MIDI note-on messages whenever they are pushed, and MIDI note-off messages when they are released. The status LEDs of the buttons can be triggered via MIDI by sending the same MIDI note numbers as the according buttons have. MIDI channel 1 is pre-selected in the Input Device, but can be changed by uploading a new firmware via MIDI.

Faders The faders send out MIDI Continuous Controller (CC) messages when they are moved. MIDI CC messages are 7 bit messages covering a data range of 128 different values. In the Input Device the value range is mapped such that it spans from $0-100$. Therefore, if the fader is moved to the uppermost position, the Input Device will send a MIDI CC message with a value of 100. As the faders are motorized, they can be moved automatically to a certain position by sending MIDI CC messages to the device. The first fader sends controller number `0x15` (`21` decimal) on MIDI channel 1. Therefore, the complete MIDI message for the first fader at the lowest position is `0xB0 0x15 0x00` (`176 21 0` decimal). If there are more faders to be used, these usually send controller numbers `0x16` up to `0x1C`.

LC Display Text can be displayed on the LCDs by sending MIDI System Exclusive (SysEx) data to the Input Device. It can receive encapsulated ASCII strings from the I3D and display them on the LCDs. The first line of display LCD1 is reserved for displaying such messages. The second line is usually configured to show the current values corresponding to the fader positions (but this can also be changed in the firmware of the Input Device). The MIDI SysEx string for displaying text must look like this:

`F0 00 00 7E 40 00 08 01 00 00` text as ASCII data `F7`

The first byte indicates that the following bytes are SysEx data. The next eight bytes are a unique identifier which tells the Input Device that the following data bytes contain text to be displayed on the LCD. The first data byte (the tenth overall byte) determines the cursor position of the text to be displayed. The remaining data bytes contain the actual ASCII encoded text. The last byte (`0xF7`) indicates the end of the SysEx message. As the first display has 2×40 characters, text to be displayed on LCD2 needs to start at cursor position `0x50` (`80` decimal).

The text to be displayed needs to be defined in the scene description. The data can be sent to the Input Device upon certain events, such as the reception of a specified MIDI note-on event corresponding to the press of a dedicated button on the Input Device.

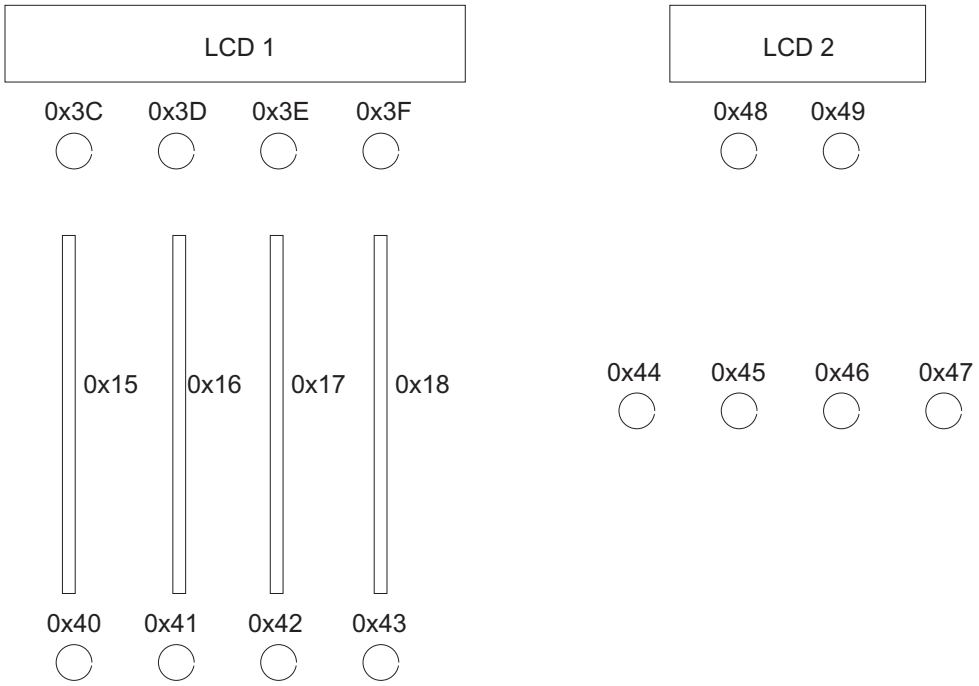


Fig. 6.3: Overview showing the MIDI note-on/note-off numbers for the buttons and the continuous control message numbers for the motorized faders on the standard layout Input Device, see fig. 6.7.

MIDI Communication: I3D $\Longleftrightarrow$ SALT

As well as between Input Device and I3D, all MIDI message types described above (MIDI note-on/note-off, CC and SysEx) can be used in the communication between I3D and SALT. MIDI data received by the I3D from the Input Device is forwarded to the SALT PC.

Additional data can be transmitted that originates from the scene itself. This can be a counter reset signal to indicate how long the test subject took to complete the rating of a test item, or a variable that contains the score achieved in an interactive test scenario (e.g. how many footballs did the subject collect during the task, how many correct choices did he make, etc.). A number of examples for such variables are presented in the discussion of subjective assessments in section 8. These variables are defined as so-called *Valuators* that are part of the scene description and that can be modified using ECMA script [16262]. This kind of data is transmitted as MIDI CC messages. Additionally, MIDI Program Change messages can be sent to control external equipment like audio mixers or video / audio cross bars, see section 6.2.1.

Upon completion of a test item (or a trial), the motorized faders can be reset to a neutral position. As the neutral position may be differently defined in different contexts (depending on the assessment itself), its value can be defined in the SALT test design dialog. Therefore, SALT emits the fader reset command to the I3D, which passes it on to the Input Device. For a fader reset, one MIDI CC message is sent per fader. If e.g. the first fader must be set to a neutral position of half the fader way, i.e. a value of 50 on a scale from $0 - 100$, then the MIDI message to be sent is `0xB0 0x15 0x32` (`176 21 50` decimal).

6.2 MIDI in the I3D

The MIDI devices present in the I3D PC need to communicate with the I3D player application. This is usually done using so-called wrapper libraries that encapsulate functionalities offered by the operating system in order to make access easier. Here, the wrapper library to be used should allow access to MIDI functionality using the C++ language, as the I3D is written in C++.

On the Windows operating system, a wrapper library written by Leslie Sanford is used [w-wrap]. It is open source software published under the LGPL[2] license. It encapsulates the MIDI functions of the Windows Multimedia API in an object oriented way and thus provides easy to use send and receive mechanisms for all types of MIDI messages for the I3D.

6.2.1 I3D MIDI Input Sensor

In analogy to the other input sensors already standardized in MPEG-4 BIFS [14496-11] and present in the I3D (key sensor and mouse sensor), a new class called MidiSensorInstance has been created. Its methods allow to use the MIDI devices that are present in the I3D host PC in the same way as for receiving input coming from the keyboard or mouse. Apart from offering a communication interface for logging events and ratings in SALT, this extension is necessary to allow interaction with and control of the audiovisual scenes under assessment via the Input Device. The MIDI sensor follows the same implementation structure as the standardized input sensors. An overview of the methods of the MidiSensorInstance class is given here.

MidiStart() is called by the constructor of the class MidiSensor. It calls GetDevCaps() to list all MIDI devices present in the system and opens or closes them as needed. These are the devices to which the SALT PC and the Input Device are connected.

ReceiveMsg() is a virtual element function of class CMIDIReceiver of the wrapper library used on the Windows platform. This method is called (as soon as an object of the class MidiSensorInstance has been created) whenever a MIDI message is received by one of the opened devices.

MidiRouter() is called by ReceiveMsg() whenever a MIDI message coming from the Input Device is received. Depending on the type of message, it is either forwarded to the Coder object of the SysCoder class (thus allowing for changes in the scene description, see below) or to the SALT PC for logging of user input coming from the Input Device.

SetIDs() is a method that allows to identify the items and their ratings in the current trial of a subjective assessment. Because the order of presentation of stimuli as well as the correspondence between fader (on the Input Device) and item itself should be randomly shuffled between trials (as well as between test subjects), both are usually modified at random by a script contained in the scene description. The SetIDs() method reconciles

[2]Lesser General Public License - the Lesser GPL permits to make use of the library in proprietary programs, see [w-lgpl].

the items' IDs and corresponding ratings before handling over a complete set of values for one trial to the SendIDs() method.

SendIDs() transfers a complete set of ratings whenever a test subject terminates the trial by pressing the "next" button. The data is contained in a MIDI SysEx message.

SendBoxDisplayText() reads a pointer to the Valuator BoxDisplayTextVAL that contains the text to be displayed on the LCDs of the Input Device.

SendControlSalt() resets an internal time counter in the SALT application to zero. The current value of the counter is logged in SALT for all repeatedly updated MIDI messages, e.g. stemming from continuous rating assessments. This time stamp can also be used to document the exact duration of an experiment.

SendAddInfo() can send additional numerical information. Three Valuators exist that can be used to compose MIDI messages. When a MIDI CC message is composed and sent, this is recorded by SALT. It may contain a high score reached by the test subject or any other numerical information. SALT does not write these messages to the log file when a MIDI Program Change (PC) message is composed and sent. MIDI PC messages are mainly thought for controlling other MIDI devices chained to the same MIDI device output that SALT is connected to - e.g. a mixing desk that needs to recall different presets, or a video cross bar that inserts a gray video screen between the trials as recommended in ITU-R BT.500-11 [itu500].

DataReceive_Ind() receives MIDI messages from MidiRouter() and passes them to the SysCoder. The SysCoder writes these values conforming to the BIFS scene description. Using the so-called Object Descriptors (ODs) of an MPEG-4 scene, these values can be used to modify the scene itself. A more detailed description of this mechanism is given in section 6.2.2.

IsYourName() is a method that every InputSensor class owns. Here it identifies the MidiSensor in the I3D. Calling this method assures that the MIDI data is only passed to the MPEG-4 scene when an instance of the MidiSensorInstance class is actually present. Generally speaking, only when IsYourName() returns *true*, a device that correlates to the Object Descriptor in a BIFS scene is available in the I3D.

MidiStop() is called by the destructor of the MidiSensorInstance class and closes all MIDI devices present in the system.

6.2.2 Use of MIDI Input Sensor in Interactive Scenes

New input sensors can be integrated into the scene by using the InputSensor node as defined in MPEG-4, see listing 6.1.

```
1 InputSensor {
2   exposedField  SFBool    enabled              TRUE
3   exposedField  SFString  buffer               ""
4   exposedField  MFString  url                  ""
5   eventOut      SFTime    eventTime
```

```
6 |}
```

Listing 6.1: Parameter fields of the MPEG-4 InputSensor node.

In order to use the MIDI sensor, the InputSensor node needs to be created and its fields need to be filled with the incoming MIDI data.

Similarly to video and audio data, all kind of user input is handled as so-called elementary streams (ES) in the MPEG-4 delivery layer, see fig. 5.1. Therefore, all user input data needs to be accessed via its object descriptors (OD). An exemplary OD for the integration of the MIDI input sensor into an MPEG-4 scene is given in listing 6.2.

```
1  AT 0 {
2    UPDATE OD [
3      ObjectDescriptor {
4        objectDescriptorID 18
5        esdescr [
6          ES_Descriptor {
7            es_id 19
8            streamPriority 16
9            decConfigDescr DecoderConfigDescriptor {
10             objectTypeIndication 255
11             streamType 9
12             upStream false
13             bufferSizeDB 10000
14             maxBitrate 0
15             avgBitrate 0
16             decSpecificInfo UIConfig {
17               deviceName "MidiSensor"
18             }
19           }
20           slConfigDescr SLConfigDescriptor {
21           }
22         }
23       ]
24     }
25   ]
26 }
```

Listing 6.2: Excerpt from a scene description showing the OD for the integration of the MIDI input sensor.

An "initial object descriptor" (contained in every scene description, not shown here) is extended by the OD contained in the `UPDATE OD` instruction `AT` time 0. In order to access this object with an input sensor, the `objectDescriptorID`, in this example 18, needs to be referenced in the *url* field of the corresponding InputSensor node, see listing 6.1. Of course, all IDs used in a scene must be unique to prevent ambiguities.

The `ES_Descriptor` contains media specific information about the ES that is attached to the OD. For video streams, the most important fields would be the decoder configuration descriptor `decConfigDescr` containing the encoding format, the maximum and average bit rates, the buffer size, and so on. For MIDI streams these settings are not relevant, maximum and average bit rates can be set to 0 without negative effects on the accuracy of timing.

Finally, the decoder specific information field `decSpecificInfo` must contain the device name of the input device that should be used to access this object. In the I3D, the MIDI input sensor is registered as "MidiSensor". The notation must be exactly the same, otherwise the device cannot be identified and an error message is produced by the I3D.

As the MIDI messages used to modify the scene description can consist of up to four MIDI data bytes, four so-called Device Data Frames (DDFs) are inserted by the SysCoder into the corresponding ES. These DDFs need to be decoded from the ES and assigned to

four Valuators. This process is shown in fig. 6.4. To fill the Valuators with the MIDI data bytes, a `REPLACE` command needs to be issued in the *buffer* field of the InputSensor node. Listing 6.3 shows the corresponding excerpt from an MPEG-4 scene description, with the *url* field set to the object descriptor ID as defined in listing 6.2. Note that although a `REPLACE ... BY 0` command is issued, the Valuators are not actually replaced by 0 but by the values contained in the corresponding DDFs. Although this is irritating at first, it is the only way to do it correctly.

The Valuators that the DDFs' contents are assigned to, need to be defined beforehand in the scene description. These Valuators can then be used to trigger other Valuators (for normalization of values), Conditionals, or proprietary functions implemented using the ECMA script language, see section 5.3.5.

```
1 InputSensor {
2   url ["18"]
3   buffer {
4     REPLACE DataByte1VAL.inSFInt32 BY 0
5     REPLACE DataByte2VAL.inSFInt32 BY 0
6     REPLACE StatusVAL.inSFInt32 BY 0
7     REPLACE ChannelVAL.inSFInt32 BY 0
8   }
9 }
```

Listing 6.3: Replacing the buffer fields with the MIDI values passed by the SysCoder.

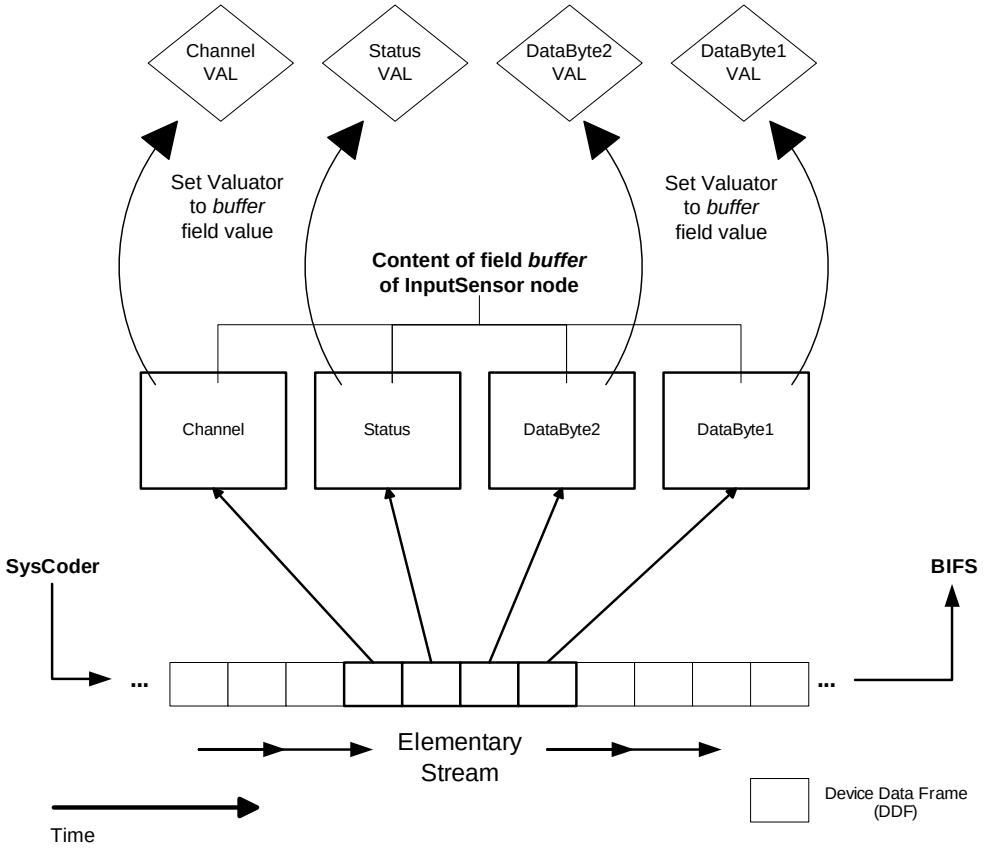


Fig. 6.4: MIDI input data contained in an ES is transformed into Valuators that can be used to modify the scene description.

On other platforms than the Windows operating system, the MIDI functionality has not been integrated yet. A solution would be to replace the Windows MIDI wrapper library

by Leslie Sanford with the RtMidi C++ classes by Gary Scavone. Both implementations provide a callback mechanism to notify the application of incoming MIDI data (instead of polling for data), a feature that is necessary to keep the computational load as low as possible. RtMidi is available across Linux (ALSA), Macintosh OS X, SGI, and Windows (Multimedia Library) operating systems [sca05, w-rtm].

6.3 Input Device

The traditional methods for performing audiovisual assessments are the so-called introspective methods that encompass perceptual and affective measurements, see section 8.1. In order to determine the level of perceived quality, test subjects need to reflect on the percepts and judge them according to their expectations or in relation to a reference stimulus. They need to fill in a questionnaire, usually by dragging a computer mouse to move virtual sliders on a computer screen or by using pen and paper, which is a process that inevitably cuts in the flow of visual percepts coming from the system under assessment. It is therefore necessary for the test subjects to "switch" between different perceptual situations (the one under assessment and the act of rating the assessed system itself). This is a process which diminishes the continuity of percepts and which therefore inevitably influences the subjects' perception.

This makes it necessary to find a method with which test subjects can deliver their ratings in such a way that the cuts in the flow of percepts are as small as possible. One important idea is to reduce the amount of distraction originating from the assessment system itself as much as possible by using a different, unused modality for the feedback channel. Whereas in the subjective quality assessments of audiovisual systems auditive and visual percepts are evaluated, haptics is a modality which is still "free" to be used in a different context. Therefore the use of the haptic channel is proposed for receiving feedback from the test subjects: they have to manually move hardware faders into a position that corresponds to their rating on a scale from 0 to 100. Of course, the scale is not necessarily a numbered one but can also consist of semantic designators, e.g. *bad* to *excellent*, as described in ITU-R BS.1284 [itu1284].

Also, the question of reliability of the Input Device is especially important. Because subjective assessments tend to be a very time consuming issue both for test subjects as well as for test personnel, it is very important that the device (and the system as a whole) functions at all times without producing random or systematic errors.

6.3.1 Usability

Another aspect to be considered is that the Input Device needs to be designed in such a way that the layout of its input elements (buttons, faders, rotary knobs, etc.) is clear and functional. Test subjects new to the device should be able to easily understand its functionality, so that the number of errors due to false handling of the device are minimized. Informal usability tests have been performed in the Usability Lab of the IMT which showed that the number of input elements needs to match the number of attributes or items to be rated. This means that for MUSHRA-like assessments (MUltiple Stimuli with Hidden Reference and Anchor, see [itu1534]), where the test subject has to compare the quality of e.g. five items, five faders should be placed on the device's surface. On the other hand, if there is only one attribute to be rated in each scene (single-stimulus method), then

there should be only one fader present. This results in the need to make the Input Device highly customizable for each assessment to be performed. It was chosen to use a design where the input elements can easily be connected and disconnected to the hardware. By using different front plates for the Input Device it can be modified to meet most subjective assessment scenarios imaginable, see fig. 6.5.

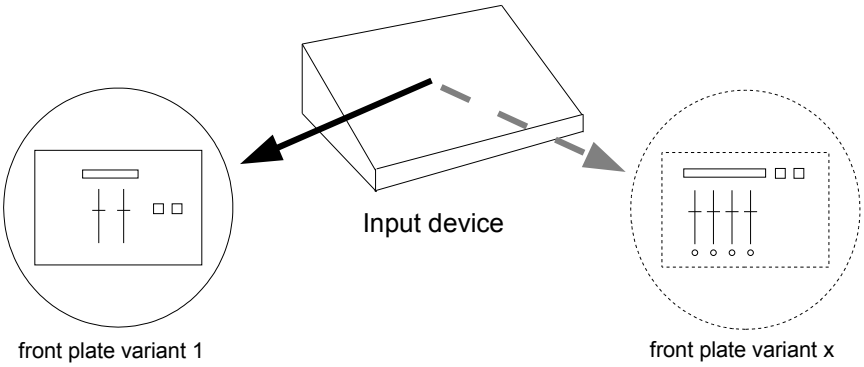


Fig. 6.5: Different front plates can be used to customize the Input Device according to the type of assessment to be performed. The basic hardware remains the same.

In the last paragraph considerations regarding the usability of the Input Device for the test subjects have been presented. But also the ease of use of the whole system for the assessment personnel should be considered. Handling errors (which might present themselves as test data errors or - even worse - test data falsifications[3]) can only be minimized if the system as a whole is easy to operate. The supervision of the system necessary during the test session should be reduced to a minimum, so that the test supervisor only needs to take very few decisions / actions like saving the data at the end of the session, when the test subject has completed all items. The configuration phase of the Input Device therefore needs to be completely separated from the assessment phase. Configuration of the system is further discussed in detail in section 6.4.

6.3.2 Hardware

The hardware design of the Input Device is based upon the open source hardware project *MIDIbox* created by Thorsten Klose [w-mio]. The decision of not developing the hardware of the Input Device from scratch, but to use a design already tested by a large number of users, was mainly based on the reliability considerations presented briefly in section 6.3. *MIDIbox* consists of a Microchip PIC microcontroller based modular hardware design with a specialized operating system (*MIOS*). Its main goal is to allow the construction of MIDI controller boxes for home-recording purposes, and the community has come up with a remarkable wealth of different controller box designs.

Most of these designs are built around the PIC 18F452 microcontroller, as is the Input Device described here. The mainboard of the *MIDIbox* system is housing the microcontroller itself, the power regulation and stabilization, the LCD connector, and the MIDI in/out connectors. It can be linked to a number of other boards providing analog and / or

[3] Whereas errors are usually detected in the process of analyzing the obtained data, falsifications are hard to notice because typically the data patterns are conserved.

digital in/outputs, as well as to graphical or character LCDs, additional MIDI in/outputs, and flash ROM memory extensions.

The basic hardware configuration of the Input Device consists of the mainboard, to which a number of daughterboards are connected, see fig. 6.6. These are:

- One motor-fader board for the connection of up to eight motorized 100mm analog faders with touch sensitivity. Any other analog input elements, e.g. joysticks or touch pads, can be connected as well.
- One LTC board which provides a second MIDI output and two LEDs indicating MIDI in/out activity.
- Four digital input boards offering a total of 128 digital input pins. Therefore, up to 128 keys or buttons can be used, or 64 incremental impulse type rotary knobs, or any combination of these.
- Four digital output boards offering a total of 128 digital outputs. These can be used to drive status LEDs indicating on/off status of keys, circular arrays of LEDs around rotary knobs indicating momentary values, seven-segment LED displays for numbers, etc.

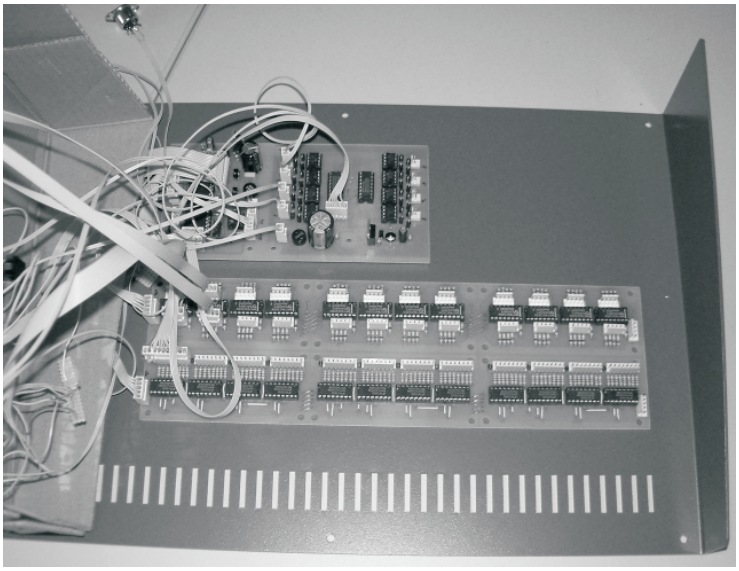


Fig. 6.6: Mainboard and motor-fader board (above) and three out of four digital input / output boards (below) of the prototype Input Device.

This basic hardware configuration remains installed inside the Input Device, independently of what front plate and what number and configuration of input elements is actually connected. Currently only a maximum of 14 inputs and 12 outputs is used in the "largest" front plate design, see fig. 6.7. Therefore, the number of digital input and output boards installed could be reduced.

6.3.3 Firmware

A new firmware has to be uploaded to the Input Device via MIDI whenever a different front plate is connected. The firmware contains all information about type and number of

input elements and the input pins these are connected to. The same goes for LEDs and character LCDs mounted on the front plate. Fig. 6.7 shows the Input Device configured for usage in a multi-stimulus test (direct comparison between three items and a reference).

Fig. 6.7: The Input Device configured with three motorized faders for a multi-stimulus test.

To use the Input Device for a pair comparison test, two motorized faders can be used, see fig. 6.8. Two buttons, one underneath each fader, toggle the item to be rendered by the I3D MPEG-4 player: e.g. item 1 is played when the left button is pressed, and when the right button is pressed, I3D switches to playback of item 2. A direct comparison of the two items is possible. The test subject can then use the two faders to enter his rating of the items. When the rating is entered, the *next scene* button can be pressed, the motorized faders are moved to a neutral position, and the test subject can start the comparison of another two items by pressing one of the two buttons beneath the faders.

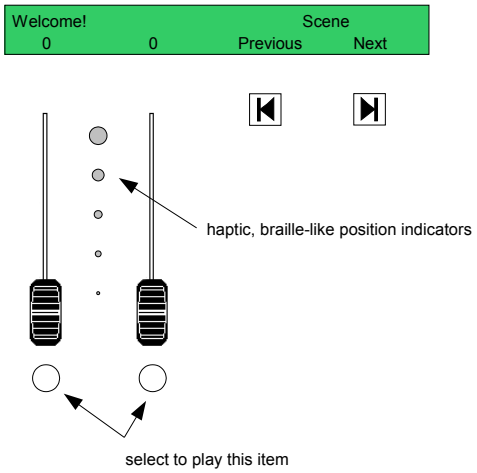


Fig. 6.8: A front plate example for a pair comparison test.

To address the problem of split attentiveness, test subjects should not necessarily have

to visually control the fader position corresponding to their rating.[4] One solution implemented is the division of the rating scale into different sections. Each time the fader is manually moved across a section border, the motor stops or shakes the fader very slightly so that test subjects experience a subtle vibration coming from the fader. Say the scale is divided into five sections, then on a scale from 0 to 100 each section is 20 scale units long. Manually moving the fader from e.g. scale level 5 to scale level 67 will then cause three short sequences of vibrations in the faders (when crossing scale levels 20, 40 and 60, respectively). Another option would be to add braille-like, protruding indicators along the faders' tracks as suggested in fig. 6.8.

6.4 SALT - Subjective Assessment Logging Tool

SALT is an acronym for *Subjective Assessment Logging Tool*. It is a cross-platform JAVA application which mainly serves three purposes: the recording of personalized test subject data, the complete logging of input provided by the test subject via the Input Device during a test session, and the automation of the Input Device's motorized faders. Furthermore, SALT provides also functionality to export logged data into formats readable by statistical analysis software like SPSS [w-spss] or R [w-r]. SALT was programmed by Stefan Holzhäuser [hha05] and Mandy Weitzel [wei07] in the course of two diploma thesis projects under the supervision of the author.

SALT provides a graphical user interface, see fig. 6.9, and needs to be configured according to the hardware it is supposed to work with. I.e., SALT needs to be configured to match the amount of motorized faders and the buttons available to the test subjects on the Input Device. On the other hand, SALT must know the structure of the subjective assessment to organize the recorded data in a sensible way. Therefore, two main paradigms exist in SALT: the *design* and the *session*. These will be explained in the following sections.

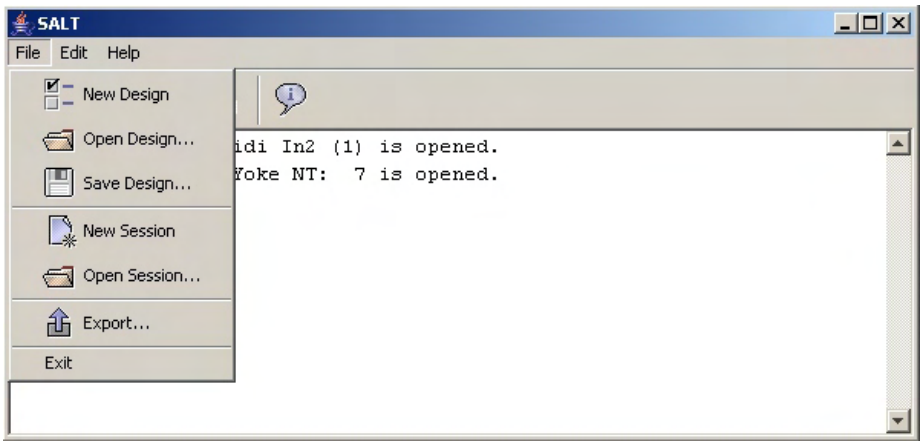


Fig. 6.9: The main menu of SALT provides commands related to the assessment design, the session organization or structure, and the export of data.

[4] Visual feedback on the exact current fader position is presented on the LCD right above the faders, either numerically or as semantic designators.

6.4.1 Assessment Design

A design specifies the general settings of a subjective assessment. Fig. 6.10 shows the design dialog GUI that can be used to enter this information. The GUI consists of a tabbed window with four tabs. In the *Additional Subject Infos* tab (upper left), additional fields containing assessment-specific subject information can be created. These fields will then appear in the pop-up window immediately after starting a session, see section 6.4.2. Apart from the parameters that are always asked to identify a test subject (name, age), here fields like *acuity of vision* or *listening experience* can be added. The relevance of these may depend on the goal and / or design of the subjective assessment itself, therefore such additional information is not always asked.

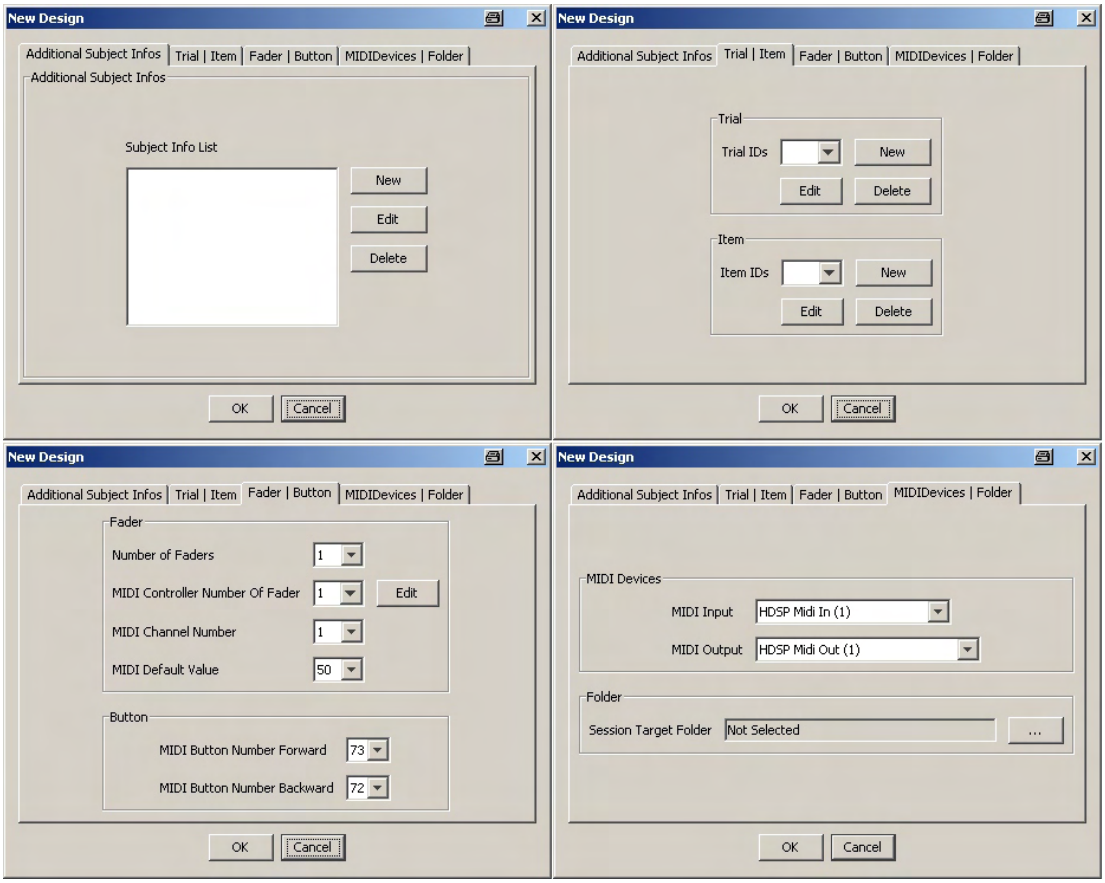


Fig. 6.10: The four tabs of the design dialog GUI that allow to specify the main cornerstones of an assessment design.

The structure of a subjective assessment is often organized in so-called *trials* and *items*. An item is an instance of the attribute to be judged or rated by the test subjects. Thus, an item can be a piece of music deteriorated by a certain amount of artifacts - a second item will then have another amount of artifacts present in the signal. An item can also be an audiovisual scene in which the room acoustic simulation is performed with a certain number of image sources - whereas another item has another number of image sources. Finally, to give another example, an item can also consist of an interactive scene with a certain difficulty of a task given to the test subject - here, items may differ in the degree of difficulty of the task.

In turn, a trial comprises a certain amount of items and groups these into a contextual

relation. A trial may group the items compared in a multi-stimulus test, or it can hold a sequence of items presented in a single-stimulus assessment (also known as Absolute Category Rating (ACR), see e.g. [itu500, itu910]). Also, a trial may only contain a single item, again depending on the structure of the assessment.

The second tab of the design dialog GUI shown in fig. 6.10 (*Trial/Item*) holds the trials and items of an assessment: each needs to have a unique ID as well as a semantic designator. Trials and items can be added, modified or deleted from a test design. During the assessment, the I3D will transmit the trial and item IDs along with the corresponding rating given by the test subject. SALT will then add the semantic designators according to the settings of this dialog.

The third tab (*Fader/Button*) specifies the number of motorized faders on the Input Device's front panel (between 1 and 8), the MIDI controller numbers they send, the MIDI channel used, and the default value that the faders are reset to in the beginning of each trial (see section 6.3.3). Furthermore, the MIDI note numbers of the "next" and "back" buttons need to be declared.

Finally, the fourth tab (*MIDIDevice/Folder*) specifies the MIDI I/O ports to be used for the communication with the I3D, and a folder can be selected to which all session data is written to. Assessment designs can be saved to be re-used at a later time. Before starting a new session, a design has to be created or loaded from file.

6.4.2 Assessment Session

When a new session is started, SALT first asks for the current subject info as specified in the *Additional Subject Infos* tab, see 6.4.1. When the data has been entered, SALT checks the availability of the MIDI ports specified, resets the motorized faders to the neutral position, and starts to record incoming MIDI data. The assessment supervisor is presented with a pop-up window in which he can select to stop the session. All the recorded data of one test subject is then stored in a log-file in eXtensible Markup Language (XML) format [w-xml]. The order of item data in the log-file corresponds to the chronological order in which items were rated. This order is usually different from subject to subject, as the presentation order is stochastically varied between subjects. Listing 6.4 gives an example of such a log-file recorded with test subject *Lucky Luke*, age *25*, *m*(ale) gender, and with *n*(o) listener experience. The example session consists of one trial designated as *Reverb_2.0s* (ID *1*) with two items designated as *1back task* (ID *1*) given a rating of *50*, and *2back task* (ID *2*) which received a rating of *75*.

```
1  <? xml version =" 1.0 " encoding ="UTF -8"?>
2  < testResults xmlns =" urn:salt:result ">
3     <! -- personalized subject data-->
4     < subjectInfo >
5        <firstName >Lucky</ firstName >
6        <lastName >Luke</ lastName >
7        <age >25</ age >
8        <xtraInfo name ="Gender">m</ xtraInfo >
9        <xtraInfo name ="Listener Experience">n</ xtraInfo >
10    </ subjectInfo >
11    <! -- subject ratings -->
12    <trial name ="Reverb_2.0s" id="1">
13       <item name ="1back task" id="1">50 </ item >
14    </ trial >
15    <trial name ="Reverb_2.0s" id="1">
16       <item name ="2back task" id="2">75 </ item >
17    </ trial >
```

```
18 |</ testResults >
```

Listing 6.4: An example XML log-file written by SALT.

When all sessions of the assessment have been recorded, the joint data of all or parts of the test population can be exported for further processing. The collected data is written to a *.sum file that uses the Character Separated Values[5] (CSV) format, a standard that can be imported into all statistical analysis programs. Previous to combining the session data in a single file, the session data is structured in the following way: ratings are sorted in ascending order, first by trial ID and then by item ID.

Fig. 6.11 shows the GUI of the Export dialog window. In the *Sessions* section, sessions to be included for data export can be selected. The *Subject Info* section allows to specify which of the previously asked additional subject infos shall be included in the exported file, and whether the items shall be identified by their ID or by their semantic designator (name). Finally, a folder can be selected in which the *.sum file is stored.

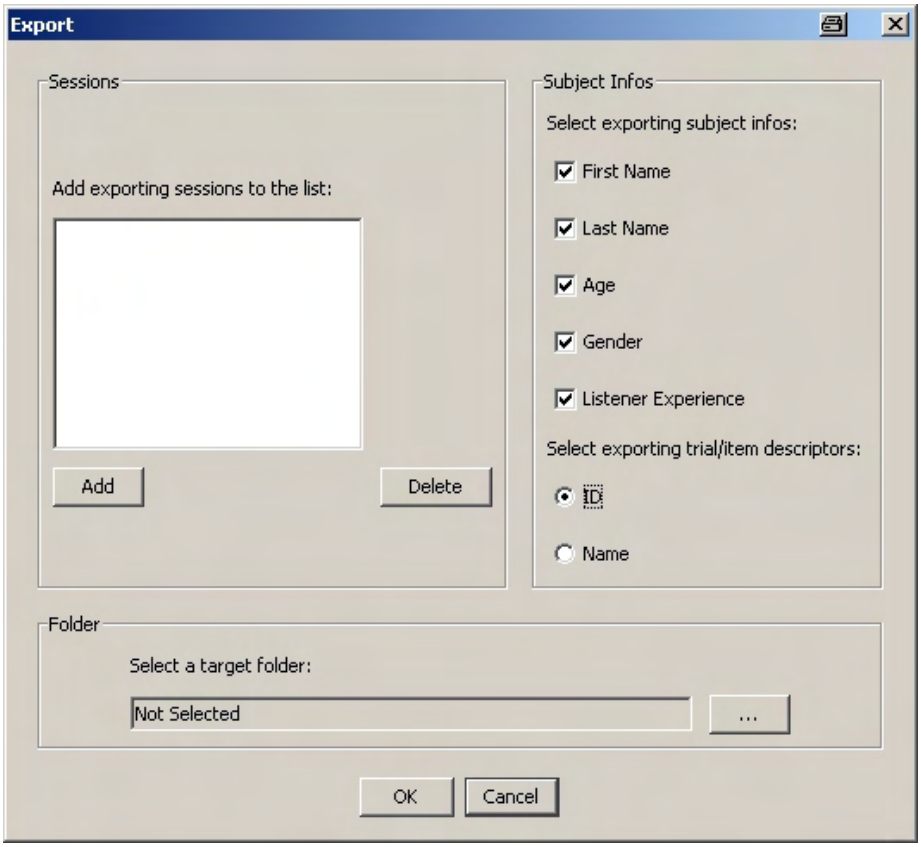


Fig. 6.11: The Export window of SALT allows to select the sessions to be included in the complete assessment data file.

[5] CSV is a computer file format dating back to early business computing applications. The entries of a spread sheet are separated by so-called *delimiters* or delimiting characters. Originally, CSV stands for *Comma Separated Values*, but is commonly used for any file format using delimiting characters to separate tabular data. For more information see the Internet Engineering Task Force (IETF) document RFC 4180 [w-ietf].

Part IV

Enforcement

7 Salience Model

As already discussed in the introduction of this thesis, the question of saliency of objects in audiovisual scenes is only recently becoming an issue of examination. Until now, many everyday application systems mainly rely on visual display and feedback to the user, with some kind of "support" in the auditory domain. As the computing power available in home application systems is constantly increasing, we see a tendency toward integrating more modalities which until now have only been available in specialized Virtual Reality (VR) systems. With this development, users can expect an enhanced degree of immersion. This is interesting in many aspects, one of them being that applications with higher immersion are generally considered more user-friendly, because they provide a feeling of personalization. Additionally, these systems better represent real life and the complexity of real-life experiences by offering information multimodally.

The general problem with this approach is that resources in consumer application systems are always limited, such that it is not feasible to perform a fully grown, detailed simulation of multimodal impressions in real-time. 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 audiovisual percept are.

In the absence of information about the history of an interactive process, an object can be considered salient when it attracts the user's visual attention more than the other objects [lan01]. This definition of salience originally valid for the visual domain can easily be extended to what might be called "multimodal 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.

Of course, a salience model requires a user model of perception, as well as it needs a task model. The user model describes familiarity of the user with the objects' properties, as attention on the properties of an object varies with the user's background. This correlates to the concept of schemata described in section 2.5.2. Whereas a picture of a human being or a human speech utterance can be considered more or less equally salient to all users, 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, so that depending on the task the user is given, his focus will shift accordingly.

Salience also depends on the physical characteristics of the objects themselves. Following the Gestalt theory introduced by Wertheimer [wer23], the most salient visual form is the one requiring the minimum of sensory information to be treated. In the auditory domain it is known that certain noises which can be characterized with properties like 'sharpness' or 'roughness' call the attention more than others [zwi99], often by skirting masking effects

in the time or frequency domain due to their spectral or temporal characteristics. Adding to this, salience can be due to spatial or temporal disposition of the objects. Thus a classification of the properties that can make an object salient in a particular context, the so-called *influence factors*, would have to be established and verified in order to draw any useful conclusions from the "multimodal salience" approach.

One of the most interesting aspects of a salience model in the context of audiovisual application systems is its dependency on the degree of interactivity that the application offers to the user. If the user is allowed to interact freely with the objects in a virtual three-dimensional scene, then it is quite easy to determine the user's focus. Obviously, the user'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, applications with fewer interaction possibilities (as well as applications which do not force the user to interact) are less likely to provide a sense of "being there" to the user. Thus, interactivity is important for realism of Virtual Environments in two different ways: first, it allows the user to do something in the virtual world, and second, it allows the application to determine the user's momentary focus. This information can then be used to enhance the auditive appearance of the object in focus, e.g. by rendering more realistic early reflections to the scene.

Still, there are obviously situations in which the application has no information about the user's current focus, so it is useful to have a multimodal salience model to classify the objects contained in the scene. Yet, no such generalized multimodal salience model exists. One of the reasons why such a model has not been devised yet is that the cost of a perfect simulation of reality in real-time is prohibitive. Still, for experiments rendering results that are representative for real-world experiences, the stimuli that are evaluated should be of the same quality as real-world stimuli.

It might be arguable whether a "perfect" reproduction of the properties of a real life experience will ever be possible in a Virtual Environment at all; let's assume a simulation as being perfect as long as there is no perceptual difference to reality detectable by the human senses. A lesser interpretation of this applies to scenes which have no counterpart in reality: their appearance needs to be plausible in every aspect, also in a sense of perfect agreement between the cues offered by the system in the different perceptual domains.

In the context of this work, 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, we shall require that the part of the virtual scene that is displayed (audiovisually) is perceived as plausible. It is thus accepted that stimuli coming from the surrounding real world (which cannot be entirely excluded in audiovisual application systems of moderate complexity) might interfere with those from the virtual scene.

Section 2.5 has introduced three models of human perception that operate on different levels of abstraction: Neisser's Perceptual Cycle, a very abstract representation of the perceptual process as a learning, comparing and exploring task; Shiffrin and Atkinson's model that tries to describe human perception on the basis of a control process that steers attention; and the BTL multisensory perceptual model as proposed by Hollier and Voelcker that operates on separated auditory and visual layer models with associated error descriptors influencing the so-called "task-related perceptual layer", a module that represents the user's background and task.

All of these models have been verified and accepted to represent certain characteristics of the human perceptual process. None of the models can be ruled out as not being

representative, therefore they all could serve as a basis for a combined perceptual model on which the above mentioned *influence factors* determine the degree of saliency of the perceived objects.

7.1 Posit: Salience Model for Interactive Applications of Moderate Complexity

A salience model would thus mainly contain (and ideally quantify) the *influence factors* that control the level of saliency of the perceived objects. It is therefore necessary to get away from a generalized salience model. A generalized salience model would be too complex and the influence factors too manifold to cope with at this state of knowledge. Instead, it is reasonable to focus on a salience model valid for interactive audiovisual applications of moderate complexity.

Fig. 7.1 shows how such a salience model may be structured. The basis of human perception are the *stimuli*. For interactive applications these are generated by the application system itself, so they will depend on a number of factors: the influence factors of level 1. They comprise the loudspeaker and visual reproduction setup, input devices for user feedback to the system, etc. Influence factors of level 1 are those related to the generation of stimuli.

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

Cognitive processing produces a response by the user. 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 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 also the most difficult to quantify: experience, expectations, and background of the user; difficulty of task (if any); degree of interactivity; type of application; etc. 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 based on the weighted sum of all influence factors of types 1-3. This quality impression cannot be directly quantified but needs additional processing to be uttered in the form of ratings on a quality scale, as semantic identifiers, and so on.

The overall quality impression is, in turn, made up from evaluating single or combined quality attributes. The scientific community has developed a number of attributes that are believed to be relevant for an overall audiovisual quality impression, see section 8.5.3, but again, a quantification of their impact is hardly possible as of now. This is because their weight not only depends on the audiovisual scene (the stimulus) under assessment, but also on the experimental evaluation itself. An attribute that is explicitly asked for will be assumed to be of higher importance by the test subject (we know from our experience that only important things are asked for in any kind of test). The subject's attention will be directed toward the attribute currently under assessment, an act that distorts unbiased perception of the overall multimodal stimulus. Therefore, the subject's reaction might be

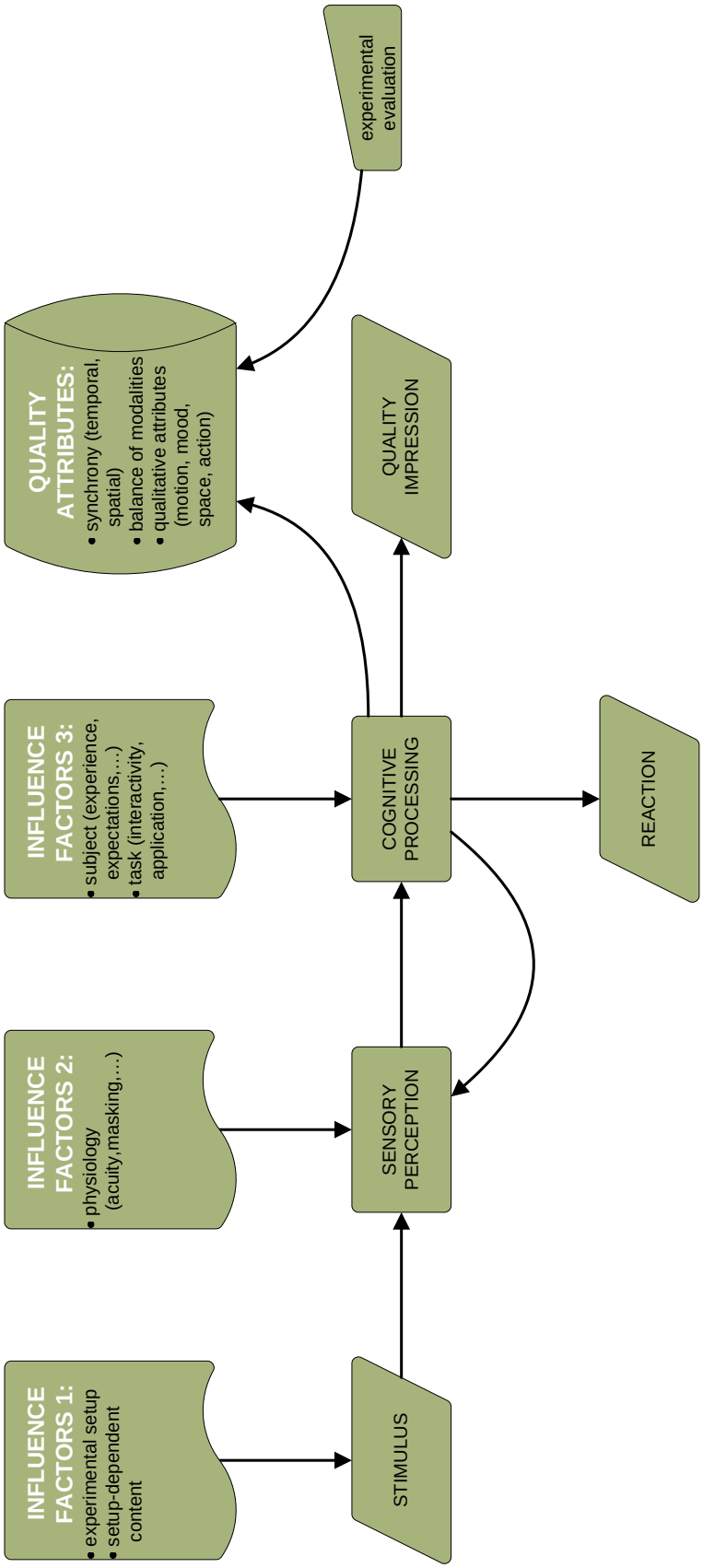


Fig. 7.1: A salience model for interactive audiovisual applications of moderate complexity.

influenced as well. This issue is further discussed in chapter 8.

So, in order to come up with reliable criteria for predicting the most important quality attributes, the influence factors at all levels have to be known. Influence factors can be determined either from scientific literature describing earlier experiments, e.g. as discussed in section 2.4, or from subjective assessments as documented in the subsequent sections of this work. The following chapter will present, apart from a discussion of assessment methodology, a number of experiments that help to verify and identify these influence factors more clearly. Whether their impact upon perceived overall quality can already be quantified remains to be seen.

8 Bimodal Evaluation

According to Woszczyk et al., the human brain combines all auditory and visual information and develops a "composite sensory image" [wos95] based on correspondences and contradictions in the modalities, see fig. 8.1. This statement is in full agreement with the perceptual model by Hollier and Voelcker discussed in section 2.5.5.

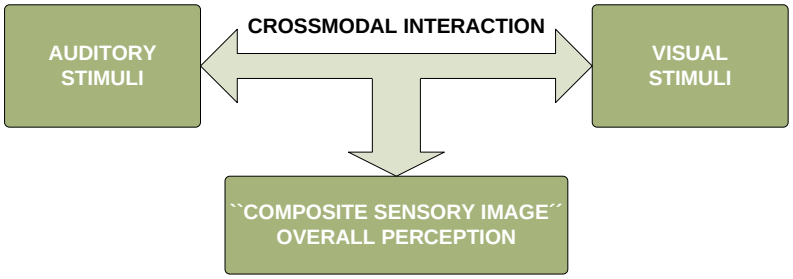


Fig. 8.1: Crossmodal interaction resulting in "composite sensory image", after [wos95].

Therefore, when trying to evaluate the overall quality of an audiovisual presentation, the assessment approach is most important. It seems necessary to assess the quality of the composite sensory image as a whole and not the auditory or visual impression separately and later "combine" the results in some way. Otherwise correspondences and contradictions in the modalities cannot be included. Yet, there are currently neither substantiated recommendations nor tools available that would allow for an easy application of such an approach.

A discussion of the dilemma, suggestions on how to solve it and a number of bimodal assessments actually performed are presented in this chapter. After a definition of the types of measurement that can be used in subjective assessments, an overview of the related recommendations issued by the international standardization bodies is given. This is followed by a discussion of the fitness of these unimodal recommendations for the bimodal (audiovisual) case. Based on the problems of the existing unimodal recommendations, a novel categorization of audiovisual assessments is suggested. This is complemented with a discussion of test methods, procedures and assessment scales suitable for audiovisual assessments. Finally, an introduction to the fundamentals of statistical analysis is given before the assessments performed are described in detail. Chapter 8 concludes with the outline of a novel mixed method approach that collects and combines quantitative and qualitative data.

8.1 Perceptual vs. Affective Measurement

As we have already seen, audiovisual perception is a highly integrative process. The percepts from the two modalities interact at some point in the processing chain, they influence each other and cannot easily be separated analytically. This affects the methods that can be applied to measure single attribute's quality or overall quality of an audiovisual scene.

Generally, in subjective assessments two types of measurements are used - the perceptual and the affective measurement. A perceptual measurement usually seeks to assess the quality of a specific, individual attribute. For audio signals, this can be the perceived spatial location of the sound source, its loudness, pitch, duration and so on. Whenever the stimulus presented is more complex, like e.g. a musical signal used in audio quality assessments, it is obvious that more than one attribute is likely to be excited. Furthermore, attributes are not always separable (they are not necessarily orthogonal), as e.g. *timbre* and *envelopment* of an orchestra recording may vary with the *perceived room size*. In these situations, it is important (yet most difficult, or even impossible) to assure that test subjects are analytical and assess only the attribute in question. Finally, taking the example of *envelopment*, not all semantic descriptors used to identify the attributes are unequivocal. Even expert listeners have different understandings of the term. Rumsey has tried to analyze this problem and comes up with suggestions [rum02], but the base line is that the correct use of terminology in most cases requires extensive training for test subjects.

In more applied unimodal as well as in bimodal (audiovisual) experiments, not only specific attributes are assessed, but additionally they often include questions related to acceptance, pleasantness, annoyance, etc. This type of assessment is called affective measurement. Also, the overall quality impression is often asked for, so the test subjects need "to use a more integrative frame of mind" as Bech and Zacharov call it [bec06]. Affective measurements are usually influenced by all aspects of the presentation, even including personal preferences, context, history, personal experience, expectations, etc.

In multimodal experiments, this brings up a rather difficult to solve problem. When assessing audiovisual quality with the aim of finding out what the most important attributes in terms of overall perceived quality are, it is not expedient to ask test subjects to concentrate on an individual attribute to assess. Assessing individual attributes will disrupt the integrative frame of mind. Yet, when asking for the overall perceived quality, it is difficult to know what exactly the attributes were that influenced test subjects to give a certain rating. In other words, an overall quality rating is provided by the test subjects, but the grounds on which that rating has come up are left unknown.

8.2 Existing Standards

The International Telecommunication Union (ITU) has come up with a number of normative suggestions of how to perform subjective assessments of perceived quality in general. These suggestions are internationally recognized and allow to compare the results of assessments carried out in different laboratories. They define the test conditions as well as the form of presentation. Suggestions on the rating scales to be used as well as on the classification of the test material are given.

Unfortunately, all these recommendations are mostly related to assessments in one single modality. Most recommendations either focus on audio or video quality alone, without taking into account possible cross-modal effects. There are only very few recommendations that relate directly to quality assessments of audiovisual program material. Furthermore, none of these recommendations takes into account possible effects caused by interaction between user and program material or scene.

Figs. 8.2 and 8.3 show the most important ITU recommendations in the context of audio and audiovisual quality. ITU-R BS.1284 provides a guide to general requirements

for performing listening tests with an overview of experimental design, selection of test subjects, methods, statistical analysis and reporting [itu1284]. ITU-R BS.1116-1 contains guidelines for the assessment of systems that introduce impairments so small as to be unnoticeable without careful control of the experimental conditions and an appropriate statistical analysis [itu1116]. It is not suited for systems with relatively large and easily

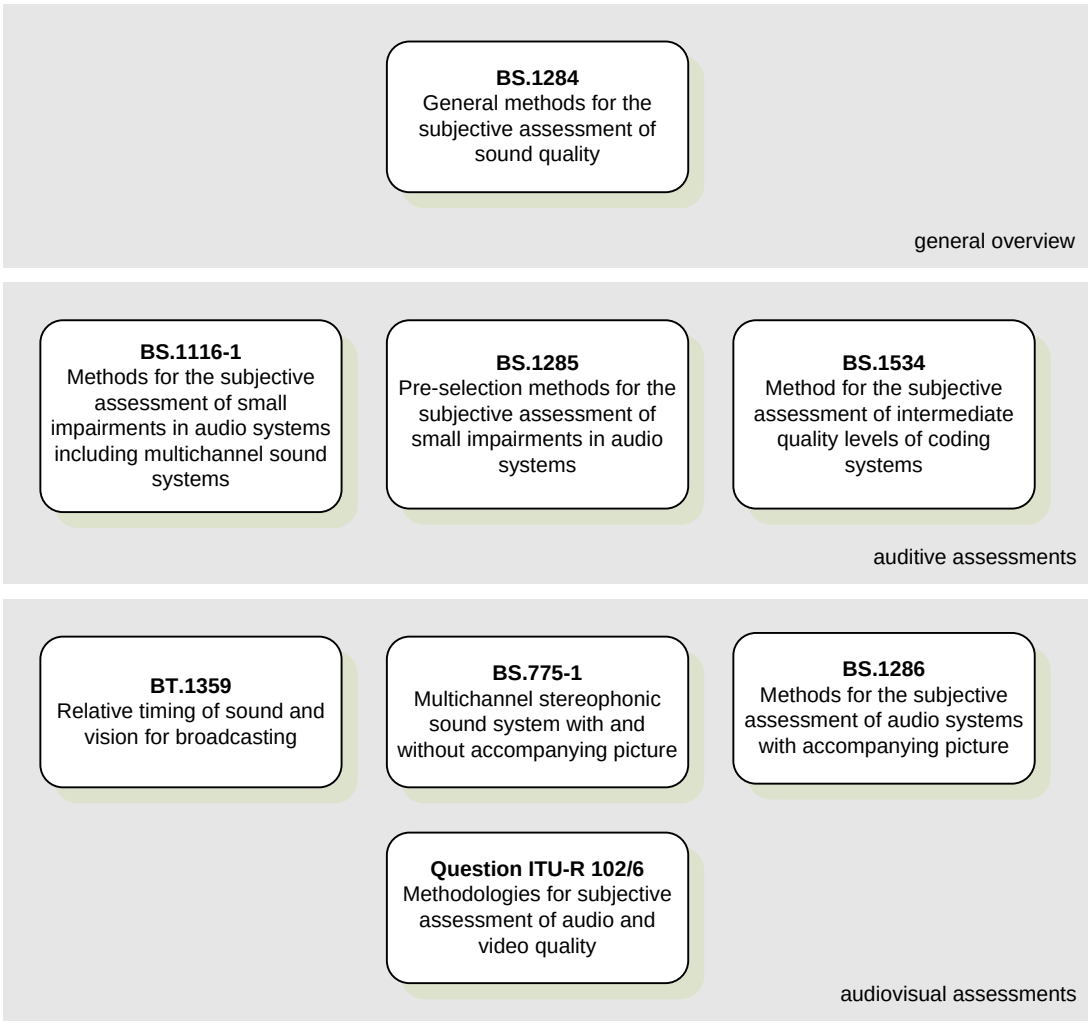


Fig. 8.2: Recommendations and questions of the ITU-R (Radiocommunication Sector) related to audio and audiovisual quality assessments.

detectable impairments. ITU-R BS.1285 [itu1285] complements ITU-R BS.1116-1, as it suggests methods for determining if impairments are small enough for ITU-R BS.1116-1. ITU-R BS.1534 handles testing of larger audio quality differences [itu1534]. Parallel testing with headphones is described, and aspects of presentation order are discussed.

ITU-R BT.500-11 provides details on methodologies for the evaluation of television picture quality [itu500]. Most important in the audiovisual context, ITU-R BT.500-11 contains recommendations for viewing distances, illumination levels, screen sizes for different resolution and aspect ratio displays, etc. ITU-R BS.775-1 makes suggestions on the reproduction setup for multichannel audio with accompanying picture [itu775]. The relation between screen size, aspect ratio and loudspeaker positions is discussed. ITU-R BS.1286 guides the testing of audio systems in the presence of an accompanying image [itu1286].

It should be applied in conjunction with recommendations ITU-R BS.1116-1, BS.1284 or BS.1285. The Appendix 1 of ITU-R BS.1286 contains a list of viewing distances and conditions for different image sizes, aspect ratios and image definitions.

ITU-R BT.1359 provides recommendations for the relative timing of television sound and vision [itu1359]. These recommendations are related to inter-modal synchrony issues discussed in section 2.4.3. Finally, Question ITU-R 102/6 requests suggestions for further standardizing the methodologies used for the subjective assessment of audio and video quality, based on the fact that the existing standards do not cover all aspects of audiovisual perception [itu102]. This is discussed in more detail in section 8.4.

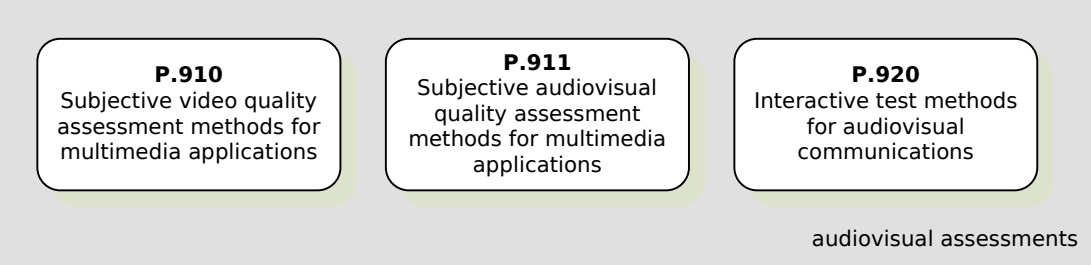


Fig. 8.3: Recommendations of the ITU-T (Telecommunication Standardization Sector) related to audiovisual quality assessments.

The Telecommunication Standardization Sector of the ITU has also come up with a number of recommendations related to audiovisual quality assessments, see fig. 8.3. ITU-R P.910 describes non-interactive subjective assessment methods that can be used to evaluate one-way overall video quality for multimedia applications such as videoconferencing and so on [itu910]. It also suggests characteristics of source sequences to be used, like duration, content, number of sequences etc. ITU-R P.911 is similar to P.910 but applies to audiovisual (instead of visual only) subjective assessments [itu911]. Both are valid for non-interactive aspects. When interactive aspects are to be examined, ITU-R P.920 provides recommendations for quantifying the impact of terminal and communication link performance on audiovisual communications [itu920]. The methodology is based on conversation-opinion tests and is designed to assess the transmission quality.

Other standardization bodies have also come up with recommendations for assessing the perceived quality of audio and visual presentations. The most important ones are the European Broadcasting Union (EBU), the American National Standards Institute (ANSI), the Audio Engineering Society (AES), and the International Electrotechnical Commission (IEC).

EBU Tech 3276 considers in detail the technical specification of listening spaces like reference listening rooms and high-quality sound control rooms [ebu3276, ebu3276a]. EBU Tech 3286 details the methods recommended by the EBU for the assessment of quality of sound program material. It was mainly developed for the evaluation of quality of classical music and includes methods and parameters [ebu3286]. EBU Tech 3286:Supplement 1 extends these recommendations to multi-channel material [ebu3286a].

The AES recommendation AES20 is a set of best practices for the subjective evaluation of loudspeaker systems that should, according to the AES, be performed in addition to objective measurements for high performance loudspeaker systems [aes20].

Two interesting related ANSI standards exist. They define the maximum permissible ambient noise level for rooms in which audiometric tests are to be conducted [ansi31], and

provide a guide for measuring the intelligibility of speech over communications systems [ansi32].

Finally, also the IEC provides recommendations for establishing, conducting and evaluating listening tests on loudspeakers [iec60268].

8.3 Fitness of Unimodal Assessment Rules for Bimodal Assessments

There have been a number of recommendations issued by the ITU and other bodies on subjective quality assessment of audio or visual systems respectively, see section 8.2. These recommendations are widely accepted and have proved to render consistent and comparable test results if applied correctly. It is therefore evident and practical to look into these documents to develop assessment rules for bimodal evaluations, departing from the unimodal case.

For the assessment of subjective quality of audio or visual presentations in the presence of the other modality, similar or even the same methods, rules and attributes as in corresponding unimodal assessments might be used under certain circumstances. This is possible whenever the subjective perception of the one modality is to be judged under continuous and invariable impact of the other modality. Of course, these situations are rather academic examples produced in the laboratory. Yet, for assessments which serve to understand the human perceptual processes from a generalized point of view, it is desirable to keep the number of variables involved to a minimum, see also section 8.4.

For truly bimodal assessments (in which both auditory and visual stimuli are varied), it is usually more difficult to apply the recommendations developed for unimodal assessments. Attributes to be rated must be chosen to represent both modalities. Detection thresholds are usually different and dynamically changing for the bimodal case, such that the rating scales might have to be modified. Feedback from the assessors needs to be collected in a way that does not interrupt the audiovisual perceptual flow (see section 6.3), a problem that can be largely ignored in the unimodal case.

It is therefore not possible to generalize statements on the applicability of certain recommendations developed for unimodal assessments to the bimodal case. Rather, common sense and the experimenter's experience should be applied when designing bimodal assessments, until recommendations truly developed for the bi- or multi-modal case exist. The ITU has issued a call to develop such recommendations for the audiovisual case back in 1999, but no generally applicable suggestions have been issued since then - see section 8.4 which explains some of the limitations of the ITU's approach.

8.4 Suggested Categorization of Assessments

In its "Question ITU-R 102/6" [itu102] dated from 1999, the study group 6 of the International Telecommunication Union (ITU) has suggested a subdivision of applications with regard to the supposed amount of interaction between the modalities. They divide applications into three groups:

- audiovisual presentations
- visual presentation in the presence of audio
  (audio presentation at a constant quality level)

- audio presentation in the presence of visual
  (visual presentation at a constant quality level)

This classification is related to specific products and tasks that these products are supposed to perform. It does not consider that the underlying perceptual processes are also influenced by a number of factors unmentioned by the ITU. These include e.g. the degree of interaction with the user that an application or device offers. Also, the background of the target group supposed to use an application or device might have to be considered to achieve valid quality acceptance data. This assumption is closely related to Neisser's model of the Perceptual Cycle described in section 2.5.2.

Therefore the author suggests to further and better differentiate between those assessments which serve to

1. understand the human perceptual processes,
2. judge the quality and the degree of optimization of a certain audiovisual application system,
3. evaluate the appropriateness of audiovisual application systems for a certain task or compare a number of these systems with each other,

see fig. 8.4.

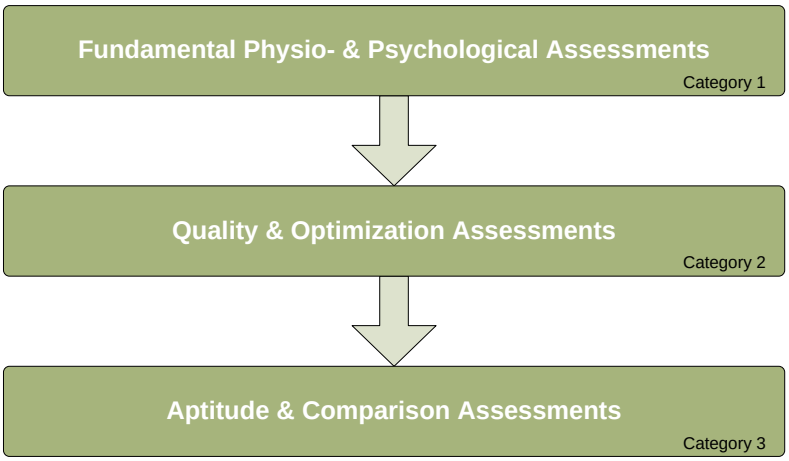


Fig. 8.4: Suggested categorization of assessments according to their objects.

The latter two categories are of high interest to manufacturers of consumer electronics products and consumers, because they allow to compare the subjective quality of products and concepts. The first category is fundamental to purposefully improve the products themselves by generating basic knowledge about the human perceptual system. All three kinds of assessments are necessary for the advancement of (interactive) audiovisual applications. The cited ITU's subdivision could be applied without further limitations to the third group of assessments. For the first group, the subdivision does not seem to be viable. For the second group, the ITU's subdivision might reflect the necessary intermediate stages in the process of quality evaluation of the system under assessment. Only when keeping one modality at a constant quality level, the effects of more or less a detailed simulation (or increase / degradation of quality, respectively) of the other can be consistently judged.

8.5 Discussion of Test Methods, Procedures and Assessment Scales

Today's assessments rely on a combination of subjective and introspective methods, e.g. with questionnaires and rating scales. These methods imply that the test subject needs to be questioned during or immediately after the presentation of an item, an act which in itself reduces the degree of presence or immersion. Because it is necessary for the test subject to reflect upon the perceived impressions before rating an experience, he is induced to consume the content in a much more conscious way, and the effects to be assessed might not occur any more. New approaches try to assess the degree of presence of test subjects using physical, physiological or behavioral investigations, but these are only starting to give relevant data [mee02]. It is therefore still necessary to ask the test subjects directly.

The most common test methods of subjective assessments can be categorized into single-stimulus (also called Absolute Category Rating, ACR) methods, pair comparison methods and multi-stimulus methods.

With single-stimulus/ACR methods, the test subjects are presented with a single audio-visual test situation or scene, the quality or properties of which are to be rated according to previously defined attributes. A test session consists of a series of so-called test items, with one item coinciding with a scene to be rated. These are presented in a randomized order, often using approaches like Latin-square order for statistical test design[1]. This way, so-called sequence effects (biases or errors due to a certain order of presentation) can be excluded. After or even during the presentation of an item the test subject gives its rating.

Pair comparison methods present pairs of items to the test subject, and these are rated in relation to each other. As with single-stimulus methods, the order in which pairs of items are presented is random. An advantage over single-stimulus methods is that pair comparison methods generally render results of higher precision and better stability for non-expert subjects. According to ITU-R BT.500 [itu500], the full scale of quality possible (*best* to *worst*) should be used, which results in test sessions being considerably more time consuming than single-stimulus methods. In order to test the reliability and continuity of test subjects, hidden reference items can be used.

Multi-stimulus methods compare a number of test items, and ratings for each item are given in relation to the other items presented in that test round or trial. For audiovisual tests, these items are always presented after each other. ITU-R BS.1116 [itu1116] describes a double-blind triple-stimulus with hidden reference method used for assessments of audio only material of only small quality impairments. Similar approaches can be used for assessments with bimodal perception, if quality differences of items are small. For bigger auditory quality differences, ITU-R BS.1534 [itu1534] describes a multi-stimulus test with reference, hidden reference and anchor (*MUSHRA*). An anchor is used to compare the quality of an item to a reference quality level. It is important that the test subject can select arbitrarily between the different items to make direct comparisons between them. Fig. 8.5 shows a schematic example for an assessment according to the *MUSHRA* method. With the buttons REF to G the test subject can select the item to be played. The markers on a continuous scale from 0-100 represent the quality rating given by the subject. Given that the REF item is of the highest possible quality, it can be assumed that item B is the hidden reference identical to the REF item.

[1] A Latin square is a square matrix of n rows and columns; cells contain n different symbols so arranged that no symbol occurs more than once in any row or column

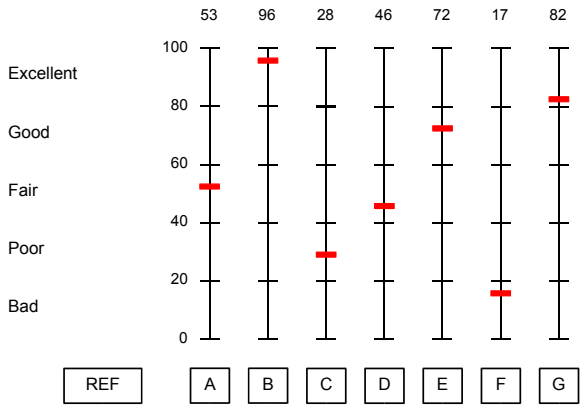


Fig. 8.5: A schematic example for an assessment according to the *MUSHRA* method. Note that item *B* appears to be the hidden reference, and that the test subject did not follow the instructions given in ITU-R BS.1534 properly.

In order to achieve correct and meaningful test results, the scale which is used to rate the quality of items is of decisive importance. Today, the original ITU-R 5-point quality and impairment scale is rarely used, because the number of ratings or marks to choose from is often too small and the differences between semantic designations like *excellent* vs. *bad* or *imperceptible* vs. *very annoying* are too big.

Therefore, the rating scales should be continuous for assessments related to human bimodal perception. Scales usually cover a range from $0-100$ and should have a length of at least $100mm$ on paper or as electronic input devices (faders), to provide test subjects with the possibility of rating very subtle differences. The use of an electronic Input Device as described in section 6.3 is highly recommended for bimodal assessments. It not only reduces the risk of errors introduced by false reading of marks on a paper scale questionnaire, but also reduces the problem of split attentiveness in the moment when the subject gives its rating of an item.

Another interesting method for the assessment of interactive audiovisual application systems is the single-stimulus performance method. The main idea is to have the quality of a system judged indirectly by examining how fast a test subject can fulfill a certain task. Examples for this are the identification of objects in visual or audiovisual presentations, or the exact determination of the direction of incidence of sound in an audiovisual system.

8.5.1 Qualitative vs. Quantitative Assessments

Coolican gives a definition of the terms *quantitative data* and *qualitative data*. Qualitative data is data "left in its original form of meaning (e.g. speech, text) and not quantified", whereas quantitative data is data "in numerical form, the results of measurement" [coo04]. As such, qualitative assessments are assessments that allow test subjects to use their own words for describing the percepts, whereas quantitative assessments work with predetermined attributes that are quantified by subjects using a given rating scale.

Usually, quantitative assessment methods are preferred, because the resulting data is easier to analyze. Unfortunately, these methods often require some kind of preparation of the test subjects. This can either consist in a lengthy and repeated vocabulary development for all kinds of descriptive analysis (DA) processes (see e.g. [bec06]), or in a training that

familiarizes test subjects with the semantic identifiers of the attributes to be rated. Other methods, like the repertory grid techniques (RGT) as revised for audio only tests by Berg and Rumsey [ber99, ber03], often require extensive and time consuming stimulus selection and separation processes. Still, quantitative methods are far more often used in assessments of audio and visual quality than qualitative methods.

According to Creswell [cre06], four basic approaches for combining quantitative and qualitative data exist:

1. *Triangulation* is based on the analysis of two independent sets of data. In a final step the results are merged. When one initial set of data is quantitative and the other is qualitative, the latter needs to be transformed into quantitative data before merging. Thus, a quantitative comparison can be performed.

2. In the *Embedded Design* approach, either the quantitative or the qualitative data collection is embedded into the other. This might be sensible when additional qualitative results are needed in an otherwise purely quantitative study.

3. The *Explanatory Design* is a two-step approach: first, a quantitative assessment is performed which is later complemented with a qualitative study, e.g. to explain unexpected results.

4. The *Exploratory Design* approach is another two-step scheme, but here the qualitative assessment is performed first. Later, a quantitative study is added, e.g. to examine the significance of factors developed in the qualitative part.

Unfortunately, the classification of these approaches is not very clear. It is often hard to say which approach has actually been followed.

In the field of audiovisual research, the application of combined methods is rarely found until today. Jumisko-Pyykkö et al. [jum07] have developed experienced quality factors on the basis of quantitative and qualitative methods, but a terminal combination of both data sets obtained in the experiment is still missing.

Section 8.9 describes an experiment in which quantitative and qualitative assessments have been combined to obtain data that goes beyond the limits of quantitative data. Section 8.13 outlines a mixed method approach that is based on a Free Choice Profiling (FCP) method and thus supersedes training of semantic identifiers.

8.5.2 Response Format and Bias

The response format defines the way in which feedback about the item under assessment is collected from the test subject. In order to collect data that can be analyzed scientifically, a number of basic requirements have to be fulfilled. According to Bech and Zacharov [bec06] these are: *objectivity* of the statements on the items, *quantification* to allow a statistical analysis of the resulting data[2], *communication* enabling an exchange of descriptions in a clearly specified format, and *scientific generalization*.

The two main scaling procedures that are usually applied are the direct and the indirect scaling. Using direct scaling techniques, test subjects are asked to directly quantify

[2] *Quantification* is actually not necessarily related to performing a *quantitative assessment*. There are methods that allow to use qualitative data as the basis for a quantitative analysis, see e.g. section 8.9. In these cases, a transformation (or quantification) of the qualitative data has to be performed.

a perceived sensation (convert it into a sensory magnitude) and to report it using a scale. These are often used for perceptual assessments. In contrast, indirect scaling techniques assess the subjects' ability to discriminate between stimuli, often applying simple preference decisions in the form of pair comparison tests. They are mainly used for affective measurements.

Scaling is always related to bias effects. These effects occur whenever subjective impressions are rated on a scale, i.e. when applying direct scaling techniques. A typical example for a bias effect is that subjects always reserve a certain portion of the scale (upper and lower parts) to allow for rating of items to come (items yet unknown). Test subjects typically want to be able to rate items of even higher or lower quality, even if these are not part of the assessment. Bech and Zacharov give an overview and a classification of the many bias effects that can occur in audio quality assessments [bec06]. This classification can easily be adopted to the bimodal case. The most prominent bias effects are:

- *contraction bias*: large differences are underestimated, small differences are overestimated
- *bias caused by (un)familiarity with units of magnitude*: most test subjects will not be able to judge the quality of e.g. an autostereoscopic display because they have never seen one before; yet they will be able to detect and rate audio artifacts when they are used to listening to recorded music
- *perceptual sensitivity*: when specific attributes are tested, these should be (physiologically) perceivable by the test subject
- *expectation bias*: the expectation of the subject to perceive certain stimuli can influence the rating; closely related to the context effect

Additionally, also context effects can have an influence on the rating. Their impact is very similar to the bias effect in that subjects tend toward giving higher or lower ratings under certain circumstances. The most prominent example is the sequence effect: depending on the presentation order, a medium quality item tends to get higher ratings when the previous items were of lower quality and vice versa.

8.5.3 Proposal of Quality Attributes for (Interactive) AV Systems

As discussed in section 8.4, the overall perceived quality cannot be split into independently processed perceptual channels for the case of assessments of the first type[3], and only to a limited degree for the second type[4]. Yet, in most test scenarios it is unwise to ask test subjects to quantify the overall quality of an item directly, because the term *overall quality* is a fuzzy one. Very often it is individually interpreted by test subjects, i.e. its meaning strongly depends on a subject's personal background.

One possible solution is to split the "overall perceived quality" into a number of independent attributes which are easier to define and detect. This implies that test subjects are familiar with the attributes and the semantic designators used to identify the attributes. Especially with untrained, "naive" subjects this can lead to substantial problems: Semantic designators used to describe audiovisual attributes may have different meanings for

[3] Experiments that serve to understand the human perceptual processes in general.

[4] Assessments that serve to judge the quality and the degree of optimization of a certain audiovisual application system.

test subjects with different background, see section 8.1. A novel approach to avoid these problems is discussed in section 8.13.

Yet, sometimes it is necessary and sensible to perform perceptual assessments by having subjects rate a number of attributes. In order to define the attributes to be assessed, it is fundamental to find a precise description of the bimodal percepts, and to delimit these percepts from each other. Larsson et al. suggest to differentiate the attributes of bimodal perception into three groups: *temporal*, *spatial* and *qualitative* aspects, see [lar03]. Fairly easy to describe are the ones related to spatial and temporal percepts. Do the locations of perceived visual and auditory sources (e.g. a virtual loudspeaker box emitting sound) coincide in the presentation? Are visual and auditory stimuli in sync?

Qualitative aspects are commonly represented by the "overall impression" of an audiovisual presentation. In scientific publications on audiovisual perception, very often the term "presence" is used to describe this. It comprises the sensation of "being there" of the human spectator, or in other words the "perceptual illusion of nonmediation", as Larsson et al. call it [lar03]. These attributes are strongly connected with human emotions.

Attributes which relate to qualitative aspects are of much higher complexity than temporal or spatial ones. Such attributes strongly depend on the application itself. Woszczyk et al. [wos95] suggest to divide these attributes into four *dimensions of perception*:

- *Motion*: subjects evaluate the illusion of physical flow and movement
- *Mood*: subjects evaluate the articulation and density of atmosphere
- *Space*: subjects evaluate the illusion of being in a projected space
- *Action*: subjects evaluate the sensation of dynamic intensity and power.

Each dimension itself is subdivided into four attributes, which include auditory as well as visual aspects:

- *Quality* (distinctness, clarity, and detail of impression): How distinct is the sensation?
- *Magnitude* (the strength of impression): How powerful is the sensation?
- *Involvement* (the emotional effect on the viewer): How involving is the sensation?
- *Balance* (relative contribution of auditory and visual sensations): How balanced are modalities - stronger sound or picture?

Using these 16 attributes (4×4), test subjects can judge sensitive and cognitive impressions and compare them to real world experiences fairly well. Therefore, these attributes allow to evaluate *emotional* parameters (as opposed to *technical* parameters), which play an important role in generating the composite sensory image.

It is very time consuming and tiring for test subjects to judge all these attributes at once, so a selection should be made before performing the assessments. Bech and Hansen have used these attributes successfully for the evaluation of audiovisual interaction in home theater systems [bec95]. They have found that the dimensions *Space* and *Action* are the most sensitive and therefore contribute the most to the perceived overall quality in their assessment. More importantly, the attributes should be selected according to the content of the test material, because attributes to be judged need to be sufficiently present in the test items. For assessments of categories two and three (see fig. 8.4), test material should be representative for real-life applications of the system under assessment.

8.5.4 Experimental Design Considerations and Test Material

Section 8.5.3 has discussed how to account for crossmodal interaction happening in bimodal perception, and which attributes to evaluate in order to assess these effects. Yet, if an (interactive) audiovisual system is to be optimized for highest performance and impact, it is necessary to find out which technical parameters are responsible for generating these impressions.

For an interactive audiovisual application in which the user can navigate through a room and interact with various sound emitting objects, it is for example interesting to find out which degree of accuracy is necessary for the room acoustic simulation when a certain degree of "visual realism" is given. Technical parameters to assess are e.g. the maximum order or total number of image sources to be used in the simulation, the number of loudspeakers necessary for good localization quality in the audiovisual case, the density of the echoes produced by a generic reverberation algorithm, etc. This question brings back the questions stated in the introduction of this work: can the accuracy of the acoustic simulation be reduced without a perceived loss of overall quality when the presentation is a bimodal (audiovisual) one, compared to the audio-only case?

Of course, the experimental design should be focused on the hypotheses (the scientific questions) that the experiment is supposed to evaluate. This means that the structure of the assessment is organized in such a way that answers to the questions under assessment can actually be derived from the experimental result. This sounds trivial in the first place, but because interactive audiovisual assessments are usually based on introspective methods (see section 8.5), careful balancing of questions, rating scales and presence of attributes can be helpful. The latter implies that the test material used should actually allow the evaluation of the attributes under assessment. This means that a sufficient number of appearances as well as amount of quality variation of these attributes must be given in the stimulus material.

Finally, depending on the category of assessment to be performed (see section 8.4), the hypothetical application simulated in the assessment (the "meta environment") possibly should resemble the target application. The target application is the application in which the result of the evaluation is going to be used. Therefore, environmental surrounding, technical setup, and ideally experience and state of mind of test subjects might be similar to what can be expected for the target application. Whether these preconditions actually influence the validity of results is still under discussion. Yet, the multitude of results from isolated cognitive laboratory experiments difficult to generalize or apply in specific applications indicates just this.

8.5.5 Test Subjects

In subjective assessments, the population of interest is represented by a selected population. The selected population is a group of subjects that have been selected randomly or according to certain criteria that are representative for the population of interest. For this thesis, most subjects that participated in the assessments were students or staff of Technische Universität Ilmenau. This theoretically confines the results obtained in this thesis to the given population. The results might not be valid for a generalized population. On the other hand, the type of application (interactive audiovisual application) used in the assessments is very likely to be adopted first by a population that is very similar to the group of subjects that participated here.

Generally speaking, depending on the category of assessment (see fig. 8.4), different groups of test subjects need to be asked. For the first category (fundamental physiological and psychological assessments), experts as well as naive assessors should take part in the assessments. Here it is important to acquire test data from a large number of test subjects in order to have a reliable basis. For the second and third category (quality and optimization assessments, aptitude and comparison assessments), we may expect an effect known from subjective audio only assessments: a smaller number of expert listeners render the same or even more reliable results as a large number of non-expert listeners.

8.5.6 Test Room - Laboratory Characteristics

All assessments described in this thesis took place at the Listening Lab of the Institute of Media Technology at Technische Universität Ilmenau. The Listening Lab is compliant with ITU-R BS.1116-1 [itu1116] and EBU 3276 [ebu3276]. It is further conforming to the GK15 and partially the GK10[5] limiting curves for noise levels in studio environments [irt95]. The Listening Lab has a total area of $120m^2$ and the dimensions of the listening area are $8.4m \times 7.6m \times 2.8m$. The reverberation time is $T_{60} = 0.34s$. It is equipped with a control room from which the assessments were conducted, and a sound-proof machine room where most of the equipment used in the assessments is installed.

Located in the Listening Lab is an acoustically transparent projecting screen. It can be lowered from the ceiling and has $2.72m$ of width providing the 4/3 format. Fig. 8.6 shows the geometry of the loudspeaker and screen setup. As can be seen, the large projecting screen roughly covers half of the overlapping binocular visual field[6], an angular area of around 50°. This coincides with the angular area of the larger screen as suggested in ITU-R BS.775 [itu775].

The influence of the screen upon the frequency response of the audio reproduction setup was measured. The measurement was done using the quasi-anechoic MLS (maximum length sequence) speaker test of the Audio Precision APWin v2.24 software with a System Two Cascade hardware and a Microtech Gefell 1/4" measurement microphone. Fig. 8.7 shows the two curves with (red) and without (blue) the projecting screen in front of the loudspeaker. The difference Δ between the two measurements is the frequency transmission curve of the screen. The $6dB/octave$ decrease in amplitude for lower frequencies is related to the narrow time window that was used to make the measurement quasi-anechoic. The distance between loudspeaker and microphone was $0.8m$ with the optional screen located in the middle at $0.4m$.

The loudspeaker setup has been derived from a regularly-spaced 12-channel setup by leaving out loudspeakers positioned at ±75° and ±135°. The reduction of loudspeaker count was due to practical software implementation reasons, as most (even professional) sound cards provide drivers that can include no more than eight output channels in one device. Also, the routing and transmission of signals using the ADAT "Lightpipe" protocol suggests the use of eight channels[7]. All loudspeakers used in the assessments described in

[5] GK10 and GK15 specify a background noise level of $10dB(A)$ and $15dB(A)$, respectively. The GK10 is only met when the projector is switched off. For these assessments - which were audiovisual tests requiring visual display provided by the projector - background noise levels met the specifications of the GK15.

[6] For a discussion of the human visual field see section 2.1.2.

[7] The ADAT optical interface can transmit eight digital channels of audio of up to 48kHz sampling frequency over one dedicated optical cable using the TOS connectors known from the S/PDIF optical

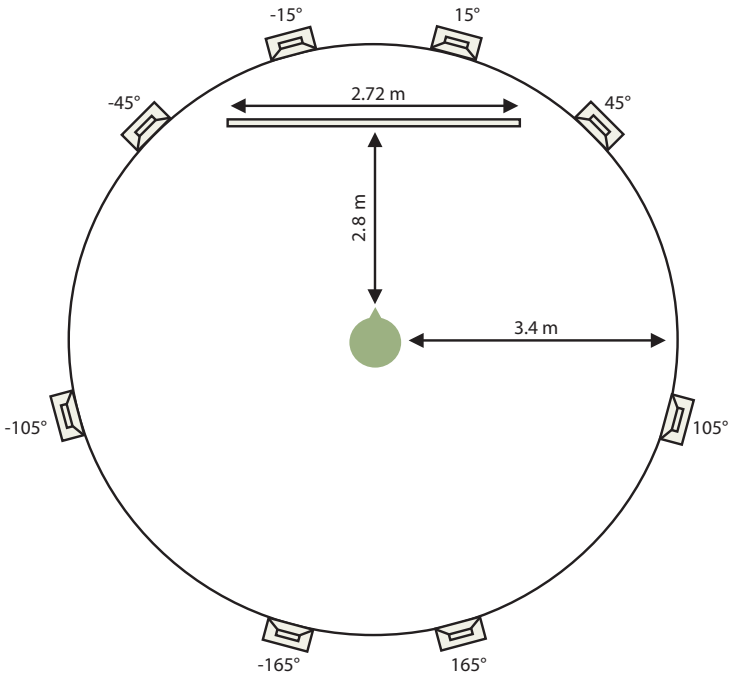


Fig. 8.6: The geometry of the loudspeaker and projecting screen setup in the Listening Lab.

this thesis were identical full-range, active loudspeakers of type Genelec 1030APM.

The somewhat unconventional distribution of the loudspeakers between front and back theoretically provides a number of benefits. Firstly, the localization is higher in the front, where the agreement between *heard* and *seen* sound source locations can be compared most easily. Secondly, there is no gap in the back, so a rotational movement will not render big changes in perceived sound quality related to phantom sources. Whether these assumptions hold true and the benefits are actually perceived was evaluated in an assessment described in section 8.7.

Among others, de Bruijn and Boone have shown that human sound localization accuracy in the vertical plane is limited [bru03]. They have found that differences in elevation between object positions in the visual and auditory domains may become as high as 22° for life-size screen projections without being noticed. Because test subjects in the assessments performed here are located at a distance of $2.8m$ from the screen (see fig. 8.6), this corresponds to a maximum allowed deviation of $1.1m$. Therefore, with a total screen height of $2.04m$ and the test subjects' head roughly located at the height of the center of the screen, elevation discrepancies do not contribute to perceived localization errors.

8.6 Fundamentals of Statistical Analysis

Statistical data obtained in a subjective assessment can be analyzed using methods from both the descriptive and the analytical statistics. Descriptive methods are used to summarize and illustrate test results, whereas analytical methods serve to verify hypotheses usually stated before the assessment.

interface. ADAT was introduced in 1991 by Alesis Corp. as a recording format for storing 8 tracks of digital audio on an S-VHS cassette tape with an accompanying optical transmission protocol (ADAT Lightpipe).

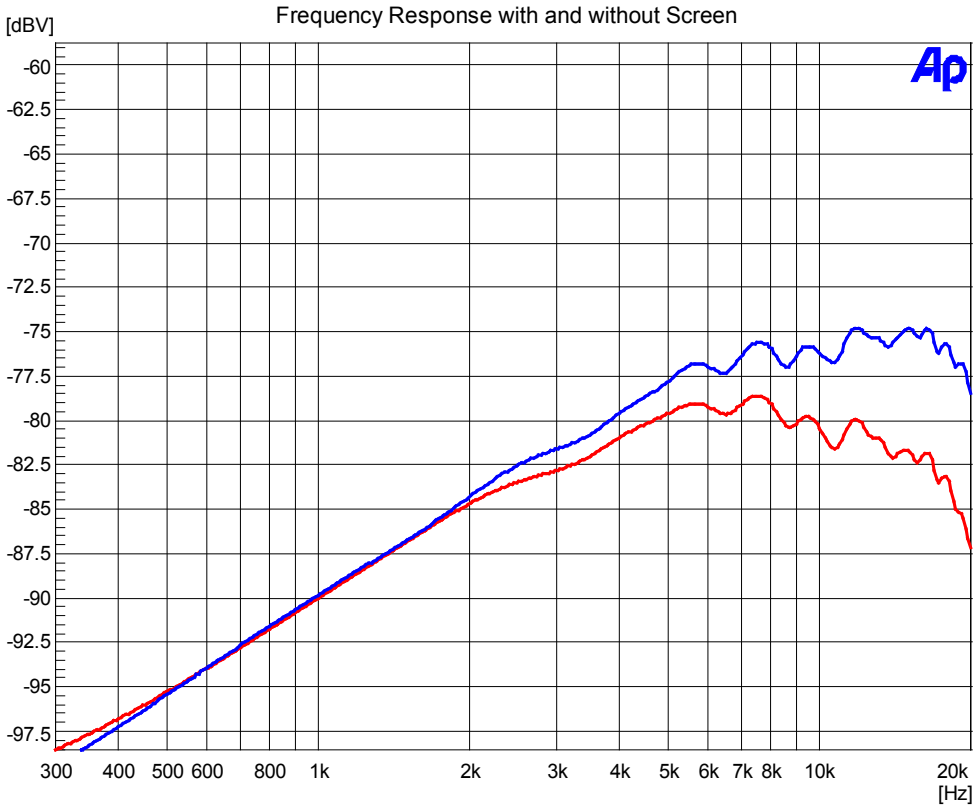


Fig. 8.7: The frequency response of the loudspeakers with (red curve) and without (blue curve) the acoustically transparent screen in front. The $6dB/octave$ decrease in amplitude for lower frequencies is caused by the narrow time window that was used to make the measurement quasi-anechoic.

Coolican suggests to check the validity of a statistical analysis method to be applied on the data collected in the assessment [coo04]. This is usually done performing three steps:

1. *Determination of the level of measurement of a variable.* The scale level (or level of measurement) is a classification that was proposed in order to describe the nature of information contained within numbers assigned to objects and, therefore, within the variable. According to this classification scheme, in statistics the kinds of descriptive statistics and significance tests that are appropriate depend on the scale levels of the variables concerned. Table 8.1 gives an overview of the four levels of measurement as proposed by Stevens in [ste46]:

   Interval scales should use equal intervals, relative proportions on the scale map to physical reality.

2. *Check for normal distribution.* Apart from visualizing the distribution by drawing a histogram, an objective revision of the distribution using the *Kolmogorov-Smirnov test* is performed. This test is also know as the *goodness-of-fit* test.

3. *Separation of dependent and independent samples.* According to Bühl [bue06], two samples are dependent when each value of one sample can be related to exactly one value of the other sample.

Scale level	**Empirical relevance**	**Statistical analysis**	**Variable (example)**
nominal	none	limited (nominal measurement)	"categorical" (gender)
ordinal	order of numbers	limited (nominal measurement, relative measurement)	"rank" (listening experience)
interval	spread of numbers	unlimited (no ratios between arbitrary pairs of numbers)	(quality rating)
ratio	ratio of numbers	unlimited	(age)

Table 8.1: Scale levels sorted by empirical relevance and statistical operations applicable.

8.6.1 Descriptive Statistics

At the descriptive level of analysis an overview of the collected data is obtained. Besides a graphical representation of the data in the form of bar charts / histograms and box plots, frequency tables are frequently computed. These contain the relative or absolute frequency of appearance of the measurements of a variable in an easy to read tabular form. The graphical representation of data usually includes mean values and associated variances of the data. Furthermore, at this level outliers are identified.

In subjective assessments the arithmetic mean $\overline{x}_{jk}$ and the standard deviation s_{jk} are the most important quantities. The *arithmetic mean* $\overline{x}$ is the average of scores found by adding them all and dividing by the total number of scores:

$$\overline{x}_{jk} = \frac{1}{N} \sum_{i=1}^{N} x_{ijk} \tag{8.1}$$

x_{ijk} = score / rating of subject i (=value i) for given condition j and stimulus k
N = sample size (total number of values i)

The main disadvantage of the mean is its sensitivity to extreme values. These give a distorted mean value. Therefore, when the distribution of values is not normal (i.e. it is not following a Gaussian curve), it is sensible to indicate the median of values. The median is the central value of a set. It is calculated by ordering the values (the samples) from smallest to largest, and the value in the middle position (or the mean of the two middle position values for an even-numbered sample size) is the *median*.

The median and the mean are identical if the distribution of values is perfectly symmetric. If the two are not equal, then the distribution is not symmetric and might by further analyzed by examining the *skewness* and *kurtosis* values. They are measures of distribution and indicate the shape of the distribution curve.

The *standard deviation* s is the measure of dispersion. It is the square root of the sum of all squared deviations divided by $N - 1$:

$$s_{jk} = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (x_{ijk} - \bar{x}_{jk})^2} \tag{8.2}$$

The *variance* is defined as the square of the standard deviation. It is another indicator

of dispersion of the data. Again, as with the mean, the standard deviation and variance are distorted by extreme values.

8.6.2 Analytical Statistics

When collecting statistical data from a selected population, one of the basic questions is how much accordance there is between the estimated mean values based on the selected population (the test subjects) and the true mean values of the population of interest? What is the relationship between the true population mean and the estimated sample mean? As Bech and Zacharov put it, this "subsequently translates into concluding whether the experimental variables have a significant influence on the perception reported by the subjects" [bec06], which is ultimately the question that we want to answer with the assessment.

This relationship between the true and estimated mean values[8] is given by the *confidence interval* (CI) of the mean. The CI has the following statistical meaning (modified from [bec06]): If a series of identical subjective assessments were conducted on a single test item using different samples from the same population of subjects and for each sample a mean and a 95% CI were calculated, then in the long run, 95% of these CIs would include the true mean value. The CI is calculated as follows:

$$\left[\bar{x}_{jk} - t(1-\frac{\alpha}{2}; n-1)\frac{s_{jk}}{\sqrt{N}}; \bar{x}_{jk} + t(1-\frac{\alpha}{2}; n-1)\frac{s_{jk}}{\sqrt{N}}\right] \tag{8.3}$$

$t(1-\frac{\alpha}{2}; n-1)$ = the $(1-\frac{\alpha}{2})$-Quantile of the t-distribution with $n-1$ degrees of freedom

α = the significance level equal to (1-(*degree of confidence*))

The *degree of confidence* is defined by the experimenter's need for certainty. It is usually set to 95%. Therefore, the width of the CI depends on the desired reliability of the estimate, the sample size N and the standard deviation s of the sample of observations. Hence, when all other things are equal, a small confidence interval indicates a higher reliability of an estimate than a large CI.

Depending on the scale level, different tests are used to accept or reject the *null hypothesis* H_0. The null hypothesis is the assumption of no effect in the population from which the samples are drawn. This means that if the null hypothesis is true, then "we would expect little difference between the samples (only that resulting from sampling error)" [coo04]. The alternative hypothesis is the assumption that an effect exists, e.g. that the populations differ. The errors committed when mistakenly accepting or rejecting the null hypothesis are called Type I error (falsely rejecting H_0 when it is true) and Type II error (falsely retaining H_0 when it is false).

To calculate the probability of the statistical effect occurring under the null hypothesis H_0, so-called significance testing is done. A significance test determines the *probability* p that the result would occur if H_0 was true. Table 8.2 shows the significance levels generally used and their meaning for acceptance or rejection of the null hypothesis. Fig. 8.8 gives an overview of appropriate two sample tests and makes suggestions on which test to use for what data and aim. For a discussion of these tests see e.g. [coo04, bor05, bec06].

For doing a lot of tests on various aspects of a set of data, each time a null hypothesis and an alpha level of 0.05 must be assumed. This is called *capitalizing on chance* [coo04]

[8] For parametric tests; for non-parametric tests, i.e. for data that cannot be described by a normal distribution, the median value is used instead, see e.g. [bec06, coo04].

Significance level	**Level of probability at which it is agreed to reject H_0. If the probability of the obtained results under H_0 is less than the set level, H_0 is rejected.**
$p \leq 0.1$ (10%)	Significance level generally considered too high for rejection of H_0. Yet, further investigation might be recommended if p under H_0 is this low for a numerical answer.
$p \leq 0.05$ (5%)	Conventional significance level.
$p \leq 0.01$ (1%)	Significance level preferred for greater confidence than that given by the conventional one. Should be set where research is controversial or unique.

Table 8.2: Significance levels for numerical answers and their meaning for acceptance or rejection of the null hypothesis. *After [coo04].*

and regarded as bad praxis, because the probability of making a Type I error is increased. To by-pass the problem one can either use a more severe level of alpha or use significance tests that are specifically designed for testing among more than two conditions. The most common statistical technique that compares variances within and between samples is the *Analysis of Variance* (ANOVA). It gives an estimate of the significance of differences between a set of means, which makes necessary three assumptions on the data [coo04]:

- The variance is homogeneous.
- The data is on an interval scale.
- The sampling distribution of means is normal. This is assumed by checking if the dependent variable is normally distributed.

ANOVA basically compares the variance between groups with the variance within groups. It is assumed that when the variance within groups is low and the variance between groups is high, then the difference between groups is important and not based on chance. This is called the F ratio statistic:

$$F = \frac{between\ groups\ variance}{within\ groups\ variance} \tag{8.4}$$

As will be seen in the subsequent sections of this thesis, data obtained in audiovisual quality assessments very often does not fulfill all of these criteria, so ANOVA can rarely be applied. A non-parametric alternative that can be used when the data is related and does not satisfy the assumptions for an F test is the *Friedman* test. It is a non-parametric rank test for significant differences between two or more related samples (a non-parametric equivalent of a repeated measures ANOVA). Data must have at least ordinal level of data measurement.

Another concept regularly tested in statistical analyzes is *correlation* between data. The relationship between two variables or the degree to which two scores match each other can be checked using tests of correlation. For data given at least at the interval level, usually *Pearson's* (r) product moment correlation is determined. It is based on the variance in two sets of scores. Whenever large deviations are paired with large deviations (or small deviations with small deviations) the resulting r is high, indicating that the two variables co-vary.

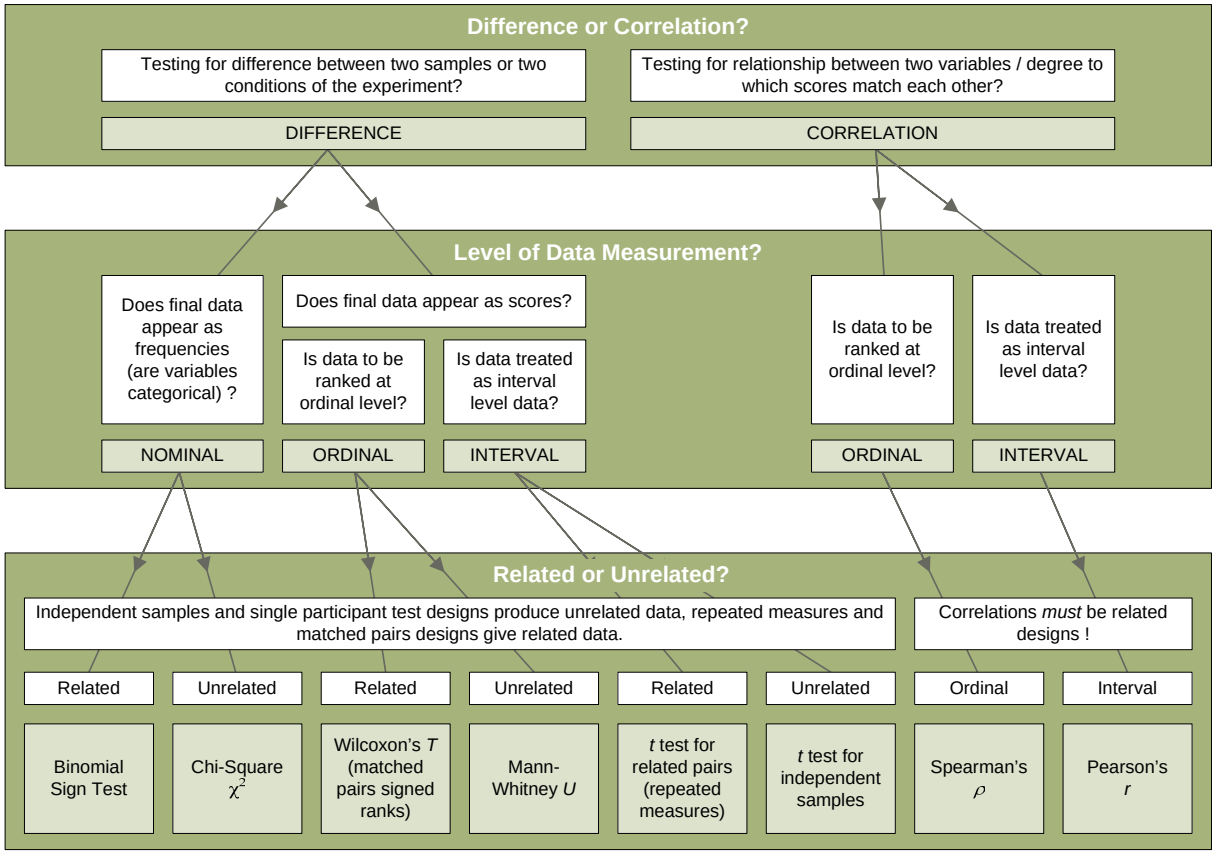


Fig. 8.8: Three questions helping to choose an appropriate two sample test, *after [coo04], modified.*

8.7 Assessment: Optimum Number of Loudspeakers

In interactive, object-based audiovisual applications the number of loudspeakers to be used for the auditory reproduction of content is rather critical. On the one hand, a higher number of loudspeakers allows a generally higher localization quality of sound events. On the other hand, each channel of audio adds to the computational costs of the application. Unlike in cinema or home cinema systems, where the content is contained in a number of audio tracks which are attached to one loudspeaker each, in object-based systems the auditory output needs to be rendered according to the reproduction setup on-the-fly. This is also due to the interactive nature of such systems - the user is often allowed to navigate freely through the scene. Therefore no pre-recorded audio tracks can be used to e.g. generate a convincing room acoustic impression, as this varies with the user's virtual position inside the scene.

This section evaluates the question of optimum number of loudspeaker channels for audiovisual application systems in which visual content is displayed on a large projecting screen. Subjective assessments have been performed comparing three different loudspeaker setups (four, five and eight channels) while varying the real-time room simulation algorithms applied.

To determine a generally valid optimum, a number of constraints need to be looked at. MPEG-4 audio allows for various degrees and types of room simulations to be included in the content. Whereas the *spatialize* flag of a sound source indicates whether it shall

be panned according to listener position and location of that sound source[9], the *auralize* flag determines a room effect to be rendered upon the sound source. This room effect can either be based on the geometric shape and absorption of the materials used for the walls (*physical approach*), or on an abstract semantic description of the room's acoustic characteristics (*Perceptual Approach*), see section 5.4.

Therefore, three typical complexities of room acoustic simulation processes were included in the assessment:

1. *Dry Audio*: The source is rendered audible without any reverberation, but is panned according to the locations of listener and source.

2. *Physical Approach*: The source is rendered audible using a room acoustic real time simulation based on the geometry of the room and the absorptive characteristics of the walls. For the simulation an algorithm based on the image source method for generating the early reflections was used as described in section 5.7. The diffuse reverberation part was generated using a nested all pass filter network as introduced by Gardner [gar92, gar98], see section 3.2.5. Panning is done according to the locations of listener, original source, and image sources.

3. *Perceptual Approach*: The source is rendered audible by a generic reverberation algorithm based on an abstract description of the room's acoustic characteristics. The Perceptual Approach is detailed in section 5.6. Again, panning is done according to the locations of listener and source.

8.7.1 Test Setup

As the audio rendering engine TANGA used in the *I3D* MPEG-4 player needs to know the number of loudspeaker channels to be used for the audio processing upon initialization, an arbitrary switching between different loudspeaker setups during the assessment was not possible. Therefore test subjects were not allowed to navigate actively in the scene during the assessment. Instead, they were presented with four different pre-recorded paths across the virtual room. Test subjects were able to switch between these loudspeaker setups while the pre-recorded paths were presented. These paths were chosen to cover a wide range of relative sound source positions, such that the source was heard in front, at the sides and from behind the test subject during the assessment. This was necessary because obviously the loudspeaker setups under assessment are not suited equally well for localizing all sound source positions. They will therefore give biased results for certain directions of incidence.

Loudspeaker Setups under Assessment The three loudspeaker setups under assessment were the following:

1. The first loudspeaker setup consisted of an eight channel setup of loudspeakers located on a circle of $6.8m$ of diameter. Loudspeakers were located at angles of 15°, 45°, 105°, and 165° from the front to both sides of the test subject, see fig. 8.9 left.

2. The second setup under assessment consisted of a "classic" five channel loudspeaker setup according to ITU-R BS.775 [itu775]. Fig. 8.9 (middle) shows the loudspeaker positions at 0° and at 30° and 105° to both sides of the test subject.

[9]Hence, a sound source with the *spatialize* flag not set does not have any specific location.

3. The third setup consisted of four loudspeakers located at 30° and 105° to both sides of the test subject, see fig. 8.9 (right side). This stripped down setup is often found in multimedia applications and (home) computer games.

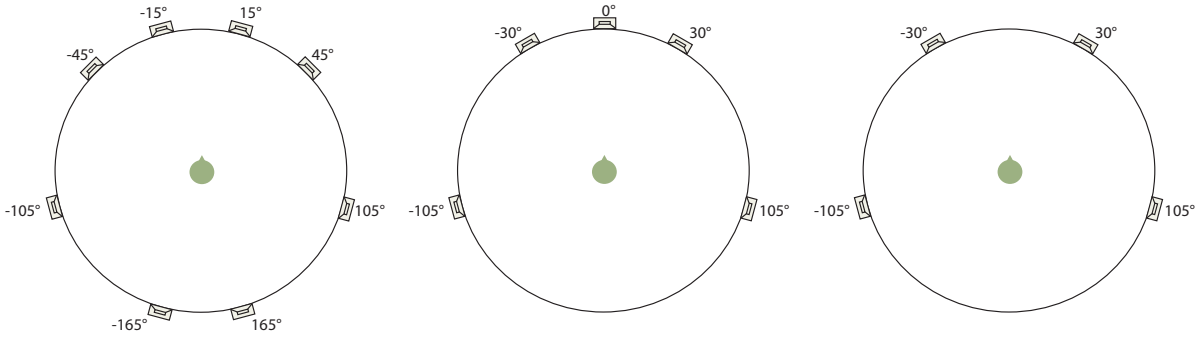


Fig. 8.9: The three loudspeaker setups under assessment, consisting of eight channels unevenly distributed between front and back (left), five channels surround setup as defined in ITU-R BS.775 (middle), and four channels as found in many computer game setups at home (right).

The sound pressure level within each scene varied depending on the virtual distance between the loudspeaker in the center of the gym and the position of the test subject in the scene. A max. SPL of 80dB(A) was measured.

Virtual Room and Navigation Paths For the assessment, a virtual room was searched after that allowed for rather unhindered, continuous movement over a prolonged period of time inside the scene. A virtual gym was created, fulfilling the requirement while at the same time being familiar to most of the test subjects. Fig. 8.10 shows a screenshot of the virtual room used in the assessment. As can be seen, an omni-directional sound source was located in the center of the gym. Typical objects for a gym (a balance beam, two goals, a ball, etc.) were added, which provided some orientation guide to the test subjects while "being moved" through the scene.

Fig. 8.11 shows the four navigation paths presented to the test subjects during the assessment. It can be seen that a number of exemplary movements have been recorded which cover incidence of direct sound from the front, from the sides and from the back of the test subject.

The projecting screen used was an acoustically transparent 4/3 format projecting screen of $2.72m$ of width, see section 8.5.6. Onto the screen a picture with PAL resolution was projected for the assessment. The conversion from the computer's VGA output to PAL format was done using a high-quality scan converter. The signal was then recorded on a video hard disk recorder together with a time code generated on an external time code generator. At the same time, the 5-channel audio signals were recorded on a hard disk recording system synchronized to the same time code.

As the navigation path was pre-defined, the three test items[10] to be compared in the assessment shared the same visual information. Therefore, the same video recording was used for all three items. Only the 8-channel audio and the 4-channel audio signals were subsequently recorded on further tracks of the hard disk recording system and aligned to be in sync with the already recorded 5-channel signals.

[10] eight, five and four channels of audio reproduced on the respective loudspeaker setup

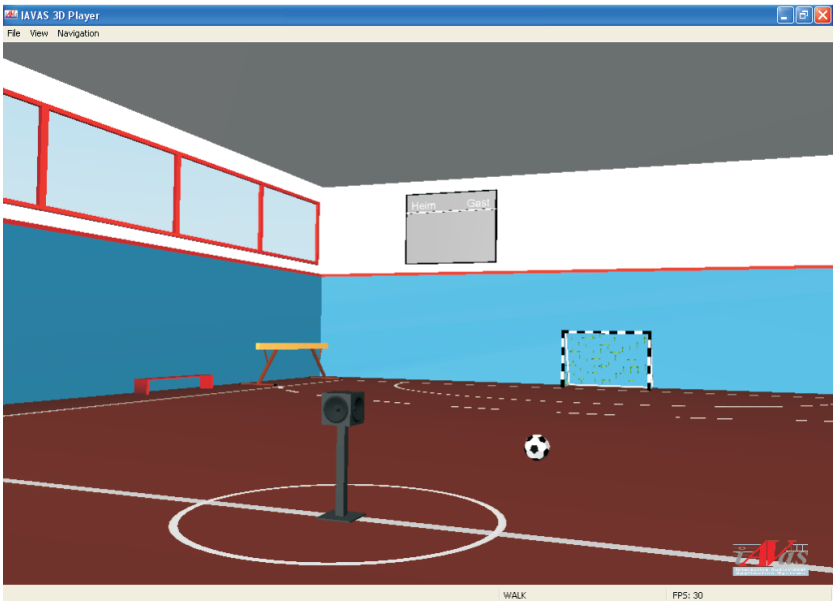


Fig. 8.10: The virtual room used for the assessment. Note the omni-directional loudspeaker in the center of the gym representing the sound source.

Synchronization of Audio and Video Upon playback, video and audio disk recorders were synced via time code. The audio disk recorder played back all 17 previously recorded tracks at the same time. These were fed separately into a digital mixing console. For each test item a separate 'scene' was created in the mixer, which was selected automatically whenever the test subject switched between the items. The mixer 'scenes' contained all necessary routing and gain information for each item. A transition time of $2ms$ was chosen to avoid audible crackles when switching between the scenes.

After aligning the items in the hard disk recording system and doing a test run of the assessment, it was noticed that in fast movements there was a time offset between auditory and visual movement. This was due to the varying complexity of the underlying real time room simulation algorithms used. Apparently, the system latency for higher numbers of loudspeaker channels was generally a bit higher than for lower output channel counts. The simulation for the 8-channel setup produced higher latencies than the one for the 5-channel or 4-channel setup. There was no immediate cure for this behavior of the audio engine at the time of the experiment. Therefore some of the items had to be slightly shifted on the time line of the audio hard disk recording system. After this operation, auditory and visual movement were in sync again, but upon switching between the items, time shifts (jumps) of up to $0.5s$ were present in the audio signal. This aggravated the test subjects' task of comparing the items' quality.

As the five channel loudspeaker setup was to be used both as a reference and hidden reference item, switching between the two did not produce any jumps, whereas switching between the reference and the other two items (eight channel and four channel loudspeaker setup) produced audible jumps. A second set of five channel audio recordings was produced with an accidental delay of up to $0.5s$ inserted into the audio simulation path to prevent test subjects from identifying the hidden reference items by listening to the jumps when switching between items. This resulted in a total number of 22 audio tracks played simultaneously in the experiment.

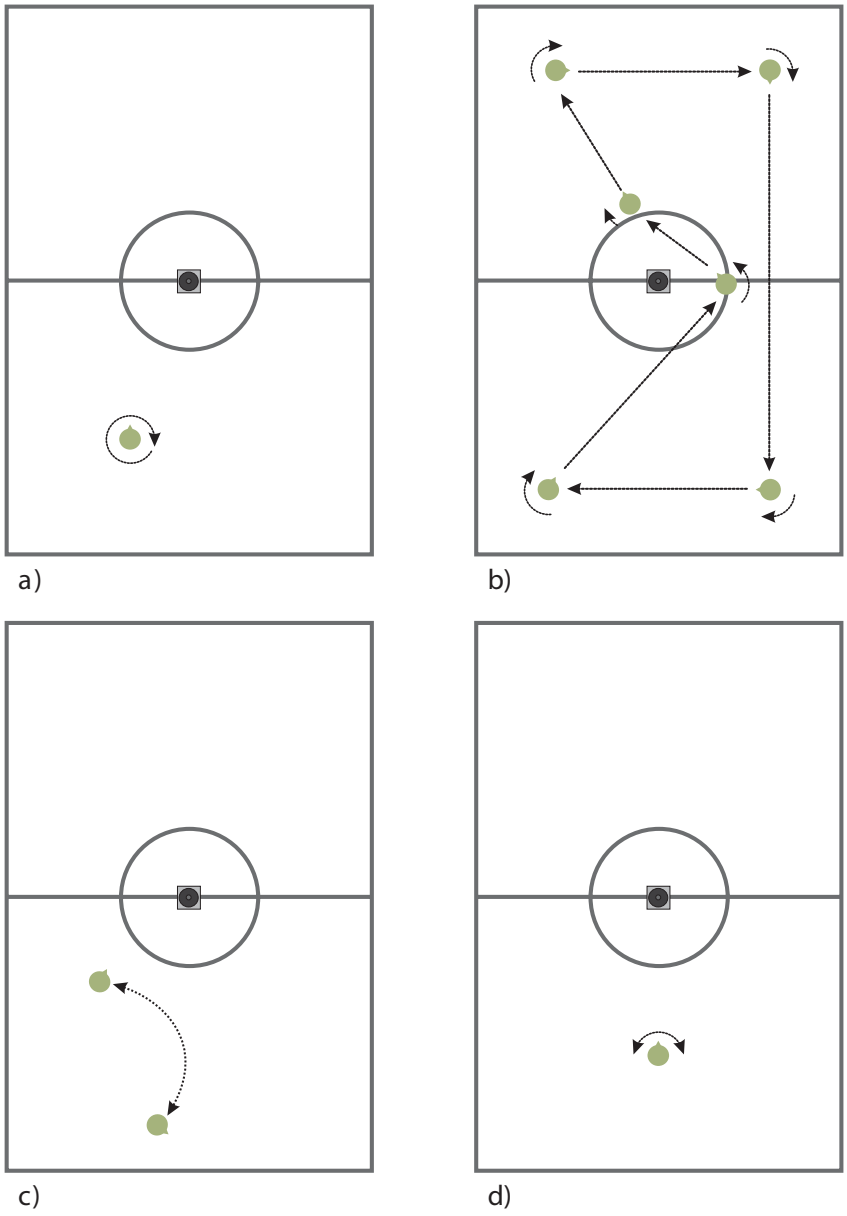


Fig. 8.11: The navigation paths the test subjects were presented with: a) a 360° turnaround, b) a walking path across the gym, c) a translation path with slight rotation of the head, d) turning the head to the left, then turning to the right. The sound source is located in the center of the gym.

8.7.2 Implementation of the Assessment

Test subjects used the Input Device described in section 6.3 to maneuver through the test and to enter their ratings. The actual configuration of the Input Device for this assessment is shown in fig. 8.12. On the left hand side, three motorized 100mm faders were located. Each fader belonged to one of the test items (4-channel, 5-channel or 8-channel loudspeaker setup) under assessment in each trial. Before the assessment, the allocation of item to fader was drawn randomly, as well as the specification of the order in which the trials were presented. This was done four times, so a total of four different sequences could be chosen from. Before each session, one sequence was randomly determined to be used for each

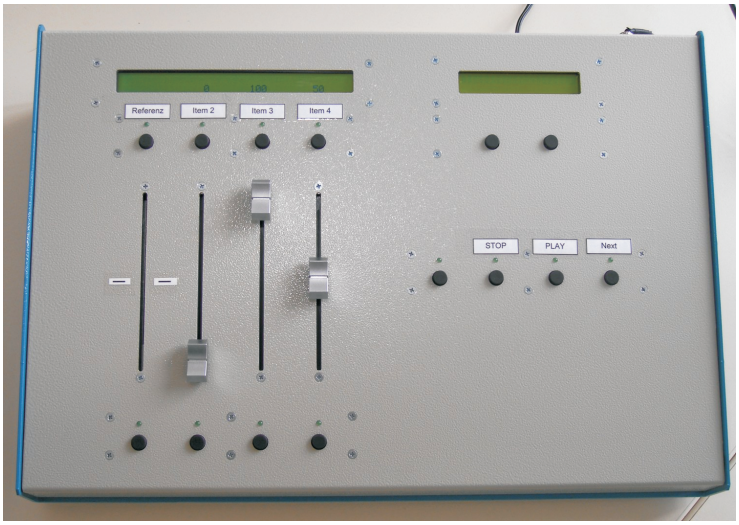


Fig. 8.12: The Input Device configured for the optimum number of loudspeaker assessment.

test subject. As the technical challenge of stepping through the trials while keeping audio and video in sync and logging the test subjects' ratings accordingly was rather high, these sequences could not be created on-the-fly but only before the experiment. Yet, with the four randomly determined sequences used, there is no indication of any sequence effects visible in the obtained data.

Above each fader, a button was located with which test subjects could select the item to be played. Each button was connected to a LED indicating which item the test subject was actually listening to. This helped avoiding a mix-up of the faders in the process of rating the items.

To the left of the three mentioned item select buttons, a fourth button was located with which a reference item could be selected. This represented a 5-channel loudspeaker setup and was also equipped with a status LED. As this was a reference item, which the other items should be compared with, there was no fader located below this button.

Above the four item select buttons, an LC display indicated the actual trial number ("*Trial x of 12.*") in the first line. The rating entered on a scale from $0 - 100$, which corresponded to the fader's position, was displayed in the second line. The display on the right hand side of the Input Device was not used.

Three more buttons on the right allowed the test subjects to start playback of the pre-recorded items, to halt the playback and repeat the trial from start, or to proceed with the next trial after having entered their ratings. Whenever the next trial was selected, the faders moved automatically into the neutral ("50") position.

The assessment consisted of 12 trials with approximately two minutes duration each. Therefore, a minimum total assessment time of about 30 minutes per test subject was calculated. Actually, the mean total time was around 40 to 45 minutes, because test subjects tended to frequently repeat trials that included motion path *b*), see fig. 8.11.

8.7.3 Assessment Task and Rating Scale

The task of the test subjects was to indicate whether the localization of the sound source in the three test items presented in each trial was better, equal or worse than in the reference item. The test subjects did not know that one of the items presented was

always a hidden reference item. They were instructed to use only three fader positions (at "*around* 0 = *worse*", "*around* 50 = *equal*", and "*around* 100 = *better*" than the reference item) for their ratings. Subjects were explicitly told not to rate the agreement between auditory and visual position of the sound source. On the other hand, they were asked not to close their eyes when judging the localization of the sound source. Thus, at least some influence of the visual stimulus could be expected whenever the sound source was located in the front. This is important to note, as the goal of the assessment was to find the optimum number of loudspeaker channels for *audiovisual* applications using large screens.

8.7.4 Analysis and Test Results

Hypotheses The null hypothesis was that the localization quality of the three loudspeaker setups under assessment would not receive different quality ratings when the auditory stimuli used were accompanied by the respective visual stimulus presented on a large screen. The alternative hypothesis was that localization quality would be influenced by the loudspeaker setup used for the presentation.

Test Panel A total of 24 test subjects participated in the assessment. They were between 23 and 34 years of age, all male except one female subject, and two third of them had only little experience in listening to multichannel sound. Eight participants had thorough experience with multichannel sound. Two of the participants can be rated as expert listeners. All of them were accustomed to the visual look and feel of the virtual scene by having played computer games at least once. Half of the participants were frequent computer gamers (first person shooter games). All reported to have normal hearing and normal or corrected to normal vision.

Data Analysis Two different methods of analysis were used to evaluate the collected data. The Friedman test was applied to test several conditions of ordinal related data. The Wilcoxon test was used for the pairwise comparison of ordinal related data, see the overview in fig. 8.8. A significance level of $p = 0.01$ was adopted to avoid the *capitalizing on chance* problem, see section 8.6.2.

Reliability The reliability of the data was evaluated between reference and hidden reference items consisting of the five channel loudspeaker setup. The varying parameters were room simulation algorithm (*Dry Audio*, *Physical Approach*, and *Perceptual Approach*), motion path (see fig. 8.11) and audio content (music vs. speech signal). No differences between the given reference and the evaluated hidden reference for the *Dry Audio* ($Z = -0.775$, $p > 0.05$, ns) and the room acoustic simulation algorithm *Perceptual Approach* ($Z = -0.277$, $p > 0.05$, ns) were reported.

In contrast to this, a significant difference appeared between the reference and hidden reference for the *Physical Approach* room simulation algorithm ($Z = -5.103$, $p < 0.001$). Therefore, in the following only the comparisons between *Dry Audio* and *Perceptual Approach* simulations will be shown. The reason for the variations in the *Physical Approach* can be found in the fact that in the algorithm implemented in the TANGA engine, early reflections and diffuse reverberation are generated separately. At the time of the experiment, the fine tuning process for the transition between these two parts was not optimized. The

algorithm had not been fully tested before the assessment took place. In fact, the *Physical Approach* room simulation algorithm rendered echoic sounding reverberation which significantly aggravated the task of rating the localization.

Effect of Loudspeaker Setup The loudspeaker setup had effect on the evaluation ($F_R = 35.0$, $df = 2$, $p < 0.001$). The eight channel loudspeaker setup was rated having the best localization with a significant difference to the other setups (8ch - 5ch: $Z = -3.163$, $p < 0.01$; 8ch - 4ch: $Z = -5.629$, $p < 0.001$). The five channel setup was rated having the second highest localization quality and the ratings differed significantly from the 4ch channel setup showing the lowest localization quality rating ($Z = -3.810$, $p < 0.001$), see fig. 8.13.

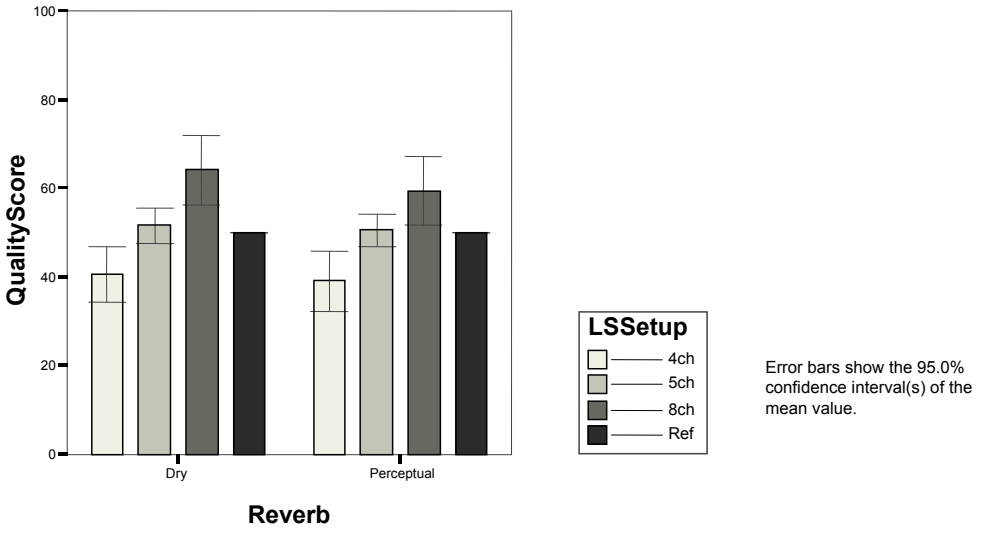


Fig. 8.13: The localization quality score of the *Dry Audio* (Dry) and the *Perceptual Approach* (Perceptual) simulations for the loudspeaker setups with four, five and eight channels. The reference setup consisted of the five channel setup and was not rated by the subjects.

Effect of Room Simulation Method There were no significant differences between the *Dry Audio* and the *Perceptual Approach* simulation methods in any of the loudspeaker setups assessed ($p > 0.05$, ns). This suggests the assumption that the *Perceptual Approach* room simulation method apparently does not influence the localization quality. This is interesting since the directions of incidence of the early reflections are not related to the room's geometry, but are always panned to a fixed angle of $\pm 30°$ from the direct sound source, see section 5.6.

Effect of Motion Paths The statistical analysis of the ratings of the loudspeaker setups with respect to the motion paths is especially interesting, since the aptitude of a loudspeaker setup for a certain task strongly depends on the supposed direction of incidence of the sound sources to be presented. If there are no sounds supposed to be coming from the back, a gap of loudspeakers in the back will hardly be noticed for the case of *Dry Audio*, see e.g. fig. 8.11 d). This is exactly what the statistical analysis shows: for *Dry Audio* played on the four channel loudspeaker setup, the motion path had significant effect on the evaluation ($F_R = 15.3$, $df = 3$, $p < 0.01$).

The differences in the ratings of the localization are significant for the pairwise comparisons of motion paths *c*) vs. *d*) ($Z = -2.977$, $p < 0.01$) and *a*) vs. *d*) ($Z = -3.127$, $p < 0.01$), see fig. 8.11. Apparently, test subjects had problems in consistently rating motion path *b*). This was also observed during the assessment, where the trials containing motion path *b*) were repeated twice as many times as the other trials. One reason for these problems can be found in the long repetition cycle of about $60s$ vs. about $20s$ to $25s$ in all other paths. As a result, the pairwise comparisons including motion path *b*) do not render significant differences in the rating of localization (*a*) vs. *b*): $Z = -1.613$, $p > 0.05, ns$; *b*) vs. *c*): $Z = -1.941$, $p > 0.05, ns$; and *b*) vs. *d*): $Z = -1.807$, $p > 0.05, ns$). Also, the analysis shows no significant difference between the localization in motion paths *a*) vs. *c*) ($Z = -0.243$, $p > 0.05$, ns), presumably because these motion paths are very similar with respect to the directions the sound source is coming from.

For the five channel setup, the effect of the motion path is not significant for the alpha-level chosen here ($F_R = 9.2$, $df = 3$, $p > 0.01, ns$), and there is clearly no effect for the eight channel setup ($F_R = 1.2$, $df = 3$, $p > 0.05, ns$).

The situation is similar (albeit more obvious) when additional early reflections and reverberation are present: for the *Perceptual Approach*, only the four loudspeaker setup shows a rather small effect of the motion path upon the test subjects' subjective perception of localization ($F_R = 13.1$, $df = 3$, $p < 0.01$). There are clearly no significant differences between the motion paths for the five channel ($F_R = 0.8$, $df = 3$, $p > 0.05, ns$) and the eight channel ($F_R = 2.8$, $df = 3$, $p > 0.05$, ns) setups.

8.7.5 Summary and Conclusions

As expected, the motion paths (and therefore the directions of incidence of the sound source) have the greatest impact on the perceived subjective quality of the loudspeaker setup used. This seems to be true independently from the room simulation algorithm used. Interestingly, the differences in localization quality are only significant for the four channel loudspeaker setup. For the five channel setup, the chosen alpha-level of 0.01 prevents the null hypothesis from being rejected. Yet, as $p < 0.05$ for the five channel setup, a Type II error cannot be ruled out completely. Using an eight channel loudspeaker setup means that localization quality is independent from the motion path (and therefore independent from the direction of incidence of the sound source), whenever an accompanying picture on a large screen is offered.

Yet, the (moving) picture of the virtual room almost always calls for a room acoustic simulation to be used in the application, as the listener expects the sound in the room to match the visual impression. Therefore, the results related to the *Perceptual Approach* room simulation are most interesting. These indicate that there are no significant differences between the motion paths in both the eight and the five channel loudspeaker setups. It can therefore be concluded that the optimum number (in terms of quality / cost ratio) of loudspeaker channels in audiovisual application systems using large screens actually seems to be five.

It remains to be evaluated whether this holds true for very high quality real time room simulation algorithms. As these are usually prohibitively cost intensive for the kind of application systems discussed here, we will not see these algorithms used in the near future.

A further result is that content can be an influence factor of level 1 (compare section 7.1): here, the suitability (and thus the perceived quality) of a certain loudspeaker setup depends

on the motion path presented. The quality of the stimulus created by the application is directly dependent on the content.

8.8 Assessment: Number of Internal Workchannels for Perceptual Approach

As described in section 5.6, the *Perceptual Approach* is standardized in MPEG-4 AudioBIFS, and a reference implementation is provided by the MPEG. Also, the reverberation module is available as a library of objects for the MAX/MSP graphical music creation environment. Yet, in internal but informal audio-only listening sessions the resulting reverberated signals repeatedly have been rated as not very realistic. The algorithm has been attested a tendency of producing harsh, sometimes metallic sounding and somewhat thin reverberation.

Thin or hollow sounding reverberation in natural environments is often attributed to small reverberant spaces such as bathrooms, in which the modal density is rather low.

A closer look at the structure of the *Perceptual Approach* algorithm (see fig. 5.13) reveals that the density of the reverberation depends on the number of internal channels, the so-called workchannels. This raises the question to what extent the density of the reverberation is actually relevant for different room sizes in the audiovisual context. If the number of internal workchannels is increased, does this lead to an increase in perceived quality and realism?

8.8.1 Test Setup

The assessment compared three versions of the *Perceptual Approach* algorithm to each other. They differed in the internal number of workchannels. Four, eight and twelve workchannels were tested in three virtual rooms of different size: a living room, a lecturing room and a sports gym. Figs. 8.14 and 8.15 show screenshots of the living room and of the lecturing room, respectively. The sports gym is shown in fig. 8.10.

Fig. 8.14: The virtual living room used for the assessment. This represented the small room.

Two different dry audio stimuli were used as sound source material, one spoken language excerpt and one music piece playing an acoustic guitar.

In this assessment, the eight channel loudspeaker setup as described in fig. 8.9 (left) was used. The projector received the image directly from the rendering computer via a

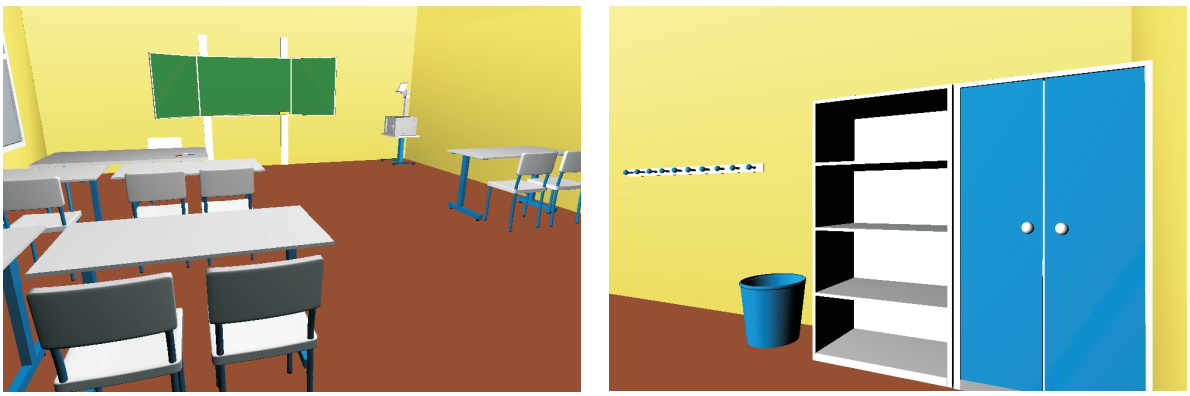


Fig. 8.15: The virtual lecturing room. The lecturing room represented the medium sized room in the assessment.

VGA cable with a resolution of $1024 \times 768 pixels$ at $75 Hz$. An eightfold anti-aliasing was computed by the graphics card to make curved lines appear more natural and smooth. No scan converter was used such that the technical image quality was as high as possible.

8.8.2 Implementation of the Assessment, Task and Rating Scale

Again, test subjects used the Input Device described in section 6.3 to enter their ratings and to perform trial after trial. The assessment was organized as a repeated pair comparison between two items. The rating scale was ranging from -50 to $+50$. No semantic designators indicating quality steps were used. ITU-R BS.1284 [itu1284] recommends the use of a horizontally oriented fader for pair comparison tests of the type described here, but at the time of the experiment the second prototype[11] of the Input Device was not fully functional yet. Therefore, a vertical fader had to be used. The upper end of the fader was marked A, the lower end B. By moving the fader into the top position (scale level 50), test subjects indicated that item A was of much higher perceived reverberation quality. Moving the fader to the bottom position (scale level -50) favored item B. The middle position (around scale level 0) was reserved for situations in which subjects thought that the quality of both items A and B was equal. Fader positions in between were used to indicate tendencies. A button above the fader labeled REPEAT allowed to repeat the items. Another button on the right side of the Input Device labeled NEXT was used to enter the test subject's rating and switch to the next trial.

Two variants of the assessment were devised. Subjects either compared items with four and eight workchannels or items with eight and twelve workchannels to each other. This comparison included all possible combinations of items (AA, AB, BA, BB), resulting in four trials per scene. Each trial consisted of two items. This way, the reliability of the participants could easily be tested. Each subject evaluated the perceived reverberation quality for all three scenes (virtual rooms) and for the two audio stimuli (music and speech), resulting in a total of 24 trials. Fig. 8.16 gives an overview of the procedure. Fig. 8.17 shows the chronological sequence of a test session.

In the beginning of a session the sequences of the trials and audio signals were randomly generated to avoid sequence effects. The sequences were different for each test subject.

[11] The second prototype of the Input Device has a face plate that can be exchanged to reduce the amount of buttons visible and to place the faders at arbitrary positions and orientations.

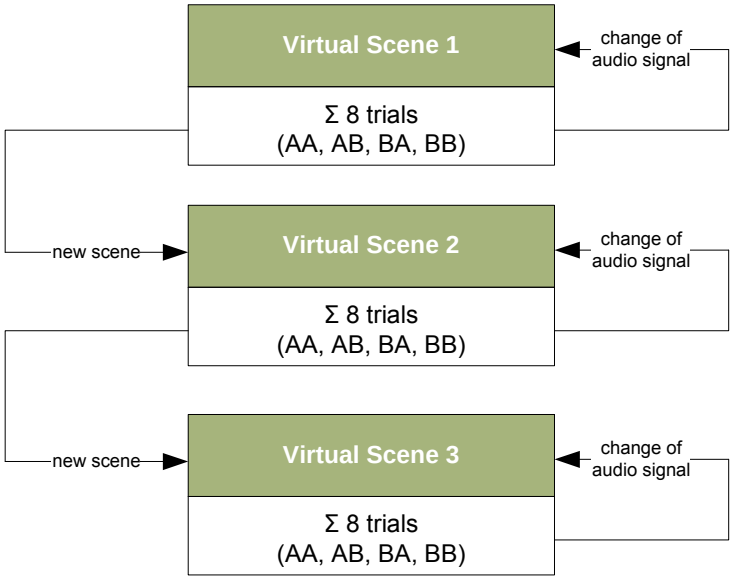


Fig. 8.16: Overview of the test procedure used in the pair-comparison assessment of the *Perceptual Approach* internal workchannel count.

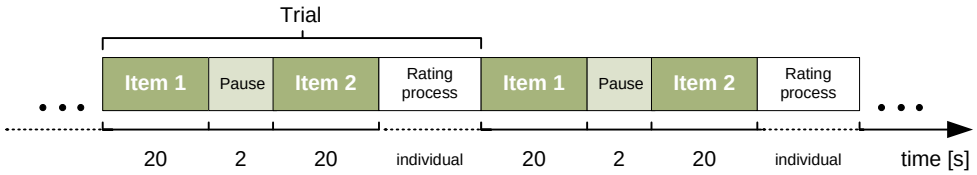


Fig. 8.17: Chronological sequence of the test procedure used in the pair-comparison assessment of the *Perceptual Approach* internal workchannel count.

Before starting the assessment, subjects were given written instructions for the assessment. After reading, questions were answered by the instructor. Subjects were also instructed to practice the navigation with the computer mouse in a short training session. This included the use of the Input Device for the rating of items. The total assessment time including the training was approximately $40 minutes$ per subject.

Test subjects were asked to actively navigate through the virtual room and to collect items within the room: a remote control in the living room, a textbook in the lecturing room, and a football in the gym. Whenever they approached the item closely enough, it was "collected" automatically[12]. The item was subsequently placed at different locations in the room, so that test subjects had to move continuously using the computer mouse as a navigation device. They were given this task in order to shift their focus from the active and conscious reception of the auditory stimulus alone to a more application-centered overall reception of both auditory and visual stimuli at the same time.

8.8.3 Analysis and Test Results

Hypotheses The null hypothesis was that subjects would not be able to perceive differences in the density of the diffuse reverberation part. The alternative hypothesis was

[12] This was done using the MPEG-4 BIFS *ProximitySensor* node described in section 5.3.1 and the ECMA script language (see section 5.3.5).

that subjects could differentiate between the number of internal workchannels applied to generate the diffuse reverberation.

Test Panel A total of 22 subjects participated (17 male, 5 female). They were either students or staff from Technische Universität Ilmenau. All test subjects reported to have normal hearing and normal or corrected to normal vision. They were between 22 and 35 years of age. Half of the participants had already gained experience in former listening tests. Of these, five can be classified as expert listeners.

Data Analysis The results of the subjective assessment could not be evaluated using statistical analysis methods, because none of the subjects delivered consistent ratings during the whole test. Therefore either their reliability has to be questioned, or the differences to be rated were too small to be significant. However, a partial result without notice of the reliability question is presented in fig. 8.18. This figure shows the averaged ratings for the large virtual room (gym) with the acoustic guitar sample being played as audio material. Obviously, the average converges to zero, so the algorithms were estimated as being of equal subjective quality in the bimodal context. This assumption is further substantiated by the fact that the group of expert listeners showed exactly the same inconsistencies in their ratings.

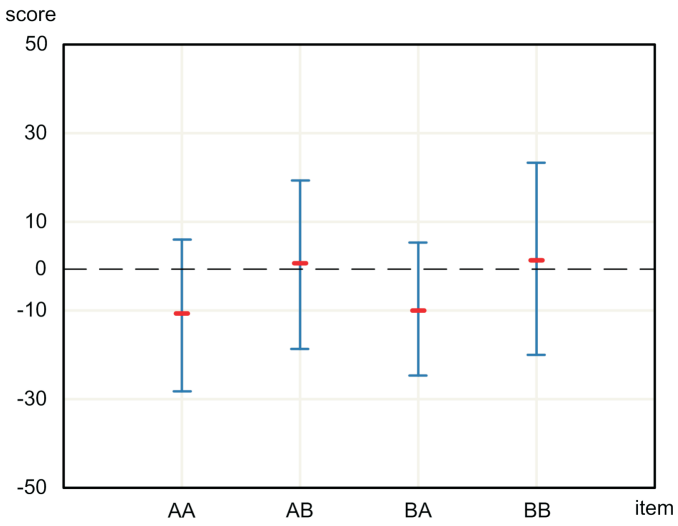


Fig. 8.18: Test results for the virtual room 'gym' and the acoustic guitar sample, average and 95% confidence intervals. Item A is the four workchannel algorithm, item B the eight workchannel variant.

8.8.4 Summary and Conclusions

The test results make clear that the test subjects were not able to identify the three versions of the algorithm under assessment. Thus an increase of the density of the diffuse reverberation part remains without perceivable improvement of the quality in bimodal perception. Therefore it can be concluded that the algorithm with four internal reverberation channels seems to be sufficient for the case of audiovisual perception and active exploration.

8.9 Assessment: Influence of Interaction on Perceived Quality

Looking at the results of the assessment described in section 8.8, it seems remarkable that a variation of the reverberation density should have gone unnoticed even with expert listeners. After *listening* to the variations of the *Perceptual Approach* algorithm used in the assessment (and actually hearing slight but clearly noticeable differences in sound coloration), it was suspected that the larger tolerance toward changes in density that showed in the assessment might be due to the *task* that test subjects had to perform. The question was whether the limited amount of interaction demanded in the assessment (collecting items by actively navigating through the virtual scene) was already enough to distract subjects from properly perceiving, comparing and rating the reverberation quality? If so, would this principle be generally applicable, thus allowing to reduce the computational complexity of the audio simulation part in interactive audiovisual applications without deteriorating the perceived overall quality? The assessment described in this section was designed as a first step towards answering this question.

8.9.1 Test Setup

The assessment compared the perceived overall quality of audiovisual scenes under different degrees of interaction. The actual amount of interaction was determined by different tasks that the test subjects had to fulfill. To make navigation easy to unexperienced subjects, a large room without obstacles (objects on the floor) was chosen - again, the gym was used, see fig. 8.10.

The assessment took place in the Listening Lab at Technische Universität Ilmenau. It was expected that subjects would move all around the virtual gym. Because the virtual sound source was located in the center of the room, direct sound was expected to be heard rather often from behind the test subjects. Therefore the eight loudspeaker setup (see fig. 8.9, left) was chosen.

The technical parameters for the visual display were identical to the preceding assessment, see section 8.8.1.

Test Material All test materials were audiovisual scenes 30*seconds* long. Two different audio contents, music (acoustic guitar) and speech (male voice), were presented with three different reverberation strengths: the lowest amount of reverberation was produced by an image source algorithm (see section 5.7) of order one, the highest by an algorithm of order three. Because the image sources were fed into the nested all pass filter network that produced the diffuse reverberation, the maximum order of image sources also influenced the amount of late reverberation. Two different audio contents were selected because of the different spectral distribution, familiarity and the preference of reverberation amount[13]. The visual content, the sports gym, was presented with two different motion paths representing a spatial movement within the virtual scene. The motion paths provided an equal number of items with main direction of incidence of direct sound from the left as from the right hand side. They were made as equal as possible between the parallel tasks.

[13] In an earlier experiment not detailed in this thesis, the author found that the preferred amount of reverberation depends on the audio content [rei06]. Apparently, for music signals a higher amount of reverberation is regarded as natural than for speech signals. This is presumably due to the often reduced intelligibility of reverberated speech compared to dry speech signals.

The experiment was organized in four parts. In the first part, demographic and psychographic data (age, gender, professionalism in video and audio handling, attendance at earlier listening experiments, playing computer games and musical instruments, and listening experience with surround sound systems) was collected with a pre-questionnaire.

The actual assessment was divided into collecting quantitative (second part) and qualitative (third part) data[14]. The fourth part consisted of a post-assessment questionnaire. Three pilot test iterations were done prior to finalizing the test setup. The average duration of the experiment was 65 minutes per subject including all four parts.

8.9.2 Quantitative Assessment: Implementation, Task and Rating Scale

Test Method In the second part of the experiment the single-stimulus method, also known as Absolute Category Rating (ACR, see section 8.5), was used. It is suitable for multimedia performance and system evaluations (compare ITU-R BT.500 [itu500], ITU-T P.910 [itu910]). Even though double- and multi-stimulus methods are powerful for high quality discrimination assessments, these would have made the quality evaluation with a parallel task become very complicated for the participants. The stimuli were viewed one by one, and overall quality was rated independently and retrospectively (e.g. ITU-R BT.500-11 [itu500]) on a continuous and unlabeled scale from 0 to 100 in a randomized presentation order.

Test subjects used the Input Device described in section 6.3 to enter their ratings and to perform trial after trial. The control panel layout of the Input Device was similar to the one described for the previous experiment: one motorized, vertical fader was used to adjust the quality rating. A button on the right side of the Input Device labeled NEXT was used to enter the test subject's rating and switch to the next trial. Test subjects were not allowed to repeat a trial.

Test Procedure The quantitative assessment procedure consisted of a quality anchoring and three quantitative evaluation tasks with a training prior to each of them, see fig. 8.19. The anchoring introduced the quality extremes of the test materials with different audio signals. The quality evaluation itself was conducted by subsequently assigning three different parallel tasks to the subjects:

1. *Listen and watch* task: Test subjects were asked to experience an automated movement through the virtual scene. No activity on their side was requested. The automated movement lasted around *30 seconds*, selected from two different predefined motion paths, see fig. 8.20.

2. *Listen and press a button* task: Again, test subjects were asked to experience an automated movement through the virtual scene. Also, the motion path presented was one of the two in fig. 8.20. This time, a football automatically appeared within the field of view. It was subsequently approached and (again automatically) collected. Then, a new football would appear, and so on. Test subjects were asked to immediately press a button on the Input Device whenever the football appeared.

3. *Listen and collect the ball* task: Test subjects were using the computer mouse to navigate freely inside the virtual gym. Their task was to collect the football that was

[14] A definition of the terms "quantitative" and "qualitative data" as well as a further discussion of the differences in methodology applied for obtaining these is given in section 8.5.1

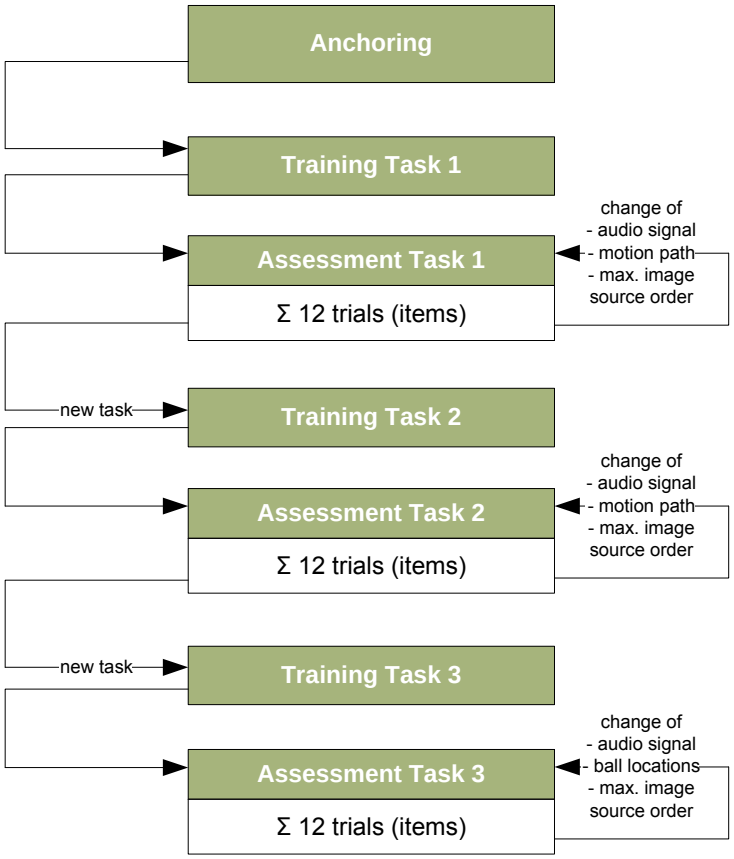


Fig. 8.19: Overview of the test procedure used in the quantitative part of the single-stimulus assessment evaluating the influence of interaction / task upon the perceived overall quality.

positioned somewhere on the floor. When they had approached the football closely enough, it was collected and re-appeared in another location. The new location was either within the field of view, or the subjects had to turn around to see the football again. They were asked to collect as many footballs as possible within a given time of *30 seconds*.

All tasks had the same evaluation instructions and the sequence of the tasks was randomized between the participants. For each task, four different sequences of items had been randomized before the test. One of these sequences was randomly selected before each task.

The post-test questionnaire was asking the following questions in written form:

1. Were you already familiar with any of the stimulus material before the experiment (music, speech, visual)?
2. Do you think that any of the presented tasks contained stimulus material of generally higher quality than other tasks? If so, indicate the order of quality!
3. Do you think that some of the stimulus material was more difficult to rate than other? If so, indicate the order of difficulty!
4. Do you think that one of the two audio examples (speech, music) was easier to rate than the other? If so, which one was easier?

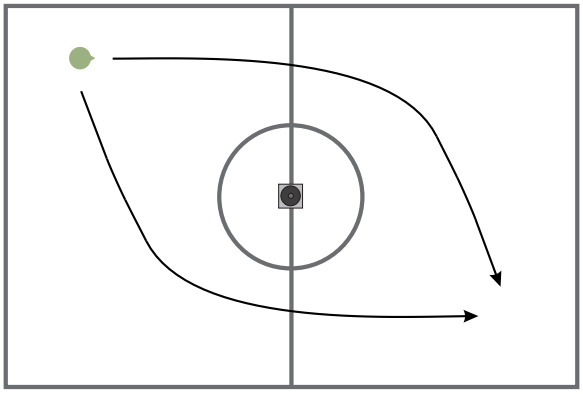


Fig. 8.20: The two pre-defined motion paths used in the first (listen and watch) and second (listen and press a button) task.

5. Do you think that one of the two audio examples (speech, music) was of generally higher audio quality than the other? If so, which one was of higher quality?

6. Do you have any general remarks? Please indicate here!

8.9.3 Analysis of the Quantitative Assessment

Hypotheses The null hypothesis was that the degree of the task would not influence the perception of overall quality; subjects would give equal ratings to the perceived quality of the scene, no matter what task they had to perform in the rating session. The alternative hypothesis was that the amount of interaction (the degree of the task) offered during the assessment would have an influence on the perceived overall quality of the scene.

Test Panel The experiment was conducted with 40 participants, mostly university students and staff, aged from 23 to 39 years ($M = 26$, $SD = 3.6$). Ten participants were female and 30 males. All participants reported to have normal hearing and normal or corrected to normal sight. 30% of the participants were experienced assessors with more than two attendances at previous subjective assessments.

Data Analysis The results were analyzed using SPSS for Windows version 13.0. Non-parametric methods of analysis were applied because the data did not reach the preconditions of normal distribution for parametric methods. Friedman's and Wilcoxon's tests were used to compare the differences between ordinal independent variables in the related design [coo04]. In the analysis of the post-test questionnaire data, Kruskall-Wallis' and Mann-Whitney U tests were used to compare differences between two groups in the unrelated design [coo04].

8.9.4 Results and Discussion

The different tasks that test subjects had to perform did not have an effect on the quality evaluation (Friedman: $\chi^2 = 3.3$, $df = 2$, $p > 0.05$, $p = 0.190$, ns) when the ratings were averaged across the maximum orders of image sources, contents and motion paths.

The maximum order of image sources had an impact on the quality evaluation (Friedman: $\chi^2 = 106.6$, $df = 2$, $p > 0.001$). The material presented with the lowest image source order

was considered the most pleasant, followed by second order, and last third order image sources. The differences were significant between all image source orders when results were averaged over other factors. (Wilcoxon: order 1 vs. order 2: $Z = -8.16$, $p < 0.001$; order 1 vs. order 3: $Z = -9.87$, $p < 0.001$; order 2 vs. order 3: $Z = -2.43$, $p < 0.05$). The results remained the same within task examination, with the exception that there were no significant differences between the second and third image source order in any of the parallel tasks (watch: $p > 0.08$, press: $p > 0.190$, collect: $p > 0.224$).

Quality evaluation was not affected by the audio content types or visual motion paths. The music and speech contents were mostly rated into the same level within each task ($p > 0.05$). An exception was that the music content was preferred over the speech content for the presentation of the first image source order (watch task: Wilcoxon $Z = -3.01$, $p > 0.01$; press the button task: Wilcoxon $Z = -2.92$, $p > 0.01$). When the contents were averaged over the other factors, music was rated as more pleasant than speech content (Wilcoxon: $Z = -2.88$, $p < 0.01$).

Two different motion paths were equally rated in each task ($p > 0.05$). The only exception appeared in the listen and press the button task with music content presented with the second image source order (Wilcoxon: $Z = -2.2$, $p > 0.03$).

Surprisingly, test subjects did not rate the audio material with the highest maximum image source order (third order) as the best quality items, but preferred a lower maximum image source order for both speech and music content. This was not expected, especially for the presentation of music content. Again, this was presumably due to an unoptimized transition phase between early reflections and diffuse reverberation[15]. Although this apparently degraded the audio quality, it does not influence the results of the assessment: The question was not whether a maximum of first or third order image sources would render higher overall perceived quality, but whether different amounts of interaction would have an influence upon the subjects' ability to discriminate between different qualities offered.

This was clearly not the case in this assessment. Section 8.9.8 further elaborates on this and the possible causes.

8.9.5 Qualitative Assessment: Implementation and Questionnaires

The third part of the experiment focused on the evaluation criteria that the test subjects had used. This was necessary because they had not been given detailed instructions on what to pay attention to in the quantitative assessment. Subjects had only been asked for the overall perceived quality of each scene, such that their motivation for giving a certain rating needed to be determined separately[16]. A semi-structured two-leveled oral interview was recorded for later analysis. It gathered data about the overall quality evaluation criteria with active (*listen and collect the ball* task) and with passive (*listen and watch* task) parallel tasks. Finally, a post-questionnaire about the experienced difficulty of the assessment and the presented quality during the tasks ended the experimental session.

The qualitative interview gathered quality evaluation criteria and reasons for impressions of quality. For this, a semi-structured interview technique was applied. Such a technique is beneficial when the research field is relatively unknown and prior expectations are not set [coo04].

[15] This is discussed in the analysis of the "optimum number of loudspeakers" assessment in section 8.7.4.
[16] For a discussion of the general problem of affective measurements see section 8.1.

Test Procedure In the first part of the interview, all participants answered questions about their overall quality evaluation criteria. This was done immediately after the quantitative assessment. The interview was asking main and supporting questions, see fig. 8.21. The main question was presented several times during the interview with slight variations.

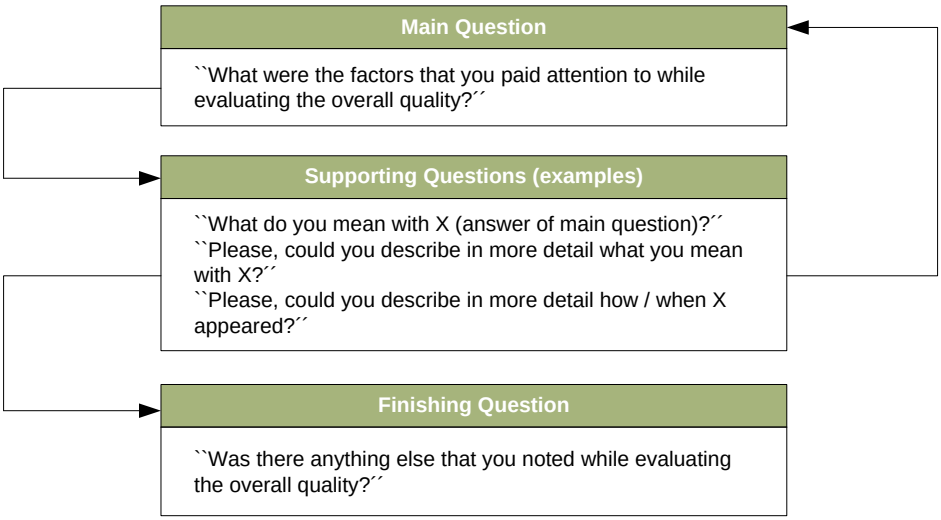


Fig. 8.21: The structure of questions asked in the general part of the semi-structured interviews.

The supporting questions clarified the answers given by the test subjects to the main question. The finishing question was asked to make sure that also those factors were named that were possibly considered irrelevant by the test subjects themselves.

Subsequently, a detailed interview was conducted with the aid of selected stimulus material. The stimulus material was presented in pairs in which task, visual or auditory variables were varied, see fig. 8.22. The paired presentation allowed to collect the differences between the stimuli that were noted by the test subjects. Half of the sample conducted the detailed interview with speech content and half with music content.

All interviews were held in German language with native German speakers and recoded later.

8.9.6 Analysis of the Qualitative Assessment

The qualitative analysis was based on Grounded Theory originally introduced by Strauss and Gorbin [str98]. The Grounded Theory approach can be applied in research areas that are little known, as in assessments evaluating a multi-modal quality experience, and whenever the research tries to understand the meaning or nature of a person's experiences. The theory or its building blocks are derived from data with systematical steps of analysis. Therefore it summarizes the central features of the data.

All analysis material was transformed into expressions in the initial coding (transcription from voice recordings). The final categories were created by two people separately (both native German speakers) in order to find the concepts and their properties.

All concepts were organized into 13 categories and all data was arranged according to these categories. In the categorization, test subjects' several mentions of the same topic were recoded only once. Coding reliability was estimated with intercoder agreement between the two researchers in all codes (Cohen's $kappa > 0.68$) [car96, har97]. After the analysis, categories were translated into English language.

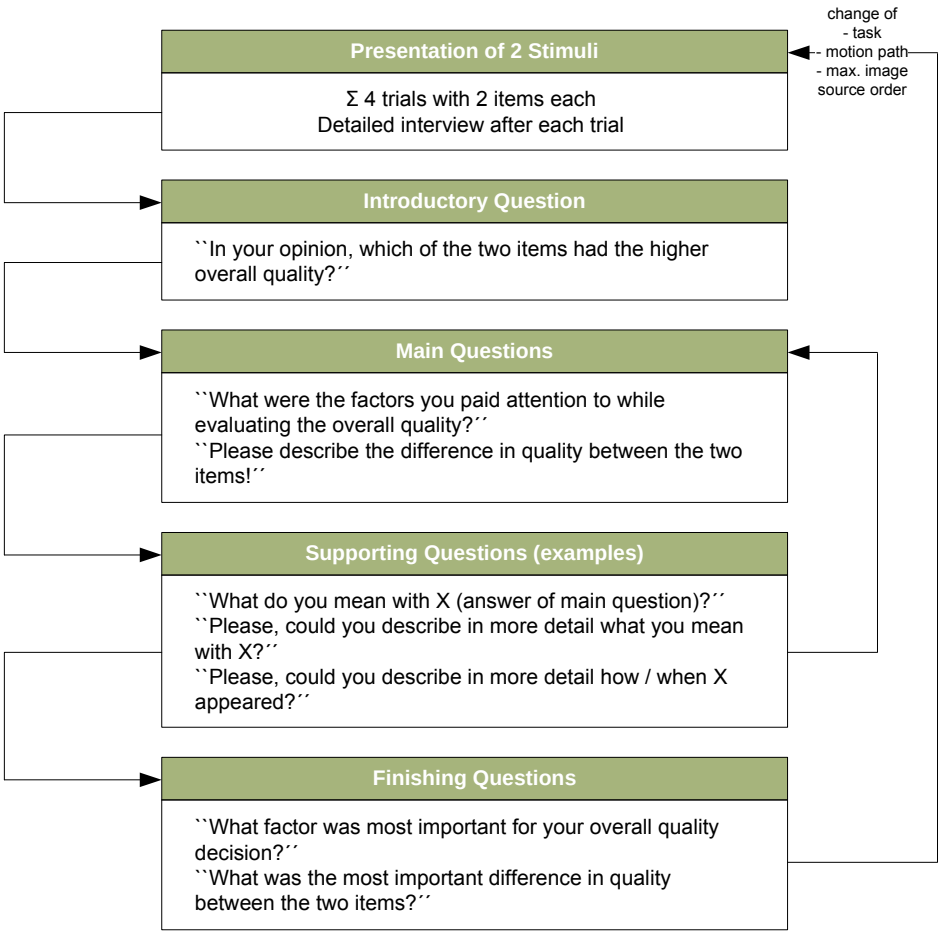


Fig. 8.22: The structure of questions asked in the detailed (pair comparison) part of the semi-structured interview.

8.9.7 Results and Discussion

Fig. 8.23 presents the overall quality evaluation criteria. The most utilized categories were *localization*, *euphony*, *strength and length of reverberation* and *dependency of sound on distance*. *Localization* included the semantic descriptions of "position of sound source", "locating", "determinability of direction", and "focusing of localization". *Strength and length of reverberation* was separated from the category *Reverberation unspecified* because test subjects often pointed out that they had paid attention to the "strength and length" on the one hand and to "coloration, sound etc. of reverberation" on the other hand - the two mentions were, although not precisely formulated by the test subjects, clearly not the same. *Dependency of sound on distance* includes the connotations of "direct sound", "reflexions off walls", and "direct / diffuse sound ratio". It comprehends every change in the audio material that was noted when navigating through the virtual room. *Audiovisual concordance* comprises "orientation in space" and "determination of own position in the room". *Perception of the setup* included the semantic descriptions of "smoothness of movement", "latency", "perception of phantom sound sources", and "sound pressure level".

Table 8.3 shows the analysis of the detailed interviews based on the evaluation of the *listen and watch* task vs. the *listen and collect the ball* task. In the pairwise presentation of stimulus material during the detailed interview, only these tasks were presented. Only

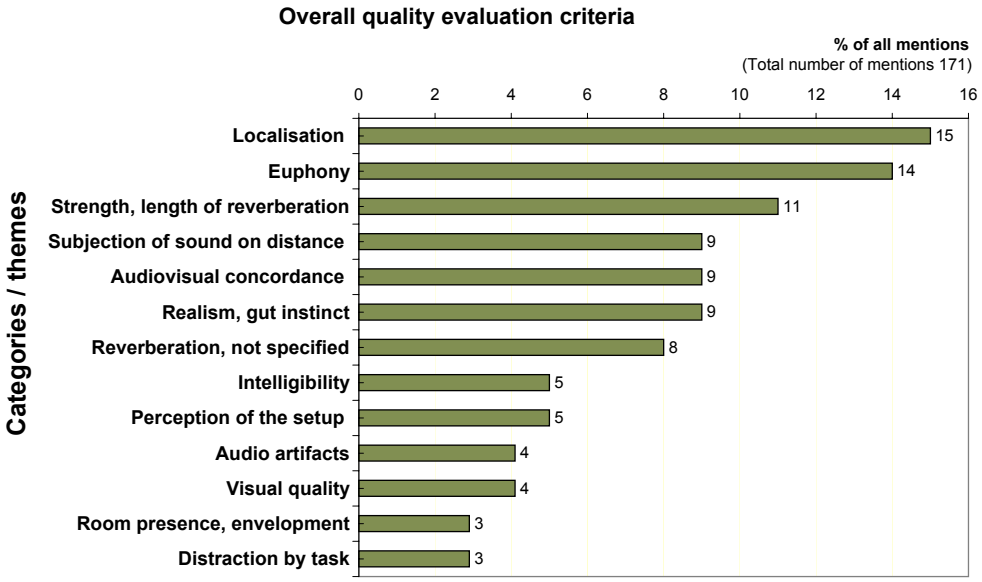


Fig. 8.23: The 13 different quality evaluation criteria categories, sorted according to the percentage of mentions in the general part of the interviews.

the five most mentioned categories are given in each column. After each presentation pair, subjects were asked which of the two items was of higher overall quality. Coinciding with the quantitative results, the maximum image source order 1 was perceived as being more pleasant than order 3 for both audio contents (Wilcoxon test: Music: $Z = -2.29$, $p > 0.05$, Speech: $Z = -3.20$, $p > 0.001$). This was independent from the task.

The results of the detailed interview support the results of the overall quality evaluation criteria obtained in the general part of the interview, see fig. 8.23. The evaluation criteria was pointed out by the test subjects using similar descriptions for both audio contents. For the *listen and watch* task, *Localization* (24%), *Euphony* (Music: 20%, Speech: 21%) and *Dependency of sound on distance* (Music: 10%, Speech: 14%) were among the most mentioned categories. These percentages emerged for the *listen and watch* task when the image source order was varied in the stimulus material presentation and when the motion path was kept constant between items. For other variations between items these percentages vary slightly. Yet, *Localization* and *Euphony* were always among the most mentioned categories.

Interestingly, *strength and length of reverberation* seems to be of more importance for perceived quality of speech than for quality of music signals. This could be related to the familiarity of the signals: we know how spoken language is supposed to sound in a room, whereas the sound of a music recording might have been influenced at various stages in the production chain. It is difficult to know how a piece of music is supposed to sound in a certain (virtual) room without a (dry signal) reference.

8.9.8 Summary and Conclusions

Unlike expected, there was no significant difference in the error rate of quality ratings between an active exploration task (user navigated freely through a virtual room, *listen and collect the ball* task) and a passive exploration task (user was being moved through the room on a pre-defined navigation path, *listen and watch* task). The quantitative analysis' result both supports and contrasts with the result of the interviews performed with the

Audio Content		Speech				Music			
Task		Watch Task		Collect Task		Watch Task		Collect Task	
Stimulus Variation		Reverb	Reverb & Path	Reverb	Reverb & Path	Reverb	Reverb & Path	Reverb	Reverb & Path
Categories									
	Localization	24%	12%	11%	10%	24%	23%	21%	16%
	Euphony	21%	24%	14%	21%	20%	17%	16%	14%
	Strength, length of reverberation	14%	14%	23%					
	Dependency of sound on distance	14%				10%	9%		10%
	Audiovisual concordance		10%			6%		9%	12%
	Realism, gut instinct		10%	9%	13%		11%	9%	
	Room presence, envelopment					10%		9%	
	Perception of audiovisual setup			9%	10%		11%		
	Reverberation (not specified)	10%			10%				8%

Table 8.3: Each column shows the five most mentioned categories for the two audio stimuli (Speech, Music) for different parallel tasks, with variation of the max. order of image sources (column *Reverb*), and with and without variation of the motion path (column *Reverb & Path*). Percentages show % of all mentions.

test subjects after the assessment. On the one hand, neither the quantitative nor the qualitative evaluation results allow to reject the null hypothesis. No influence of difficulty of task (in the form of varying amounts of user interaction) on the subjects' ability to discriminate between different qualities could be substantiated.

On the other hand, a considerable part of the test panel claimed that they had been distracted by the navigation task and therefore found it more difficult to rate the quality in the self-motion (interactive) situation.

There are three possible explanations for the insignificant effect of interaction on the perceived overall quality:

1. The overall quality differences offered were too obvious. The quality steps between the items were too big. It was therefore too easy for the test subjects to differentiate between the quality levels.

2. User interaction took place in a different modality than the quality rating. Whereas the interactive navigation task involved haptics and visual, the quality ratings were based on audiovisual percepts. These modalities might be largely independently pre-processed in the human brain (as opposed to what e.g. Coen [coe01] argues), therefore no significant influence has been detected in the experiment.

3. The task of self-motion in a virtual room was not demanding enough. It was too easy for the test subjects, so they could still concentrate enough on the overall quality rating even during active exploration when collecting the football.

All of these attempted explanations are closely related to potential (lack of) capacity limits, see section 2.5.4. The subsequent sections describe a number of experiments that were performed to find out which of the above mentioned reasons represents the most probable explanation for the surprising results in this experiment.

8.10 Assessment: Influence of a Working Memory Task on Perceived Quality

The result of the experiment described in the preceding section contrasts with other publications. A number of studies (e.g. [zie03] and [kas03]) have focused on the fact that the amount of involvement in a task apparently can have a positive effect on the perceived quality. In these studies it was observed with some test subjects that the more they were involved in the task, the less they noted impairments in the quality of the simultaneously presented audio. The perceived quality was not directly dependent on the degree of the impairment.

In the experiments described in these publications, subjects were asked to evaluate the audio quality of a multichannel audio system while playing a computer game. It was observed that a divided attention between auditory and visual modalities has a significant effect on the rating of audio quality. The authors report that in the active gaming condition subjects rated the same audio quality higher than in the passive looking-at-a-video condition. However, these results were only valid for some listeners and for some experimental conditions. Furthermore, the sample sizes of these experiments were too small (six and seven participants, respectively) to render statistically meaningful results.

As mentioned before, the previous experiment suggested three alternative explanations for the fact that there seemed to be no significant influence of the task upon the quality

rating. The first and third explanation are closely related: following the theories of divided attention, attention is defined as an information selection process that is characterized by limited information processing resources (see e.g. [pas99, sty97]). This means that smaller differences in stimuli can be detected if there is more processing capacity available, i.e. a parallel task is less demanding, as well as vice versa. This reduces the remaining independent factors, leaving two possible explanations to be further examined:

- Whereas the user interaction was related to visual and haptic modalities, the quality rating was based on audiovisual percepts.

- The subjects' task of self-motion was not demanding enough, making the quality differences too obvious.

In order to examine these explanations, both were addressed with this experiment: on the one hand, user interaction, rating process and tasks aimed at sharing the same modality. On the other hand test subjects were confronted with a mentally complex task.

8.10.1 A Scalable Working Memory Task

Whereas the experiment in section 8.9 had attracted subjects' attention to a visual task while varying auditory quality, here a task related to the auditory modality was chosen. The so-called auditory n-back task ensured that the interactive task and the sound quality rating were performed in the same (auditory) modality.

The n-back task is one of the most popular paradigms for studies of working memory [owe05]. Typically, the subject is required to monitor a series of stimuli and to indicate whether or not the stimulus currently presented is the same as the one presented n trials before. n is a pre-specified integer, usually 1, 2 or 3. The task requires on-line monitoring, updating, and manipulation of information remembered by the subject and is therefore assumed to put great demand on working memory.

In the experiment a series of numbers spoken in German language were used as auditory stimuli in the n-back task. Subjects had to press a button labeled "equal" when the numbers presented were the same, and a button labeled "different" when they were not. An example of 1-back and 2-back working memory tasks with correct responses is shown in fig. 8.24.

In experiments related to working memory the difficulty of the task (*degree*) has to be raised carefully. On the one hand, subjects should not be frustrated with a task far too difficult for them. On the other hand, subjects in this experiment should not be able to concentrate fully on the quality rating during the active exploration part.

Therefore, the actual selection of the degree of the task was the result of a pilot experiment. Here, five subjects had to try different tasks: Navigation in a virtual scene without additional task, navigation with 1-back task, navigation with 2-back task, and navigation with 3-back task. As expected, subjects felt more stress with the increasing degree of the task. However, to concentrate on the 3-back task while navigating through the virtual room turned out to be almost impossible. Pilot test subjects felt frustrated because they were not able to complete this very demanding task. Therefore the 3-back task could not be used in the later experiment.

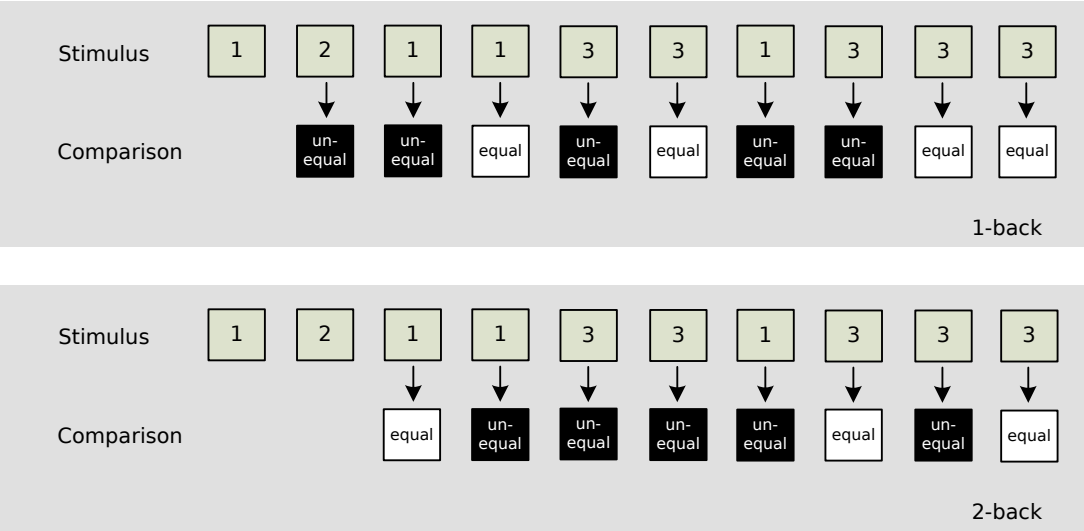


Fig. 8.24: Presented stimuli and correct answers ("Comparison") for a 1-back and 2-back continuous-matching-task.

8.10.2 Test Setup

Selection of Quality Attribute The criteria for selecting the quality attribute used in this experiment resulted from the suggestions obtained in the previous experiment (section 8.9) and can be summarized as follows:

- The evaluation of the overall perceived quality is a multidimensional process.
- The differences between quality levels should not be too obvious. If they are, the influence of user interaction on the quality rating may remain unnoticeable.

A number of significant attributes had been collected in the qualitative part of the previous assessment. In this experiment, the purpose was to have subjects concentrating on a single, well-defined quality attribute. Marginal quality steps between the items were to be examined. Therefore the length of reverberation was chosen as attribute, because it had turned out to be one of the prominent attributes in the interviews of the previous experiment. Furthermore, the length of reverberation is clearly defined and easily compared. The acoustic real-time rendering in the test was performed using the MPEG-4 Audio *Perceptual Approach* as described in section 5.6. The algorithm variant with eight internal reverberation channels was used. Contrary to geometry based implementations, the Perceptual Approach is controlled via a number of so-called perceptual parameters. One of these parameters, the *lateReverberance*, corresponds to the length of reverberation.

In a pilot experiment with five participants the reverberation times to be used for the later test were determined to be $1.0s$, $1.5s$, $2.0s$, $2.5s$ and $3.0s$. These were easily differentiated by the pilot subjects.

Test Material The test material content was, according to the bimodal context of the experiment, of audiovisual type. The visual content consisted of a three-dimensional virtual scene, which represented the foyer of the main lecturing building of Technische Universität Ilmenau, see fig. 8.25.

Fig. 8.25: View of the virtual foyer used in the experiment.

Technical Setup The auditory signals consisted of a sequence of numbers, spoken by a male voice in German language, with a total length of 20 seconds. The sequence included numbers from one to four. A number followed three seconds after the previous number. To avoid learning effects the sequences offered to each subject were varied in the course of the experiment.

The assessment took place in the Listening Lab at Technische Universität Ilmenau. The loudspeaker setup used was the eight channel setup described in fig. 8.9 (left). The reproduction setup for the visual content was identical to the one described in section 8.8.1.

8.10.3 Implementation of the Assessment, Task and Rating Scale

Test Method The pair comparison method, see section 8.5, was chosen to compare a modified item in relation to a reference item. The modification was performed on the attribute *length of reverberation time (RT)* that was varied between trials. The items presented in the experiment had five different reverberation times, with 1.0, 1.5, 2.0, 2.5 and 3.0 seconds. The medium reverberation time (2.0s) represented the reference item.

In each session the subjects evaluated five trials. First the reference item was played, followed after a small break by the item under evaluation with a modified reverberation time. A timing diagram is shown in fig. 8.26. The pairs were presented in random order for each subject. Because of time constraints each pair of items was presented only once per session.

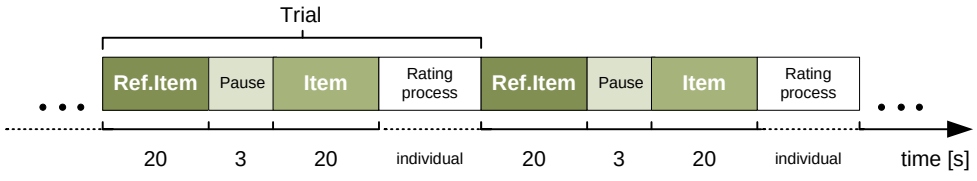


Fig. 8.26: Chronological sequence of the test procedure using the pair-comparison method.

Rating Scale For the rating of the reverberation time subjects used the Input Device described in section 6.3. The control panel of the Input Device was equipped with a single

vertical slider with a continuous scale from 0 to 100. Only the center position of the scale was marked, see fig. 8.27. Subjects were asked to rate shorter reverberation times with lower scale values, higher reverberation times with higher scale values. The extremes (1.0*s* and 3.0*s*) corresponded to the lower and upper limit of the scale, whereas the center position of the scale represented the reverberation time of the reference item. The items were rated retrospectively. Two buttons labeled "equal" and "different" were located in the central part on the right hand side of the control panel. These were used by the subjects to indicate the results of the continuous n-back comparisons during the presentation of items. The "next" button would read the current slider position, transmit it to the logging tool, move the slider to the neutral position and start the next trial.

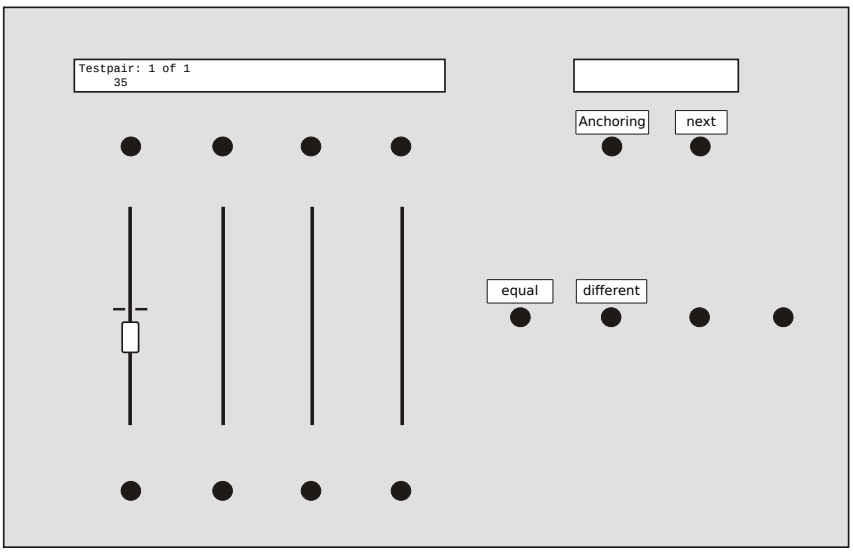


Fig. 8.27: The control panel layout of the Input Device as used in the divided attention / working memory assessment.

Test Procedure The experiment took roughly 45 minutes per subject and was divided into an anchoring phase and three evaluation sessions. The schematic procedure is shown in fig. 8.28.

The anchoring consisted of the presentation of the full quality range of stimuli that the subjects were going to experience during the experiment. While the subjects were watching a static picture of the visual scene, the reference item and the quality extremes of the test material were played twice, interrupted by a break. This procedure allowed test subjects to familiarize themselves with the range of the reverberation times to be presented in the assessment.

The anchoring was followed by the assessment sessions with prior training. Subjects went through three sessions with different degrees of the interaction task, while the process of quality evaluation (rating) was kept the same. The tasks varied in the degree of interaction, including navigation only, navigation with 1-back working memory and navigation with 2-back working memory. Navigation consisted in actively moving around the virtual room by using a computer mouse. Each subject was allowed to move freely without restrictions regarding speed and complexity of the movement. All test subjects navigated actively during the assessment - none of them did not move around. Yet, the actual amount of navigation activity was not recorded and thus not taken into account in the analysis. This

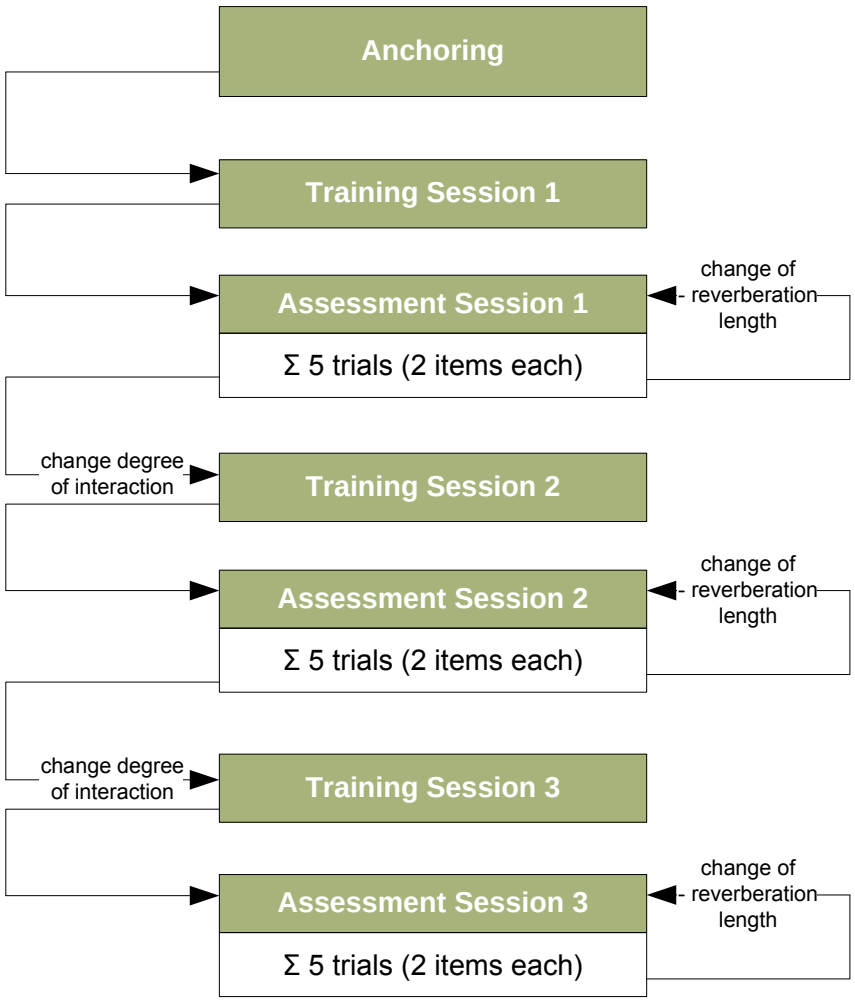


Fig. 8.28: Test procedure with anchoring and three evaluation sessions.

was because subjects expectedly had different degrees of experience with the kind of mouse navigation used here, thus making a direct comparison very questionable. The presentation order of the tasks was randomized between subjects.

In the beginning of each session, test subjects were presented with instructions in written form. These described the test procedure, the interaction tasks and the evaluation process. They were able to ask for additional explanations from the experimenter if necessary.

Furthermore, all subjects completed a training before each evaluation session in order to familiarize themselves with the test procedure and to practice navigation in the virtual room.

8.10.4 Analysis and Test Results

Hypotheses The null hypothesis of the experiment was that the interaction tasks would have no effect on the users' ability of perceiving the reverberation time correctly. The alternative hypotheses were that 1-back working memory task and 2-back working memory task respectively weakened the users' ability of perceiving the reverberation time correctly, compared to the navigation only task. This means that subjects would evaluate inconsistently or would make incorrect ratings with an increasing difficulty of the task (degree of n-back task).

Test Panel A total of 21 subjects participated in the experiment. Seven of these participants were females and 14 males. The age of the subjects ranged from 21 to 36 years ($M = 26$, $SD = 3.7$). All participants were students or staff of Technische Universität Ilmenau. All reported to have normal hearing and normal or corrected to normal sight.

71% of the participants had already experience in subjective testing. Therefore they belonged to the category of initiated assessors. Only a minority of the panel could classify as naive assessors.

Data Analysis The obtained results were analyzed using SPSS for Windows version 13.0. The binomial sign test, a nonparametric test of analysis, was applied. The correctness of a subject's rating score was binomial: It was either right or wrong.

The ratings of each subject were analyzed in terms of their correctness. The correctness of the ratings was defined as follows: An answer was considered correct when in an evaluation session five rating scores for five reverberation times (from 1.0s to 3.0s) were given in an ascending order. The answer was considered incorrect whenever the five rating scores were not in ascending order (including two items receiving the same score). Table 8.4 gives one example for the definition of a correct answer (navigation task) and two examples for incorrect answers (1-back task, 2-back task).

Reverberation Time [s]	**1.0**	**1.5**	**2.0**	**2.5**	**3.0**	**Correct / Incorrect**
Navigation Task	3	23	49	89	96	Correct
1-back Task	3	49	23	89	96	Incorrect
2-back Task	3	49	49	89	96	Incorrect

Table 8.4: Analysis in terms of correctness: Examples for correct (navigation task) and incorrect (1-back and 2-back task) answers.

The percents of correct answers in perceiving reverberation time of the 21 participants in each condition of the experiment are shown in table 8.5. The data was analyzed using the binomial sign test and an alpha level of 0.05.

Condition	**Navigation**	**1-back Task**	**2-back Task**
Percent	0.71	0.57	0.43

Table 8.5: The percents of correct answers in perceiving reverberation time in navigation, 1-back working memory and 2-back working memory conditions.

It was tested whether subjects' ratings during 1-back and 2-back task included less correct answers than 0.71. Here, 0.71 was the binomial test hypothesized value from the navigation only condition. This value represents the control level. The participants' correct percent of 0.57 in the 1-back condition was not significantly different from the navigation only condition, $p = 0.125$ (one-tailed test). Yet, in the 2-back condition the participants' correct percent of 0.43 was significantly lower than in the navigation only condition, $p = 0.007$ (one-tailed test).

To find the cause of the result presented above (more errors in the 2-back condition than in the navigation only condition), a paired sample t test was used in a further analysis. This statistical method is used to compare two dependent samples, see fig. 8.8. Pairs of two adjacent reverberation times were grouped to check the difference of rating scores between them within a certain n-back condition, see table 8.6. Whenever the paired levels are

not significantly different the test subjects were not able to discriminate precisely between these reverberation times.

Reverberation Pair	Reverberation Time [s]
1	1.0 - 1.5
2	1.5 - 2.0
3	2.0 - 2.5
4	2.5 - 3.0

Table 8.6: The reverberation pairs consist of two adjacent reverberation times.

The results show that all reverberation pairs in navigation only and in 1-back condition were perceived as different. This difference was significant ($p < 0.01$). However, in the 2-back task, ratings of reverberation times $2.0s$ and $2.5s$ were not perceived as different ($t(20) = -2.08$, $p > 0.05$, ns). All other pairs in the 2-back condition were perceived as significantly different ($p < 0.01$). This indicates that subjects were unable to discriminate the reverberation times $2.0s$ and $2.5s$ in the 2-back task. This is what causes the increasing number of errors in the n-back task detected with the binomial test, see tables 8.5 and 8.7.

Absolute Number of Errors The absolute number of rating errors was dependent on the degree (difficulty) of the task. Whereas in the navigation only task the total number of errors per test subject was 0.29, in the 1-back task it was 0.48 and in the 2-back task it was 0.71, see table 8.7.

Condition	Navigation	1-back Task	2-back Task
Errors/Subject	0.29	0.48	0.71

Table 8.7: The absolute number of rating errors in navigation, 1-back working memory and 2-back working memory conditions.

As can be seen, the n-back task is a well suited method for scaling the amount of distraction. The correlation between distraction and degree (difficulty) of task might well be used in future applications to predict the level of divided attention in the user.

Subjects' Comments The majority of the test subjects claimed that the interaction tasks had a distracting effect. Especially the 2-back working memory task prevented them from performing the quality evaluation thoroughly. A general remark was that in most evaluating sessions they did not feel confident.

Finally, subjects were asked about the exact time at which they had usually made their decision on the perceived length of reverberation. This was asked to gain some knowledge about the different strategies test subjects had used to cope with both the interaction task and the opinion-forming process at the same time. Some subjects were able to develop strategies which allowed them to avoid the amount of distraction originally planned. For some participants the first number presented after the reference item was decisive for the rating. They claimed that at this moment they could remember the reference reverberation time very well. Other subjects neglected the working memory task for a short time to be able to concentrate on the formation of opinion. One subject claimed to have used the first part of the spoken number (a single-syllable word) to cope with the working memory task, and the decay part of the spoken word to evaluate the reverberation time.

8.10.5 Summary and Conclusions

In this experiment, subjects were asked to correctly rate the length of reverberation in comparison to a reference reverberation length while being distracted by an auditory n-back working memory task. Unlike in previously published experiments, both the parameter to be rated and the distracting task were related to the same modality. The analysis of the data obtained indicates that the precision with which auditory parameters can be rated by humans is dependent on the degree of distraction present in the same modality.

This result further confirms and specifies the findings of Zielinski, Kassier, Rumsey, Bech et al. in [zie03, kas03]: Whereas cross-modal division of attention only renders a small significant effect and apparently depends on the experimental conditions (apart from being listener-specific), with inner-modal distraction test subjects would predictably commit errors in their ratings.

If we conclude that these errors are related to the precision with which the stimuli are processed by the human perceptual system, then this effect might be exploited in future audiovisual applications: By offering attractive and interesting interactivity options equivalent to what in this experiment is called "task", a user could be distracted from the process of permanently rating the quality of a scene and scanning it for deficiencies in terms of scene realism, thus resulting in a higher overall quality impression.

8.11 Assessment: Influence of a Working Memory Task on Perceived Quality, II

The previous experiment indicated that a working memory task performed in the auditory modality can influence the ability of discriminating between auditory percepts of different quality (here: different lengths of reverberation). Unfortunately, the experiment was designed in such a way that subjects were able to develop strategies that allowed them to temporarily evade this influence. Therefore, a variation of the same experiment was implemented and performed. Subjects were motivated to concentrate on the n-back task by counting and displaying the errors they would commit. Additionally, the point in time at which the stimulus would be changed from reference to test item was randomized.

8.11.1 Test Setup

The quality attribute was the same as in the preceding experiment, see section 8.10.2: subjects were asked to identify the length of reverberation in comparison to a reference length. The stimulus material (visual scene: entrance hall of the main lecturing building at Technische Universität Ilmenau, audio material: spoken numbers ranging between *one* and *four*) was identical to the preceding experiment, too.

Technical Setup The experiment was conducted in the Listening Lab at Technische Universität Ilmenau. The loudspeaker setup used was the eight channel setup described in fig. 8.9 (left). The reproduction setup for the visual content was identical to the one described in section 8.8.1. Subjects were seated in a chair mounted on a platform of $0.40m$ of elevation during the experiment, such that their eyes were at the same height as the center of the screen. Loudspeakers were mounted on loudspeaker stands at the height of the subjects' ears.

8.11.2 Implementation of the Assessment, Task and Rating Scale

Test Method A test method based on the Degradation Category Rating (DCR), specified in Recommendation ITU-T P.911 [itu911], was chosen to compare a modified item with a reference item in terms of length of reverberation time (RT). The test items varied in five different reverberation times, with 1.0s, 1.5s, 2.0s, 2.5s and 3.0s. The reference item consisted of the medium length of reverberation time (2.0s).

Within a sequence of spoken numbers with a total length of 30 seconds, the length of reverberation slope was varied at a random point in time. In the beginning of a trial always the reference item was played. During the presentation the length of reverberation was changed: the audio presentation was switched to one of the test items, switched to the anchor item, or the reference item was kept unchanged during the transition phase. The exact point in time of the transition between the reference and the test item was changed randomly. Transition took always place in an area of transition (see fig. 8.29) which began after the first two numbers and ended before the penultimate number of the sequence. Therefore each trial always began with the reference item and ended with the item to be rated by the subjects.

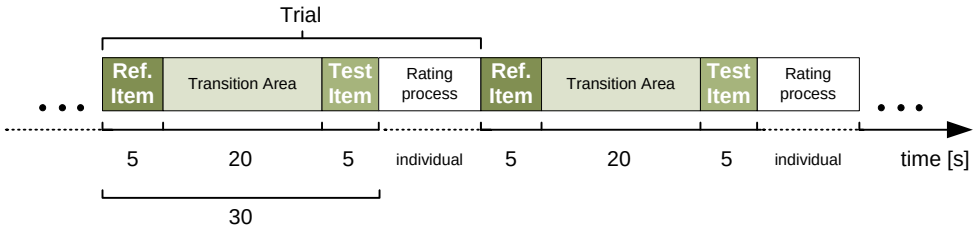


Fig. 8.29: Chronological sequence of the test procedure (Modified Degradation Category Rating) used to present the stimuli.

Rating Scale After the presentation of a sequence of numbers, subjects had to rate the perceived length of reverberation time of the modified item in relation to the reference item on a five-level scale. The scale values and the semantic identifiers of the rating scale in English as well as the German translations actually used in the experiment are listed in table 8.8. Subjects were instructed only to use these levels and no values in between.

Scale value	English identifier	German identifier
100	much longer	viel länger
75	longer	etwas länger
50	equal	gleich
25	shorter	etwas kürzer
0	much shorter	viel kürzer

Table 8.8: Five-level scale with scale values, English identifier and German language translated identifier as used in the experiment.

Test Procedure The experiment was divided into an anchoring phase and three evaluation sessions with different degrees of the interaction task: “navigation only”, “navigation with 1-back working memory task” and “navigation with 2-back working memory task”.

These were the same as described in the previous experiment, see section 8.10.1. The process of quality evaluation (rating, see fig. 8.29) was kept the same throughout the whole assessment.

Navigation consisted in actively moving around the virtual room by using a computer mouse, in the same way as described in the previous experiment.

Before the experiment, subjects were presented with written instructions. These included descriptions of the test procedure, the rating method, the attribute to be evaluated and the anchoring process. In case of subjects' questions, additional information was given orally by the experimenter.

Subsequently, subjects could familiarize themselves with the test items presented in the evaluation sessions and the rating scale while watching a static picture of the visual scene in an anchoring process. For this, five buttons were provided on the Input Device which were labeled according to the levels of the rating scale. Upon the press of a button, a spoken number with the corresponding reverberation time was played. In addition, the spoken number could be varied between 1, 2, 3 and 4 by pressing the "number" button. The duration of the anchoring process was individually controlled by each test subject. Fig. 8.30 shows the control panel layout of the Input Device used in this assessment.

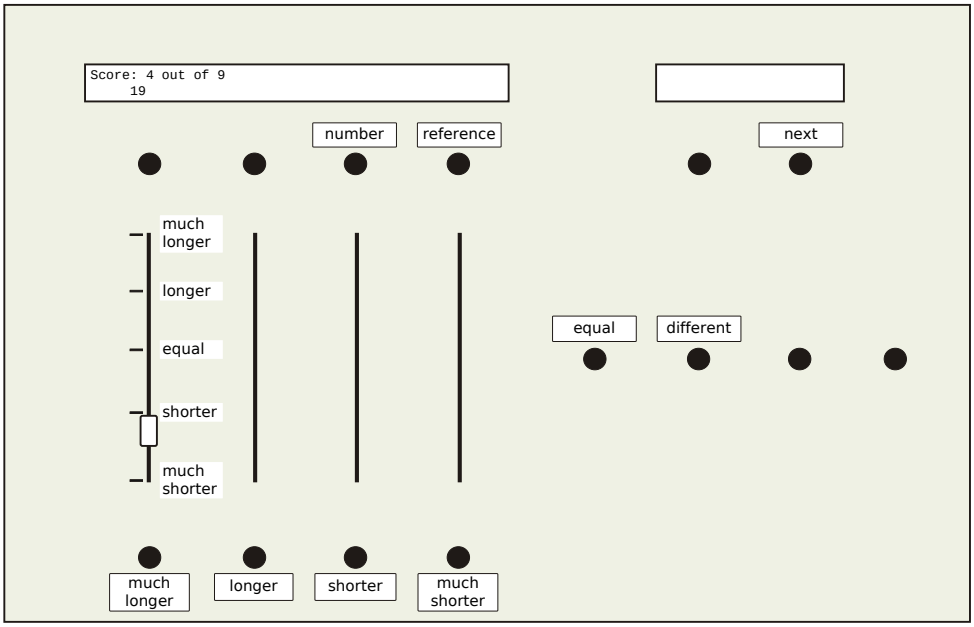


Fig. 8.30: Schematic view of the panel layout of the Input Device as used in the second divided attention / working memory assessment.

After the anchoring, three evaluation sessions followed which were presented in random order. Instructions and a training session preceded each of these sessions. Subjects were briefed in the instructions to concentrate on the reverberation length, the navigation and the n-back task. Each correct answer in the n-back task increased their score by one point. Their goal was to reach the highest possible score. In the training session, test subjects could practice both the n-back task and the navigation with the computer mouse.

In each evaluation session subjects performed eight trials. The presentation order of trials was randomized between the subjects. Because of time constraints only the ratings of the reverberation times of $1.5s$, $2.0s$ and $2.5s$ were presented twice. The duration of the experiment totaled in 40 minutes per test subject. Fig. 8.31 shows the schematic procedure

of the experiment.

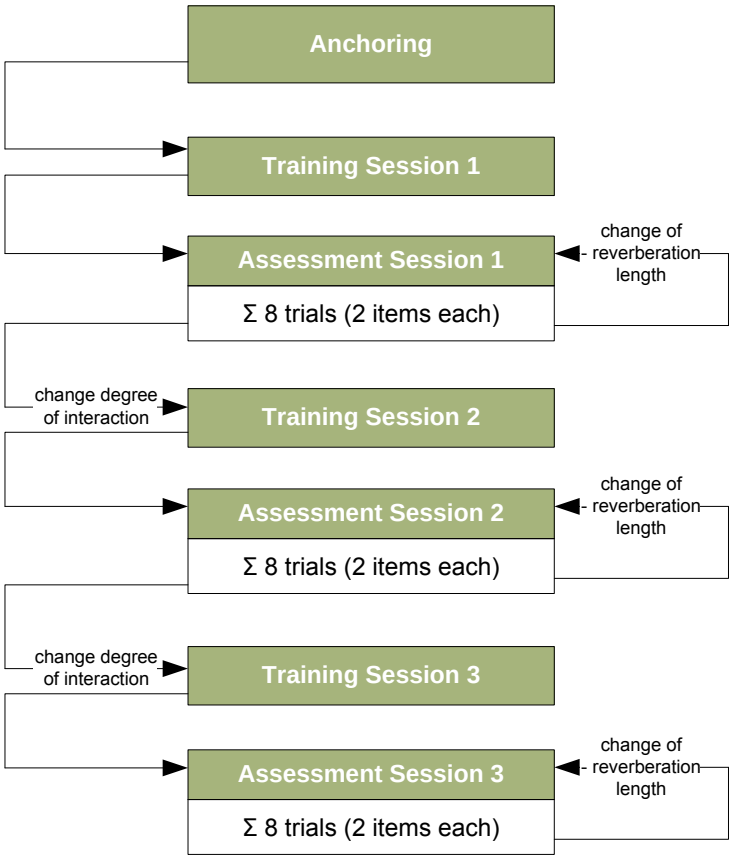


Fig. 8.31: Example procedure of the second working memory task experiment. The evaluation sessions were presented in random order.

8.11.3 Analysis and Test Results

Hypotheses The null hypothesis was that the interaction tasks would have no effect on the users' ability of perceiving the reverberation time correctly. The alternative hypothesis was that with an increasing degree of the interaction task (from "navigation only", to "navigation with 1-back task", to "navigation with 2-back task"), subjects would become more unconfident with evaluating reverberation differences and make incorrect ratings.

Test Panel 21 test subjects participated in the experiment. The majority of the subjects were students and scientific staff of Technische Universität Ilmenau. Five of the participants were females and 16 males. Subjects' age had the mean value $M = 25.1$ with a standard deviation $SD = 3.85$. All reported to have normal hearing and normal or corrected to normal sight. Regarding the listening experience, six of the subjects could be categorized as initiated assessors and 15 subjects as naive assessors. The group of initiated assessors had already participated in preceding unimodal and bimodal subjective assessments in which they had gained abilities and knowledge in rating the quality of auditory displays.

Data Analysis I The ratings of each test subject were transformed into correct and incorrect answers and were statistically analyzed using nonparametric test procedures. Two methods were used for the error analysis:

1. In the first analysis (variant 1), a rating was considered incorrect whenever the expected scale level was not met. E.g. if the reverberation time was rated as being 2.0s by the subject, but actually was only 1.5s, then this was considered an error.

2. To reduce the amount of random incorrect ratings in the second analysis (variant 2), a tolerance range of one scale level was introduced. Thus a rating was considered incorrect whenever it was wrong by two scale levels or more: e.g. if the reverberation time was rated as being 2.0s by the subject, but actually was only 1.5s, then this was *not* considered an error in variant 2. The rating was still within the tolerance range.

The error analyzes resulted in the relative frequency of incorrect ratings which are shown in fig. 8.32 and fig. 8.33. Generally, longer reverberation times were more difficult to rate than shorter reverberation times.

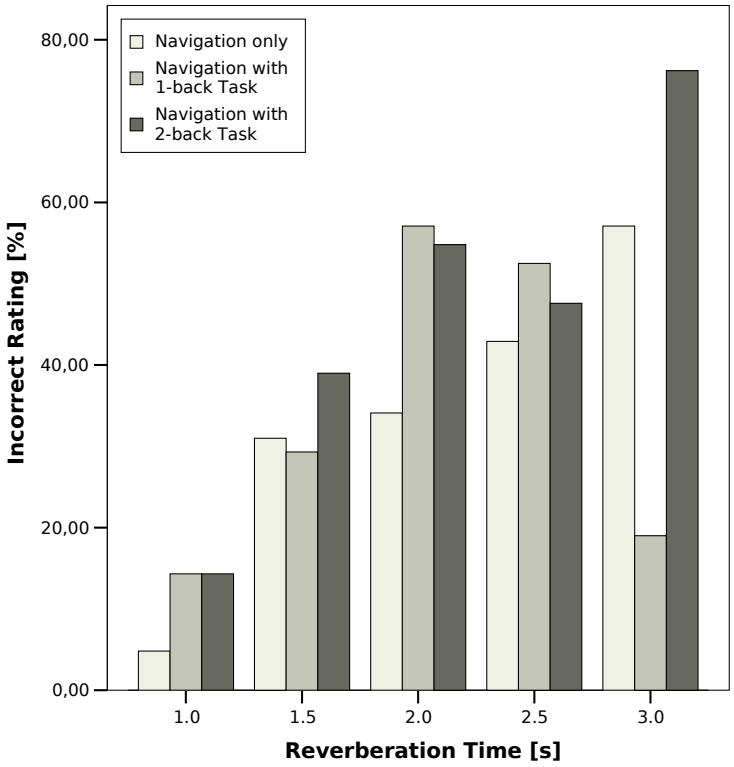


Fig. 8.32: Incorrect ratings across all subjects for different reverberation times using error analysis variant 1 (a rating was considered incorrect when the expected scale level was not met).

The incorrect ratings of all reverberation times in variant 1 and variant 2 were summarized respectively to a mean value (see table 8.9) and tested for significant differences among each other.

In the previous experiment described in section 8.10, the binomial test was used to verify significances. This statistical test was also applied in the further analysis of this assessment's data. The binomial sign test delivered a significant difference for the distribution

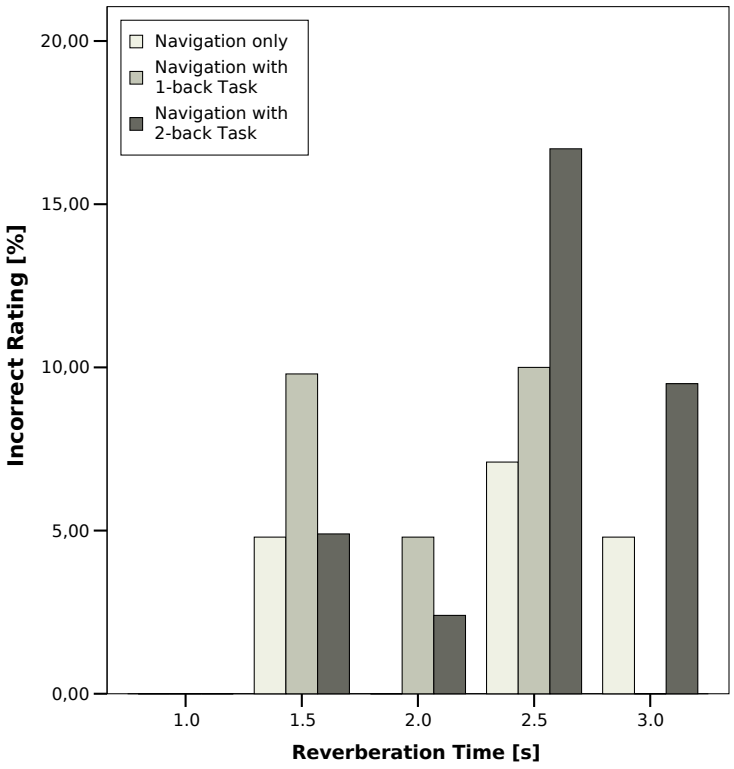


Fig. 8.33: Incorrect ratings across all subjects for different reverberation times using error analysis variant 2 (a rating was considered incorrect when it was wrong by two scale levels or more). Note that compared to fig. 8.32 the scale is enlarged for clarity.

	Variant 1	**Variant 2**
Navigation only	34.7 %	3,6 %
Navigation with 1-back task	38,8 %	6,1 %
Navigation with 2-back task	46,7 %	7,2 %

Table 8.9: Incorrect ratings summarized for all reverberation times in error analysis variant 1 and variant 2.

of correct and incorrect ratings in both error analyzes between the "navigation only" and "navigation with 2-back task" ($p < 0.025$)[17]. The results of the binomial sign test for variant 1 and 2 of the error analysis performed with SPSS can be found in table 8.10 and table 8.11.

The total percentage of errors for the n-back tasks was, compared to a similar assessment described in section 8.10, between 5% and 10% lower (variant 1, compare tables 8.9 and 8.5). This can be attributed to a learning effect in the test subjects: a quarter of the sample (five subjects) participated in both assessments. Actually, in both n-back tasks these five subjects showed 1/4 fewer errors than the rest of the subjects.

[17] Here, the significance level alpha was lowered from 0.05 to 0.025 because two comparisons were made ("navigation only" vs. "navigation with 1-back task" and "navigation only" vs. "navigation with 2-back task"). Thus, the significance level was halved to avoid capitalizing on chance, see section 8.6.2 and [coo07].

Binomial Test

		Category	N	Observed Prop.	Test Prop.	Asymp. Sig. (1-tailed)
Navigation with 1-back Task	correct	,00	101	,61	,66	,113[a,b]
	incorrect	1,00	64	,39		
	Total		165	1,00		
Navigation with 2-back Task	correct	,00	93	,56	,66	,004[a,b]
	incorrect	1,00	74	,44		
	Total		167	1,00		

a. Alternative hypothesis states that the proportion of cases in the first group < ,66.

b. Based on Z Approximation.

Table 8.10: SPSS results of binomial test for variant 1 of error analysis.

Binomial Test

		Category	N	Observed Prop.	Test Prop.	Asymp. Sig. (1-tailed)
Navigation with 1-back Task	correct	,00	155	,939	,964	,076[a,b]
	incorrect	1,00	10	,061		
	Total		165	1,000		
Navigation with 2-back Task	correct	,00	155	,928	,964	,018[a,b]
	incorrect	1,00	12	,072		
	Total		167	1,000		

a. Alternative hypothesis states that the proportion of cases in the first group < ,964.

b. Based on Z Approximation.

Table 8.11: SPSS results of binomial test for variant 2 of error analysis.

Data Analysis II In order to further substantiate the results from the binomial sign test, the data was re-organized and re-scored such that a small error of one scale level was assigned one error point, whereas larger errors of two or three scale levels were awarded with two or three error points, respectively. This was possible because in the modified test design (compare with section 8.10) a five-level scale was used and subjects were instructed only to use these levels. This ranking system is favorable because it gives "much more subtle data and a difference will emerge as significant if it's there" [coo07].

After re-scoring the data, reliability of the test subjects was checked by comparing the scores for the repeated items ($1.5s$, $2.0s$, and $2.5s$). One subject was excluded because he showed inconsistent ratings in all three conditions. This reduced the sample size to 20. Normal distribution could not be assumed, such that parametric tests like ANOVA could not be used. Instead, Wilcoxon's T test (matched pairs signed ranks, see fig. 8.8) was used to evaluate the differences between the three conditions "navigation only", "navigation with 1-back task", and "navigation with 2-back task". The Friedman test was not used here because Friedman only evaluates a total difference over all conditions. The assessment here was designed to test the effect of split attention over a range of scaled perceptual loads. Thus, the evidence is contained only within the difference between "navigation with n-back task" condition and "navigation only" condition. The actual load was an arbitrary value based on the restrictions associated with the scalability of the n-back task: n can only be a natural number. Therefore, there is only limited information contained in the difference between "navigation with 1-back task" condition and "navigation with 2-back task" condition; it was evaluated only for the sake of completeness.

As can be seen from table 8.12, the results collected in the binomial test were certified. The difference between "navigation only" and "navigation with 1-back task" was not significant ($T = 63.5$, $p > 0.05$, ns), as well as the difference between "navigation with 1-back

Test Statistics[b]

	ErrorPoints1Back - ErrorPointsNavigation	ErrorPoints2Back - ErrorPointsNavigation	ErrorPoints2Back - ErrorPoints1Back
Z	-,622[a]	-2,718[a]	-1,266[a]
Asymp. Sig. (2-tailed)	,534	,007	,206

a. Based on negative ranks
b. Wilcoxon Signed Ranks Test

Table 8.12: SPSS results (test statistics) of the Wilcoxon's T test.

task" and "navigation with 2-back task" ($T = 64$, $p > 0.05$, ns). Yet, a highly significant difference was shown for the "navigation only" condition vs. the "navigation with 2-back task" condition ($T = 20$, $p \leq 0.01$). The total of the ranks where subjects committed more errors in the "navigation with 2-back task" condition was 133 and the total for the "navigation only" condition was 20, see table 8.13.

Ranks

		N	Mean Rank	Sum of Ranks
ErrorPoints1Back - ErrorPointsNavigation	Negative Ranks	7[a]	9,07	63,50
	Positive Ranks	10[b]	8,95	89,50
	Ties	3[c]		
	Total	20		
ErrorPoints2Back - ErrorPointsNavigation	Negative Ranks	5[d]	4,00	20,00
	Positive Ranks	12[e]	11,08	133,00
	Ties	3[f]		
	Total	20		
ErrorPoints2Back - ErrorPoints1Back	Negative Ranks	6[g]	10,67	64,00
	Positive Ranks	13[h]	9,69	126,00
	Ties	1[i]		
	Total	20		

a. ErrorPoints1Back < ErrorPointsNavigation
b. ErrorPoints1Back > ErrorPointsNavigation
c. ErrorPoints1Back = ErrorPointsNavigation
d. ErrorPoints2Back < ErrorPointsNavigation
e. ErrorPoints2Back > ErrorPointsNavigation
f. ErrorPoints2Back = ErrorPointsNavigation
g. ErrorPoints2Back < ErrorPoints1Back
h. ErrorPoints2Back > ErrorPoints1Back
i. ErrorPoints2Back = ErrorPoints1Back

Table 8.13: SPSS results (ranks) of the Wilcoxon's T test.

By looking at the recorded score during the 1-back task and the 2-back task, the active involvement of the test subjects can be verified. The scores obtained across all subjects against the duration of an evaluation session are shown in fig. 8.34 and fig. 8.35. The highest possible score was nine in the 1-back task and eight in the 2-back task. The Pearson correlation between the duration of an evaluation session (trial) and an increasing score in the 2-back task was found to be significant ($r(6) = 0.63$, $p < 0.05$). Again, this is a clear indicator of a learning effect.

8.11.4 Summary and Conclusions

A subjective assessment was performed to evaluate the effect of user interaction (in the shape of an auditory n-back task) upon the perceived length of reverberation time. It was hypothesized that with an increasing degree of the interaction task (from "navigation only", to "navigation with 1-back task", to "navigation with 2-back task") subjects would become more unconfident with evaluating reverberation time differences and make more incorrect ratings.

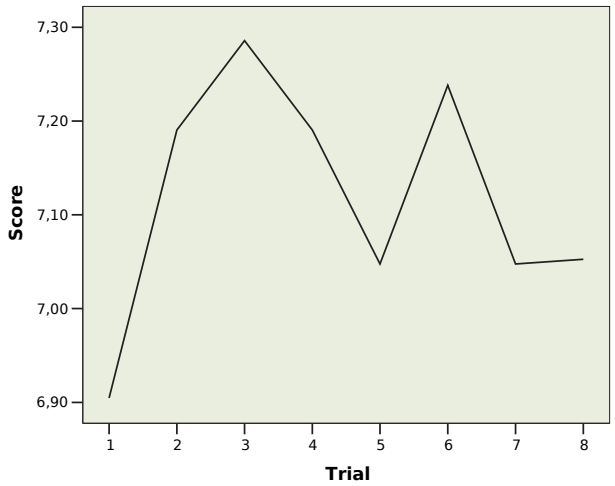


Fig. 8.34: Mean score in the 1-back task against the duration of the evaluation session (trials 1-8). The highest possible score was nine.

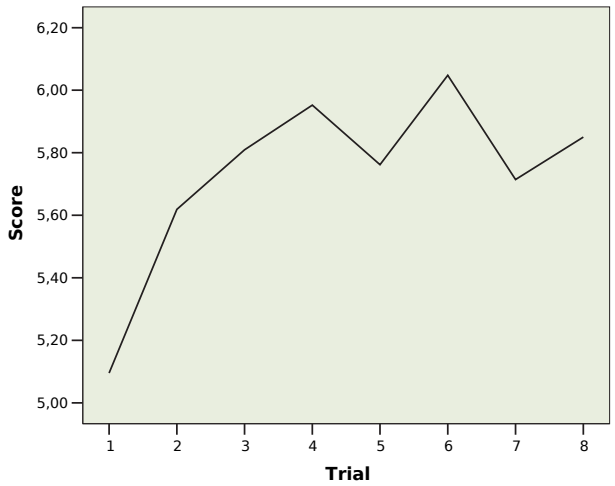


Fig. 8.35: Mean score in the 2-back task against the duration of the evaluation session (trials 1-8). Here, the highest possible score was eight.

This was verified using two methods of data analysis, the binomial sign test and Wilcoxon's T for related data. For the higher degree of distraction ("navigation with 2-back task") a significant effect was observed for both methods[18]. The results of the assessment indicate that an additional working memory task taking place in the same modality as the main target can have a significant influence upon the ability to discriminate between different qualities of percepts - if the work load is high enough. This, combined with the results of section 8.9, supports the claim that capacity limits appear to be more severe when multiple stimuli are presented in the same modality compared with multiple modalities (see section 2.5.4).

These findings suggest the conclusion that inner-modal division of attention can disturb the discrimination of perceived quality - if only the attention required by the secondary

[18] Although it might seem obvious that the method should not influence the result of the analysis (acceptance or rejection of the hypothesis), in practice this is not necessarily true. Because the two methods used here are based on different levels of data (nominal vs. ordinal data), they do not provide the same sensitivity to the effect under test. Thus, the results may very well differ without one being plain "wrong".

task is large enough to reach capacity limits.

Audiovisual assessment data very often does not fulfill all of the ANOVA criteria, such that non-parametric test methods have to be used. The advantage of these methods is that the dependent variables not necessarily have to be normally distributed, thus requiring a lower sample size. They are equally powerful but require a thorough test design.

The learning effect that was observed within this assessment, as well as between this and the preceding assessment, indicates that Neisser's *Perceptual Cycle* as described in section 2.5.2 is in fact a valid model describing the mechanism of human perception qualitatively. Beyond that, no quantitative conclusions can be drawn from the data obtained.

8.12 Assessment: Influence of Visual Interaction on Perceived Audio Quality

This experiment investigates the possibility of cross-modal influence of interaction upon perceived quality. Whereas in the previous two assessments the influence of interaction was checked within the same modality, here the influence of a visual (-motion) task upon the perceived audio quality was evaluated. This experiment is borrowing from what Zielinski et al. [zie03] and Kassier et al. [kas03] have described, but the test panel is significantly larger (31 test subjects opposed to 6 and 7, respectively), thus allowing a profound statistical analysis.

8.12.1 Test Setup

The assessment took place in the Listening Lab at Technische Universität Ilmenau. Reproduction of the monophonic audio signal was done via two active, full range monitor loudspeakers located behind an acoustically transparent projection screen at the $\pm 15°$ positions, see fig. 8.36. The SPL at the listener's position was measured to be $76dB$ during the experiment.

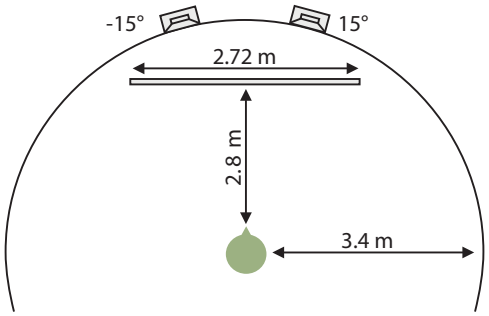


Fig. 8.36: Test setup with the position of the listener, the loudspeakers and the projecting screen.

Subjects were seated in a chair mounted on a platform of $0.40m$ of elevation during the experiment, such that their eyes were at the same height as the center of the screen. Loudspeakers were mounted on loudspeaker stands at the height of the subjects' ears. The visual reproduction setup was identical to the setup described in section 8.8.1, see also fig. 8.36.

8.12.2 Implementation of the Assessment, Task and Rating Scale

Interactive Scenario A computer game based on the MPEG-4 BIFS format was created using the ECMA script language and the scene control mechanisms provided by MPEG-4, see section 5.3. It was designed to assess the effect of divided attention in the evaluation of audio quality during involvement in a visual task. In this game two different types of objects moved through the virtual room in random directions: donuts and snowballs. Fig. 8.37 shows a screenshot. Subjects had to collect selected flying objects (donuts) by running into them and avoid the collision with other objects (snowballs). For the navigation, test subjects used the left and right arrow keys of a computer keyboard. Movement was only possible to the sides at a fixed distance to the wall on the other end of the room.

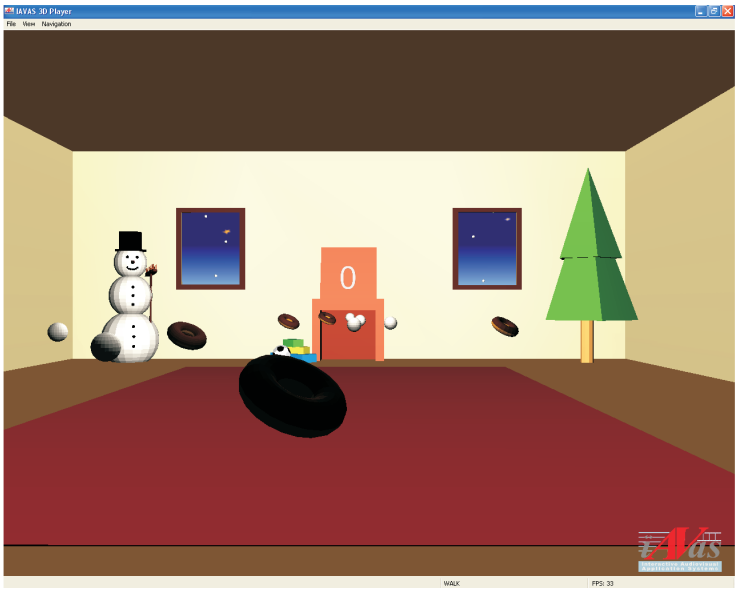


Fig. 8.37: Screenshot showing the interactive game scenario in the experiment. Subjects had to catch the snowballs and evade collisions with the donuts.

A game score was recorded for each subject to verify subjects' involvement in the game and to prod the subjects to actively play the game. By collecting the right object (donut) the score was increased by one point, whereas a collision with a snowball decreased the score by one point. The actual game score was displayed in the visual scene near the source of the flying objects.

For the experiment, each subject carried out a passive and an active session. The active session consisted in playing the computer game and evaluating the audio quality. This session was designed to cause a division of attention between the rating of the audio quality and the involvement in a computer game. 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 directed to the auditory display.

Auditory Scenario A typical background music for a computer game was chosen for the auditory presentation during the game. The audio quality degradations were realized by modifying the tonal quality. The original signal ($16kHz$) was low-pass filtered using three different cut-off frequencies $f_c = 11kHz$, $12kHz$ and $13kHz$. Additionally, an anchor with a low-pass filtering at the cut-off frequency $f_c = 4kHz$ was created. Thus three test items,

one anchor item and a reference item (corresponding to the original full range signal) were presented to the test subjects in the experiment.

The experiment was performed using a test method (see fig. 8.38) based on the Degradation Category Rating (DCR), which is standardized in recommendation ITU-T P.911 [itu911]. In the beginning of a trial the reference item was played. During this presenta-

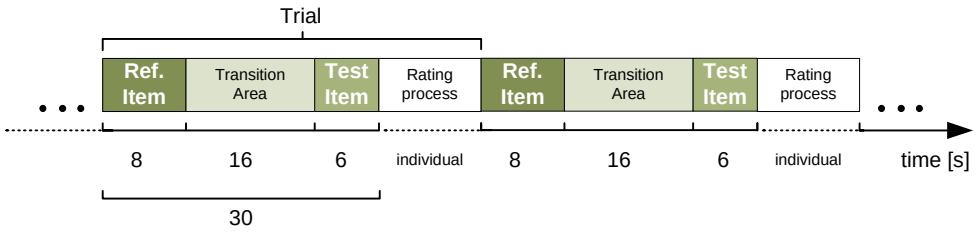


Fig. 8.38: Chronological sequence of the test procedure used to examine a possible influence of visual interaction upon perceived audio quality. A modified Degradation Category Rating was adopted to present and rate the audio material.

tion the tonal quality was changed: the audio presentation was switched to one of the test items, switched to the anchor item, or the reference item was kept unchanged during the transition phase. The exact point in time of the transition between the reference and the test item was changed randomly. Transition took always place in an *area of transition* (see fig. 8.38) which began after the first eight seconds and ended before the last six seconds of the game. Therefore each trial always began with the reference item and ended with the item to be rated by the test subject.

Input Device Fig. 8.39 shows a schematic view of the Input Device used during the assessment. In the anchoring phase of the experiment subjects could use the two buttons 'imperceptible' and 'very annoying' to familiarize themselves with the quality extremes. During the experiment, the 'start' button was pressed to start a trial. By sliding the fader into the corresponding vertical position, the test subject could make a rating on the perceived quality and transmit it to the system by pressing the 'rating' button. The motorized fader was automatically moved into a neutral position before each trial.

Rating Scale After the presentation of the bimodal stimulus (at the end of a gaming trial), subjects had to rate the perceived tonal quality degradation using the standardized ITU-T P.911 [itu911] five-level impairment scale. The scale values and the semantic identifiers of the rating scale as well as the German translations actually used in the experiment are listed in table 8.14.

Scale value	Standardized identifier	German identifier
100	imperceptible	nicht wahrnehmbar
75	perceptible, but not annoying	wahrnehmbar, aber nicht störend
50	slightly annoying	etwas störend
25	annoying	störend
0	very annoying	sehr störend

Table 8.14: Five-level impairment scale with scale values, standardized identifier and German language translated identifier as used in the experiment.

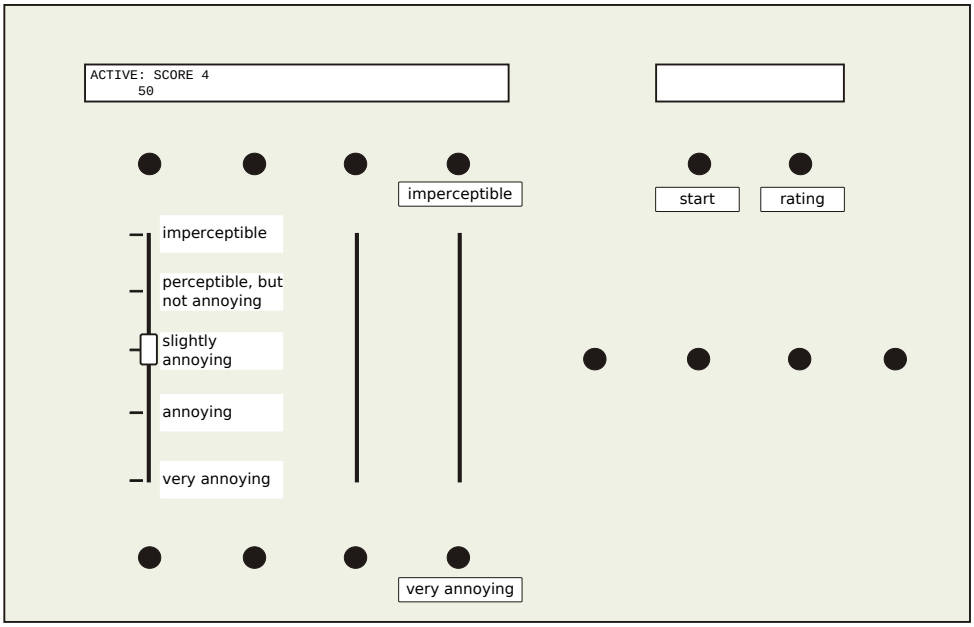


Fig. 8.39: Schematic view of the Input Device used in the assessment.

Test Procedure It the beginning of the experiment subjects were presented written instructions which included descriptions of the test procedure, the rating method and the attribute to be evaluated. In case of subjects' additional questions, these were answered by the test supervisor.

Subsequently, subjects could familiarize themselves with the reference item and the anchor item in an anchoring process. The duration of the anchoring was individually controlled by each subject.

After this the passive and active test sessions followed, see fig. 8.40. The three different test items were presented four times each, the anchor and reference items two times each, resulting in a total of 16 trials per session. The order of appearance was changed randomly. Instructions and a training session preceded each session. Subjects were briefed in the instructions to concentrate on the game in the active session and to concentrate on the auditory display in the passive session. The training was used to familiarize subjects with the test method, as well as to practice the lateral movement with the arrow keys of the keyboard in the gaming session.

8.12.3 Analysis and Test Results

Hypotheses The null hypothesis was that a test subject playing an interactive audiovisual game would perceive degradations in the audio quality as being equally disturbing as when only passively watching a presentation of the same game. The alternative hypothesis was that the distraction caused by actively playing the game would decrease the perceived audio quality degradation across the modalities in comparison to passively watching the presentation.

Test Panel A total of 32 subjects participated in the experiment. The majority of the participants were students and scientific staff of Technische Universität Ilmenau. Seven of the participants were females and 25 males (age $M = 25.7$, $SD = 5.36$). Regarding

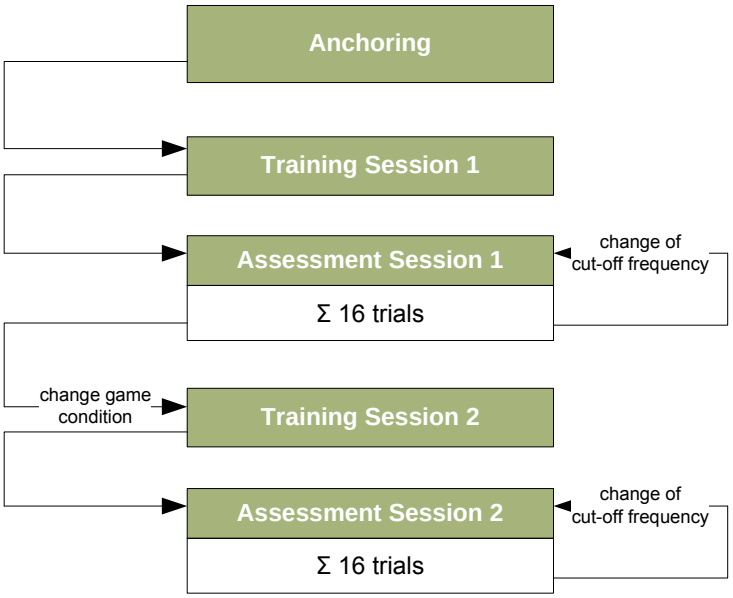


Fig. 8.40: Example procedure of the cross-modal division of attention experiment. The succession of active and passive sessions was determined at random.

the listening experience, 20 subjects belonged to the category of *initiated assessors* and 12 subjects classified as *naive assessors*. The group of initiated assessors had already gained abilities and knowledge in rating the quality of auditory displays in preceding unimodal and bimodal subjective assessments. All participants reported normal hearing and normal or corrected to normal visual acuity.

Data Analysis The standard deviation between repetitions of identical items in the passive session (*No Game* condition) served as an indicator for the reliability of the test subjects. A large standard deviation is usually an indicator for an unreliable subject. The mean standard deviation across all subjects was found to be half a step on the five-level impairment scale ($SD = 11.2$ scale values, see table 8.14). No test subjects had to be excluded from the analysis.

The results of all subjects were summarized for the different cut-off frequencies in the passive session (*No Game* condition) and the active session (*Game* condition). The resulting quality ratings are given in fig. 8.41.

The bar chart shows that the perceived quality of items with the cut-off frequencies $f_c = 4kHz$, $11kHz$, $12kHz$ and $13kHz$ on average received better audio quality scores in the active session (*Game* condition) than in the passive session (*No Game* condition).

This tendency was analyzed using tests of significance. Because the Kolmogorov-Smirnov test shows a significant departure from normality ($p < 0.05$), nonparametric tests of analysis were applied. For the cut-off frequency $f_c = 12kHz$ the 95% confidence intervals of passive (*No Game* condition) and active (*Game* condition) items overlap. In normally distributed data this is an indicator for not significant variances. As we are dealing with not-normally distributed ratings the confidence intervals presented are not a reliable criterion.

Instead, the Wilcoxon T test which compares two dependent samples was applied. The Wilcoxon T test showed that the quality ratings of the active session vary significantly from the ratings of the passive session for cut-off frequencies up to $12kHz$. A significant

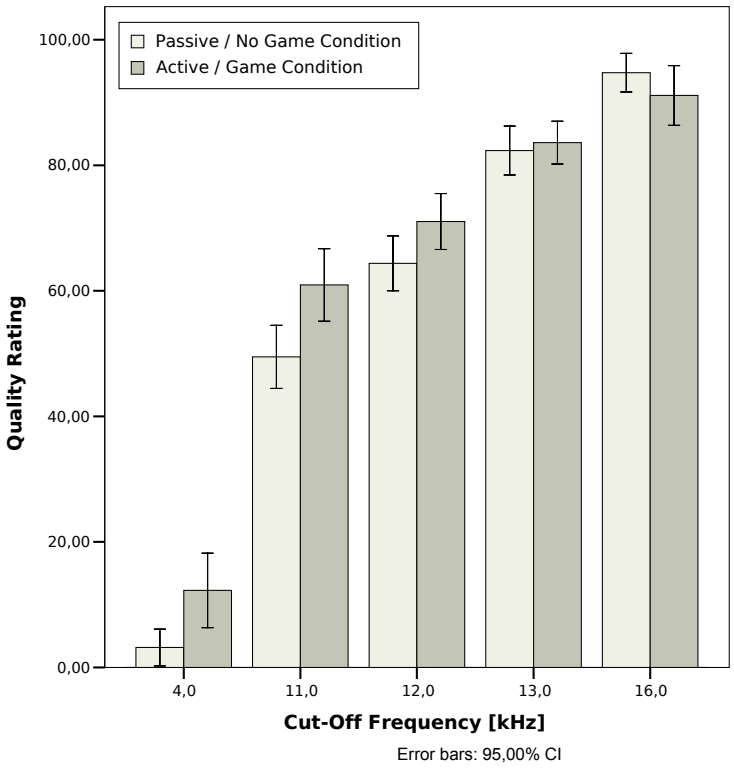


Fig. 8.41: Audio quality ratings for passive session (*No Game* condition) and active session (*Game* condition) for different cut-off frequencies (error bars show 95.0% confidence interval of mean).

decrease in rating correctness was shown for the *Game* condition in comparison to the *No Game* condition for the anchor item ($T = 37$, $p \leq 0.01$), the cut-off frequency $f_c = 11kHz$ ($T = 452.50$, $p \leq 0.01$), and the cut-off frequency $f_c = 12kHz$ ($T = 812$, $p \leq 0.01$), see table 8.15.

For the cut-off frequency of $13kHz$ and the reference item, no significant differences were found ($T = 630.50$ and $T = 75$, resp., $p > 0.05$, ns).

Fig. 8.42 shows the rating differences between the active (*Game* condition) and the passive (*No Game* condition) session for different cut-off frequencies. What can be seen is that subjects in general rated the perceived audio quality degradation as less perceptible in the active session (*Game* condition) than in the passive session (*No Game* condition). Fig. 8.43 shows that the majority of the subjects rated equal audio quality higher in the

Test Statistics[c]

	Anchor (Game) - Anchor (No Game)	Item 11 kHz (Game) - Item 11 kHz (No Game)	Item 12 kHz (Game) - Item 12 kHz (No Game)	Item 13 kHz (Game) - Item 13 kHz (No Game)	Reference (Game) - Reference (No Game)
Z	-2,628[b]	-3,541[b]	-2,623[b]	-,552[b]	-1,161[a]
Asymp. Sig. (2-tailed)	,009	,000	,009	,581	,246

a. Based on positive ranks.

b. Based on negative ranks.

c. Wilcoxon Signed Ranks Test

Table 8.15: SPSS generated results of the Wilcoxon test.

active session (*Game* condition) than in the passive session (*No Game* condition). Yet, there were also some subjects which, when involved in the game, rated the audio quality as being lower.

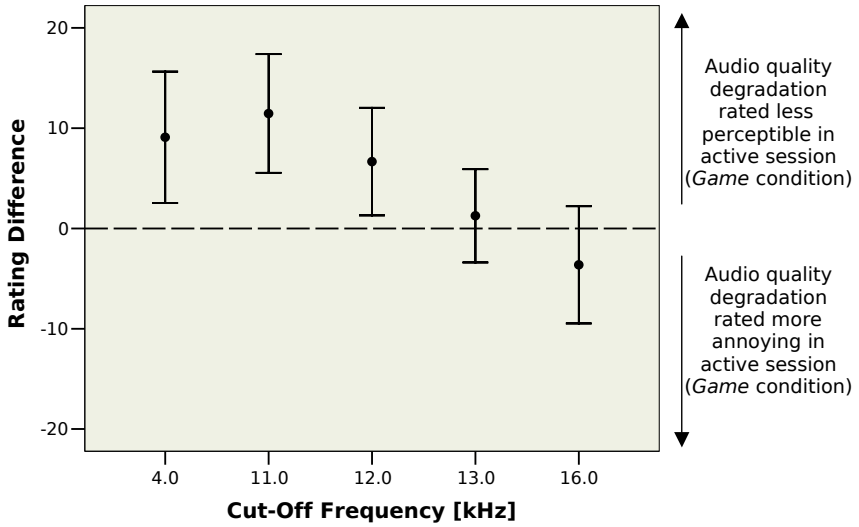


Fig. 8.42: Rating differences between the active and the passive session for different cut-off frequencies (error bars show 95.0% confidence interval of mean). The $4kHz$ entry is the anchor item, the $16kHz$ entry denotes the reference item.

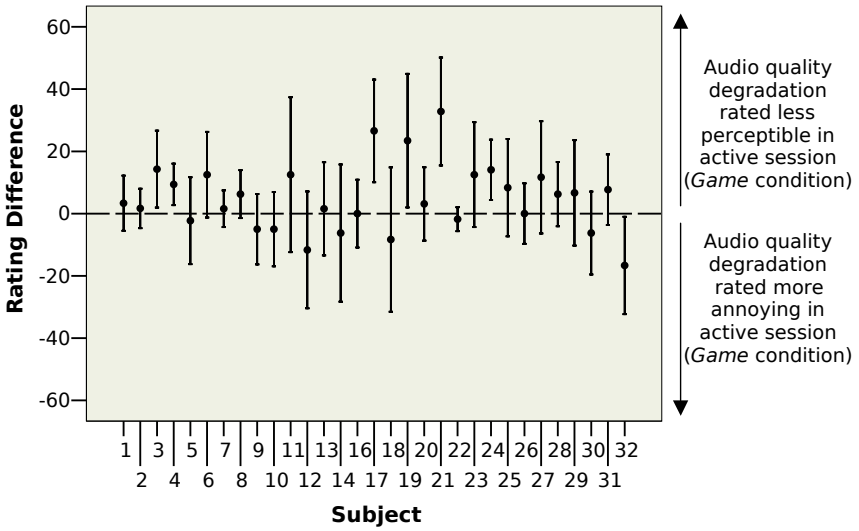


Fig. 8.43: Rating differences between the active and the passive session for individual subjects (error bars show 95.0% confidence interval of mean).

The score across all subjects in each trial is shown in fig. 8.44. The Pearson correlation between the duration of an evaluation session (trial) and an increasing game score was highly significant ($r(16) = 0.64, p \leq 0.01$). The SPSS results of Pearson's test of correlation are summarized in table 8.16. Two additional trials containing the reference item were added in the active session (*Game* condition) before and after the evaluation (trials 1 and 18, see fig. 8.44) to verify a learning effect in the game playing. No rating scores were recorded for these trials.

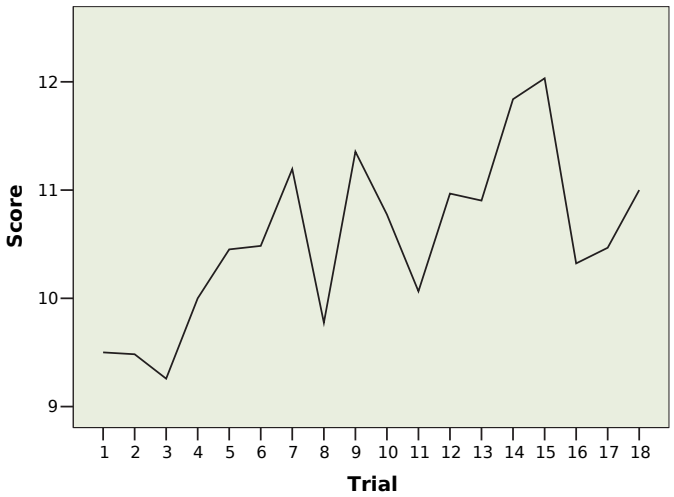


Fig. 8.44: Mean game score against the duration of the experiment.

Correlations

		Score	Trial
Score	Pearson Correlation	1	,635**
	Sig. (1-tailed)		,002
	N	18	18
Trial	Pearson Correlation	,635**	1
	Sig. (1-tailed)	,002	
	N	18	18

**. Correlation is significant at the 0.01 level (1-tailed).

Table 8.16: Results of correlation test between the duration of the experiment (trial) and an increasing game score.

A correlation between the game score and a high rating difference between the active (*Game* condition) and the passive (*No Game* condition) session could not be substantiated using Pearson's correlation calculation ($r(30) = 0.019$, $p > 0.05$, ns, see table 8.17).

8.12.4 Summary and Conclusions

The statistical analysis shows that the ratings of the tonal quality degradations in the active session differ from those in the passive session. The low-pass filtering in the active session (*Game* condition) was rated as being less perceptible. This effect was found to be significant at the cut-off frequencies of $f_c = 4kHz$, $11kHz$ and $12kHz$. There is also a

Correlations

		Score	Rating Difference
Score	Pearson Correlation	1	-,019
	Sig. (1-tailed)		,460
	N	32	31
Rating Difference	Pearson Correlation	-,019	1
	Sig. (1-tailed)	,460	
	N	31	31

Table 8.17: Results of correlation test between a high rating difference between the active and the passive session and a high game score.

difference between the ratings in the active and the passive session at the cut-off frequency of $f_c = 13kHz$, but an effect towards an increased perceived quality was not found to be significant. Probably this filtering was not readily discriminable from the original signal with $f_c = 13kHz$ vs. $f_c = 16kHz$, which is also suggested by the overlapping 95% confidence intervals of the two items in the active session (Game condition), see fig. 8.41. Apparently, in this frequency range the influence of the screen upon the stimulus starts to take effect. This is further substantiated by the measurement of the screen's frequency transmission curve, see fig. 8.7.

Nonetheless, the experiment shows that an influence of interaction performed in one modality (visual-haptic) upon another modality (here: auditive) is possible. Thus, cross-modal influences exist, suggesting that the perceptual model by Hollier and Voelcker (see section 2.5.5) is a valid qualitative approximation.

By recording the game score it was possible to verify the active involvement of the subjects in the computer game. The increasing game score over time (duration of the whole experiment) indicates a learning effect in the subjects, see fig. 8.44. Again, this supports Neisser's model of the *Perceptual Cycle*.

8.13 A Mixed Method Approach

Section 8.9 reported an assessment on the influence of interaction on perceived quality, in which both quantitative and qualitative methods were applied. A first effort was made to combine the two methodologies, producing quantified results without losing information about test subjects' rationale.

Yet, it was noticed that going through all recorded interviews, transcribing them and extracting keywords which could only then be grouped and analyzed to form the resulting categories, is not a viable method for large sets of data. This is even more true when taking into account that transcription and coding errors do occur, such that the same work needs to be done by at least two people independently. Only this way coding reliability can be determined. Therefore, assessments in the form of interviews can only be recommended for collecting additional information.

At this early stage in methodology research it is not possible to say which of the two methodologies performs better in terms of speed and reliability for audiovisual assessments: the traditional attribute-driven quantitative assessment (which requires training for correct usage of semantic descriptors), or the quantitative assessment asking for overall-quality, complemented by a semi-structured interview gathering additional qualitative data.

There are, however, methods that neither require a pre-defined set of attributes nor semi-structured interviews to elicit both, overall perceived quality and decisive attributes. These methods are relatively new to research in the field of perceived quality of technical appliances. In areas that are often considered more peregrine by traditional engineers, like food science, they have been applied with great success over the last decade.

Here, a mixed method approach is introduced that has been used in two independent experiments related to perceived audiovisual quality [wei07a, str07] supervised by the author. The sample sizes were 10 and 20 subjects, respectively. The data has not been analyzed completely yet, such that no results are given. Only the methodology is quickly introduced to give an overview of a third alternative method of data collection for bimodal assessments.

8.13.1 Internal Preference Mapping

The quantitative part of the mixed method approach is based on an Absolute Category Rating (ACR) process (single stimulus method) as suggested in ITU-T P.911 [itu911]. The collected data is analyzed using ANOVA or, if the ANOVA criteria are not met, an equivalent non-parametric method.

In the second part of the quantitative analysis an Internal Preference Mapping (IPM) is performed. The IPM corresponds to a Principal Components Analysis (PCA) and thus delivers an overview of the preference of test subjects for certain items. The PCA reduces multidimensional sets of data to lower dimensions for analysis, thus identifying which items contain the most salient attributes in the quality perception of test subjects.

8.13.2 Free Choice Profiling

The qualitative analysis is based on the Free Choice Profiling (FCP) method. In this method, test subjects themselves develop their own, personalized set of attributes. Complex circumstances like preference or taste are described using well acquainted vocabulary [jac91]. This approach generally minimizes the problem of attaching the correct meaning to a given semantic identifier, a problem that is especially relevant when working with untrained ("naive") assessors.

A common problem with FCP is that subjects need to be acquainted with the method: usually they need to be introduced to the process of developing the vocabulary. In the two experiments performed, this was done presenting the test subjects with different kinds of cookies and asking them to describe the features they liked or disliked about the cookies. Immediately after this they were asked to do the same for a selection of different audiovisual test scenes that would also be used in the final assessment. For each scene the attributes were collected on a separate sheet of paper. Repetitions were allowed.

From the attributes, individual questionnaires were developed for each test subject. These contained the personal attributes along with a horizontal line of $10cm$ of length marked with *min* and *max* on the two ends, respectively. In a second session that served as a training session, subjects were again presented with the audiovisual scenes. They were asked to rate these scenes using the personalized questionnaires that contained their own personalized attributes. Furthermore, this session allowed subjects to refine their vocabulary: subjects were asked to delete, modify or add attributes if necessary and to write down a short definition for each attribute. This definition is helpful at a later stage in the analysis process.

Finally, when subjects are acquainted with the method, the qualitative data is collected in a third session.

8.13.3 Generalized Procrustes Analysis

The data collected with the FCP method is analyzed using the Generalized Procrustes Analysis (GPA) introduced by Gower [gow75]. To analyze data based on individual attributes it is necessary to make these so-called profilings comparable. In the GPA this is called to create a *Consensus*. It is assumed that all sets of attributes generated by the test subjects describe the same perceptual space. Yet, the dimensionality of the attribute space may vary between subjects - a configuration of m attributes describes an m-dimensional attribute space. Assuming that all configurations relate to the same items,

the GPA minimizes the goodness-of-fit criterion by rotating, scaling and dislocating the individual configurations to meet an average configuration, the Consensus [gow75].

8.13.4 External Preference Mapping

To make the method a mixed method, the results from both the quantitative (ACR) and the qualitative assessments (FCP) need to be combined. This can be done using an External Preference Mapping (EPM). The EPM maps two different sets of data into a common space. It is based on the Partial Least Square regression (PLS). The PLS in turn is similar to the PCA in that it tries to reduce the multidimensionality of spaces to a lower dimension while maintaining the most important characteristics. Here, the two multidimensional sets of data that are fused are the quantitative and the qualitative data sets.

8.13.5 Summary

Supporting the preferences of items resulting from the ACR assessment with the attributes generated in the FCP assessment is a promising way of obtaining quantified estimations of perceived quality along with the underlying quality attributes. Still, the mixed method approach suggested here needs further evaluation. E.g., criteria for determining the necessary output dimensionality of the PLS need to be verified. The process of analysis is rather complicated and lengthy - no monolithic software solutions exist. Alternative procedures for the GPA have been suggested by Kunert and Qannari [kun99] that promise simpler algorithms with similar results.

Part V

Summary and Conclusions

9 Conclusions, Discussion and Further Work

The dissertation thesis at hand addresses the question of human audiovisual perception in interactive application systems of moderate complexity. One of the main benefits of audiovisual application systems is that they easily convey the temporal and spatial relationship of different objects. In these systems the computing power available is limited whereas a high overall quality impression is desired. Because the perceived overall quality is not directly related to the simulation accuracy, a new paradigm of improving such systems has been introduced. By better integrating the auditory and visual modalities and by exploiting effects of divided attention caused by user interactivity with the application, the perceived overall quality can be maintained or increased even when the simulation depth of the auditory part is reduced. Only those auditory stimuli that contribute to the perceived overall quality impression have to be simulated in-depth.

In order to separate a priori those attributes of an audiovisual scene that are decisive for the perceived overall quality from those that go unnoticed, a model of the human perceptual process is necessary. A perceptual model that would satisfy the needs of interactive audiovisual application systems does not yet exist. It is therefore necessary to develop such a model.

This was started by summarizing the main characteristics of the human auditory and visual systems with a focus on the perception of space. The current state of knowledge regarding the mechanisms of cognition and the processing of percepts in humans was analyzed from a neurophysiological perspective. This was complemented with an overview of the most important perceptual effects known in human audiovisual perception. From these findings it became clear that attention plays a dominant role in the perceptual process. A number of models at different levels of abstraction have been analyzed that try to describe how attention influences and steers perception. The most prominent is Neisser's Perceptual Cycle. The characteristic trait they all have in common is that they are purely qualitative. To make use of them in technical applications, a quantification of the models' parameters is needed. Such a quantification is only possible by collecting extensive sets of data by means of subjective assessments.

The interactive audiovisual assessments performed in the course of this work aim at answering different types of research questions. First of all, the results show that the system devised for performing these assessments is actually functional, provides straight-forward usability for both, test subjects and supervisors, and integrates a number of complex technologies to form a reliable test bench. Second of all, it was for the first time possible to perform a thoroughly inter-coordinated sequence of audiovisual tests using the same test platform, based on the recommendations originally authored for unimodal subjective assessments. The results show the deficiencies of these recommendations for the bimodal and multi-modal case, but they also show that large parts can be used as a basis for modified recommendations. Thirdly, the results themselves deliver some insight into the very complex processes of human audiovisual perception. These aspects are summarized and discussed in the following, along with suggestions for further work.

9.1 Assessment System

The assessment system created consists of three main parts: the I3D MPEG-4 based interactive audiovisual scene renderer, the Input Device for test subject's haptic feedback, and the SALT logging and exporting tool for the obtained data. All three elements provide clearly defined interfaces. They could be exchanged with other elements providing the same or extended functionality without provoking malfunctions to the system as a whole.

I3D The I3D's audio functionality, namely the TANGA real-time rendering engine which has been the focus of the author's work in the MPEG-4 context, is the first modular audio engine available in an MPEG-4 player. Its development has been supervised and guided by the author. The specifications follow the needs of today's audiovisual scene representation paradigms. The room acoustic simulation features that include a simplified image source model and the so-called *Perceptual Approach* are unique. The experiments performed have shown that in the simplified image source model, a careful matching of Early Reflections (ER) and Diffuse Reverberation (DR) parts is necessary to produce a convincing overall room acoustic impression. This is more important than the sheer number or order of image sources contributing to the ERs. Although the computational load is comparatively high, the image source model based computation allows for immanent changes in the ER pattern. These changes (or the lack of) were clearly perceived by test subjects in the audiovisual assessments performed.

A direct comparison of perceived quality between the MPEG-4 *Perceptual Approach* and the *Physical Approach* was not performed. Although from an academic point of view such a comparison is attractive and would allow to quantify the influence of changes in the ER pattern provided by the *Physical Approach*, its implementation is more than difficult. In order to assure that subjects actually based their quality ratings on the ER patterns and not on the overall sound of the reverberation, the *Perceptual Approach* parameters would need a careful fine-tuning to resemble the sound of the *Physical Approach*. Because not all of these parameters are orthogonal, this is a cumbersome and lengthy task. No automatism for this task exists.

Occlusion and obstruction effects have been identified to be very important for a convincing audiovisual impression. Unfortunately, the methods that exist for the detection of obstruction are either very coarse (e.g. using BoundingSpheres) or computationally expensive. In the course of this work it was not possible to come up with more efficient algorithms. Yet, an alternative method for the room acoustic computation was identified and described in detail that immanently provides acoustic obstruction as part of the rendering result: the beam tracing method. For rectangular-shaped rooms without large obstructing objects the image source method provides very useful results. For more complex rooms or systems of (acoustically) interconnected rooms, the beam tracing method has very high potential.

The first implementation of the newly-standardized MPEG-4 *WideSound* node has shown that the use of distributed sound sources can enhance the perceived naturalness of sound sources compared to the usual point sound sources. This has been found in informal audiovisual test sessions. Therefore, more research work needs to go into this interesting topic in the future before recommendations can be given.

As the TANGA engine is multi-processor / multi-thread capable, it is very well primed for the use on future processors. It has been shown to run very well on both Intel and

AMD dual core processors. From the specifications, it can be assumed that it runs equally well on quad core processors of the same manufacturers. Whether it performs equally well on the octal core Cell processor (as used e.g. in the Sony Playstation 3) depends on the structure of that processor, the available cache and RAM sizes, and so on. Further research is necessary to come up with sound evidence concerning this matter.

Because TANGA can run on Windows and Linux based Operating Systems (OS) and can also run without the MPEG-4 context (standalone as a command line application), it could be used in a number of different contexts. Its application is not reduced to audio rendering within the MPEG-4 specifications. In fact, all audio rendering Components have been debugged and tested outside the MPEG-4 context in the first place. A TANGA Component Graph (CG) can easily be generated directly in the form of C++ code instructions instead of deriving it from the MPEG-4 scene description. Of course, it would also be possible to devise a Graphical User Interface (GUI) for generating the CG if needed.

In areas in which the MPEG-4 standard does not provide sufficient means of external control, the I3D has been enhanced beyond the standard (e.g. the MIDI functionality implemented). The MIDI extension allows to modify in real-time the scene description via (nearly) arbitrary input devices. Also, the exchange of scene control and status data is possible. This for the first time allowed to use an MPEG-4 player as a rendering device for assessments evaluating the perceived quality of interactive audiovisual scenes. This is especially interesting since the MPEG-4 paradigm of object-based scene description is widely believed to be one of the main innovations to the way how audiovisual content will be produced, distributed and consumed in the future.

Input Device The haptic Input Device as a means of collecting quality feedback from test subjects has proven to work efficiently and flawlessly. Test subjects were, according to their own estimation, able to concentrate fully on the audiovisual percepts presented in the assessments. Especially for sequences of single-stimulus items (ACR) this was observed. The usage of the Input Device was regarded as straight-forward and intuitive.

Thanks to a careful selection of the microcontroller platform and the underlying MIDI operating system running on the microcontroller, the Input Device is stable and worked without failure during the whole assessment period. The concept of exchangeable front plates is helpful to further reduce test subjects' confusion, but has not been applied consequently in the course of the assessments. In fact, it was shown that simply removing the unnecessary faders already simplified the front plate to an extent that completely avoided any possible confusion. For the buttons, clearly marking them and leaving those buttons that were not needed without functionality was sufficient with the test subjects that participated. This might be different for test populations that are considerably less technophile.

The MIDI-based communication structure proved to be simple to employ and stable. The MIDI protocol is easy to implement and fast enough for the purpose.

SALT The JAVA-based Subjective Assessment Logging Tool (SALT) has been used extensively to record the test subjects' ratings and all other events (high scores, presentation order of items and trials, subject related data) during the assessments. Its GUI is straightforward and helps to prepare different types of assessments. The fact that complete assessment setups can be saved and restored at a later time has considerably helped in performing assessments in a lab that was frequently used by other experimenters in

between the test sessions.

The data export worked comfortably and flawlessly. The common process of pasting data from different test subjects to form a set of data for the whole test population is done automatically. Because most statistical analysis tools can import EXCEL tables, no further export formats need to be implemented at this time.

Together with the Input Device, SALT for the first time provides an integrated hardware / software system for easy collection, unscrambling[1] and export of assessment data. It is not limited to the field of audiovisual assessments, but can also be applied in traditional unimodal quality tests.

9.2 Assessment Recommendations

The assessments performed have shown that the existing recommendations issued by the international bodies of ITU, EBU, AES and IEC provide basic guidelines that can be transferred to the bimodal case. This is true for all recommendations related to the test setup itself and the test methodologies to use. Yet, it is important to note that the focus of these recommendations is on the evaluation of perceived quality of simple commercial systems. It is therefore necessary to specify the type of assessment to be performed before blindly following any such recommendation. In this work, a novel categorization of assessments into three categories has been suggested, see section 8.4. The existing recommendations certainly do not cover assessments of the first category (assessments that serve to understand the human perceptual processes), because they simply ignore the mechanisms of mutual influence that the auditive and visual modalities can have upon the perceived quality.

Also, the existing recommendations stem from a period in which such commercial systems did not (or only to a very limited degree) provide means of interaction with a user. None of the recommendations suggests how to consider and include interaction possibilities in the course of evaluation of such systems. That there is an - albeit small - effect of interaction (or, more precisely, of task) upon the perceived quality has been shown in this work. Therefore it would be favorable if guidelines for the evaluation of application-specific interaction features existed. Yet, it remains unclear how a generalized set of recommendations could cover the whole field.

What is clearly missing in the existing recommendations is a meaningful set of audiovisual quality attributes. The problem with these is that they are strongly context dependent and may change their individual meaning with a test subject's personal background. The assessment described in section 8.9 has introduced an alternative method that does not rely on potentially fuzzy attributes. Yet, the workload imposed upon the experimenter(s) in such an approach is almost prohibitive for larger sample sizes.

The Mixed Method approach outlined in section 8.13 needs further evaluation. It consists of a combination of Absolute Category Rating (ACR) and Internal Preference Mapping (IPM) / Principal Components Analysis (PCA) with Free Choice Profiling (FCP) and Generalized Procrustes Analysis (GPA) by means of an External Preference Mapping (EPM) / Partial Least Square regression (PLS). Two promising experiments have been performed using this methodology, but a number of issues remain open that need further

[1] The term "scrambled data" here refers to the fact that items are usually presented in random order. The related ratings are also stored in the order of presentation, therefore they need to be "unscrambled" or sorted before the data is exported.

research: especially the questions of validity and reliability of the method call for further examination before it can be accepted as a new approach in the assessment of perceived audiovisual quality.

It has been shown that data collected in audiovisual assessments very often does not fulfill all of the ANOVA criteria. This is in correspondence with other author's findings, e.g. [kas03], and calls for the use of non-parametric methods of analysis. Apparently, the use of non-parametric methods of analysis is often evaded by claiming "close to normal distribution" of scores or residuals and applying ANOVA anyway. Improved guidelines and recommendations might help to overcome this self-imposed inaccuracy.

9.3 Audiovisual Perception

Chapter 8 has presented six subjective assessments of perceived audiovisual quality that were performed using the assessment system developed in the course of this work. The initial motivation for this work was to find methods and schemes of reducing the computational costs in the audio simulation part of interactive audiovisual scenes without deteriorating the perceived overall quality. Consequently, the assessments aimed at verifying different toeholds: Sections 8.7 and 8.8 focus on a possible reduction of algorithmic complexity. The simplifications assessed are directly related to the computational load that the real-time rendering of audio imposes on the processor. It has been shown that the number of loudspeakers necessary in interactive audiovisual application systems of moderate complexity using a VBAP panning approach depends on the content itself. As a rule of thumb, the well-known five-channel setup defined in ITU-R BS.775 [itu775] is suitable for interactive application systems of moderate complexity. The *Perceptual Approach* algorithm as specified in MPEG-4 Scene Description can be simplified to use only four internal workchannels without degrading the overall perceived quality in the audiovisual context.

Sections 8.9 to 8.11 focus on the effect that interaction with the audiovisual application or scene might have on the perceived overall quality. Here the general assumption was that by offering an attractive interactive content or by assigning the user a challenging task, that user would become more involved and thus experience a subjectively higher overall quality. As section 8.9 shows, this is not generally the case. However, when both task and main varying (or salient) quality attribute were located in the same modality, such an effect could be substantiated. Apparently, inner-modal influence is significantly greater than cross-modal influence. This is also suggested by the common theories of capacity limits in human attention.

Finally, section 8.12 showed that also cross-modal influence of interaction is possible when stimuli and interaction are 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 a salience model for interactive audiovisual applications of moderate complexity. Now, these influence factors need to be quantified.

9.4 General Conclusion

A new paradigm of "multi-modal perceptual coding" promises an improved overall perceived quality in interactive audiovisual application systems without further increasing the simulation depth. Instead, the available computing power is distributed according to

the priorities of the human perceptual processes, resulting in an optimum quality / cost ratio. Object-based description schemes like MPEG-4 Scene Description lend themselves especially well for this approach. Multi-modal perceptual coding needs a model of human multi-modal perception to work properly. Such a model does not yet exist. In this work, a first step toward the creation of a perceptual model valid for interactive audiovisual application systems of moderate complexity has been made. This was achieved by identifying and verifying in a series of subjective assessments the main influence factors that compose the salience model of human perception for such applications. These influence factors need further quantification.

A system that is able to perform subjective assessments of perceived audiovisual quality has been developed and put to use. For the first time, such a system has been based on the object- and scene-concept of MPEG-4. One of the main benefits of that concept is that auditory and visual characteristics of an object are inherently related, a scheme that corresponds to the mechanisms of human perception. The existing recommendations for unimodal subjective assessments can be transferred in part to the multi-modal case. The methodologies need to be revised to include both quantitative and qualitative methods as well as combinations of the two. Suggestions for this have been made in this dissertation. A total of eight subjective assessments has been performed using the assessment system. The obtained data was fully analyzed for six of them. The results indicate that cross-modal influence of interaction and task is possible, depending on the afore mentioned influence factors. An inner-modal influence has been verified for the type of application at hand.

List of Figures

List of Tables

List of Abbreviations

$F0$	Fundamental Frequency
T_{60}	Reverberation Time
2D	two-dimensional
3D	three-dimensional
AABIFS	Advanced AudioBIFS
ACR	Absolute Category Rating
ADAT	Alesis Digital Audio Tape
AES	Audio Engineering Society
AI	Primary Auditory Cortex
ALSA	Advanced Linux Sound Architecture
AMEI	Association of Musical Electronics Industry
ANOVA	Analysis of Variance
ANSI	American National Standards Institute
API	Application Programming Interface
ASCII	American Standard Code for Information Interchange
ASIO	Audio Streaming Input Output
BA	Brodmann Area
BFS	Breadth-First Search
BIFS	BInary Format for Scenes
BRIR	Binaural Room Impulse Response
BRS	Binaural Room Scanning
BSI	Background Spatial Impression
BSP	Binary Space Partitioning
BTL	British Telecom Labs
BTL	British Telecom Labs
CC	Continuous Controller
CI	Confidence Interval
CLAM	C++ Library for Audio and Music
CPP	Controlled Parallel Processing
CPU	Central Processing Unit
CSI	Continuous Spatial Impression
CSV	Character Separated Values
DA	Descriptive Analysis
DAC	Digital Analog Converter
DCCR	Dynamic Component Cluster Rendering
DCPR	Dynamic Component Parallel Rendering
DCR	Degradation Category Rating
DDF	Device Data Frame
DR	Diffuse Reverberation
DSP	Digital Signal Processor
EBU	European Broadcasting Union
ECMA	European Computer Manufacturers Association

EPM	External Preference Mapping
ER	Early Reflections
ES	Elementary Stream
ESI	Early Spatial Impression
FCP	Free Choice Profiling
FDAM	Final Draft AMendment
FIR	Finite Impulse Response
FLE	Flash-Lag Effect
fMRI	Functional Magnetic Resonance Imaging
GPA	Generalized Procrustes Analysis
GPS	Global Positioning System
GPTI	Graph Processing Time Index
GUI	Graphical User Interface
HRIR	Head Related Impulse Response
HRTF	Head Related Transfer Function
HUT	Helsinki University of Technology
I3D	Interactive 3D (MPEG-4 player)
IASIG	Interactive Audio Special Interest Group
IAVAS	Interactive Audio Visual Application Systems
IC	Inferior Colliculus
IC	Integrated Circuit
IEC	International Electrotechnical Commission
IEM	Institute of Electronic Music and Acoustics
IETF	Internet Engineering Task Force
IIR	Infinite Impulse Response
ILD	Interaural Level Difference
IM1	Implementation Model One
IMT	Institute of Media Technology
IP	Internet Protocol
IPM	Internal Preference Mapping
IR	Impulse Response
IRCAM	Institut de Recherche et Coordination Acoustique/Musique
IRT	Institut für Rundfunktechnik
ISM	Image Source Method
ISO	International Standards Organization
ITD	Interaural Time Difference
ITU	International Telecommunication Union
LCD	Liquid Crystal Display
LGN	Lateral Geniculate Nucleus
LGPL	Lesser General Public License
LOD	Level of Detail
LSO	Lateral Superior Olivary Nucleus
LTC	Linear Time Code
LTI	Linear Time Invariant
MFLOPS	Million Floating Point Operations Per Second
MIDI	Musical Instrument Digital Interface
MIT	Massachusetts Institute of Technology

MLE	Maximum-Likelihood Estimation
MLS	Maximum Length Sequence
MMA	MIDI Manufacturers Association
MP3	MPEG-1 Layer 3 data format
MPEG	Moving Picture Experts Group
MSO	Medial Superior Olivary Nucleus
MUSHRA	MUltiple Stimuli with Hidden Reference and Anchor
NP	Non-deterministic Polynomial time
OD	Object Descriptor
PA	Public Address
PAL	Phase Alternating Line
PCA	Principal Components Analysis
PLS	Partial Least Square regression
PTI	Processing Time Index
RGT	Repertory Grid Technique
RIR	Room Impulse Response
RMRS	Realtime Multichannel Room Simulator
RMS	Root Mean Square
RTP	Real-time Transport Protocol
SA	Structured Audio
SALT	Subjective Assessment Logging Tool
SAOL	Structured Audio Orchestra Language
SDM	Systems Decoder Model
SFC	Signal Flow Chart
SysEx	System Exclusive
TANGA	The Advanced Next Generation Audio
UDP	User Datagram Protocol
V1	Primary Visual Cortex
V2	Visual Association Cortex
VBAP	Vector Base Amplitude Panning
VE	Virtual Environment
VR	Virtual Reality
VRML	Virtual Reality Modeling Language
WFS	Wave Field Synthesis
WMM	Windows MultiMedia
XML	Extensible Markup Language

Bibliography

[14496-1] Int. Std. (IS) ISO/IEC 14496-1:2004, Information technology - Coding of audio-visual objects - Part 1: Systems, 3rd Ed., Geneva, Switzerland, 2004.

[14496-3] Int. Std. (IS) ISO/IEC 14496-3:2007, Information technology - Coding of audio-visual objects - Part 3: Audio, 4rth Ed. Appr. Draft, N9075, Geneva, Switzerland, 2007.

[14496-5] Int. Std. (IS) ISO/IEC 14496-5:2001, Information technology - Coding of audio-visual objects - Part 5: Reference software, Geneva, Switzerland, 2001.

[14496-11] Int. Std. (IS) ISO/IEC 14496-11:2004, Information technology - Coding of audio-visual objects - Part 11: Scene description and Application engine, Geneva, Switzerland, 2004.

[14496-11.1] Int. Std. (IS) ISO/IEC 14496-11.1:2004, Information technology - Coding of audio-visual objects - Part 11.1: MPEG-4 Systems Node Semantics, Geneva, Switzerland, 2004.

[14772-1] Inst. Std. (IS) ISO/IEC 14772-1:1997, Virtual Reality Modeling Language (VRML), Geneva, Switzerland, 1997.

[16262] European Computer Manufacturers Association: "ECMAScript Language Specification", 3rd Edition, Geneva, December 1999, *also available as* Int. Std. (IS) ISO/IEC 16262:1999, Information Technology - ECMAScript language specification, Geneva, Switzerland, 1999.

[aes20] AES 20, AES Recommended Practice for Professional Audio - Subjective Evaluation of Loudspeakers, Audio Engineering Society, New York, USA, 1996.

[ait90] Aitkin, Lindsay: "The Auditory Cortex", Chapman and Hall, London, 1990, ISBN 0-412-32490-3.

[ala99] Alais, David; Blake, Randolph: "Neural Strength of Visual Attention Gauged by Motion Adaptation", Nature Neuroscience, vol.2, no.11, 1999, pp 1015-1018.

[ala03] Alais, David; Burr, David: "The 'Flash-Lag' Effect Occurs in Audition and Cross-Modally", Current Biology, Vol. 13, January 8, 2003, pp 59-63.

[ala04] Alais, David; Burr, David: "The Ventriloquist Effect Results from Near-Optimal Bimodal Integration", Current Biology, Vol. 14, February 3, 2004, pp 257-262.

[ama04] Amatriain, Xavier: "An Object-Oriented Metamodel for Digital Signal Processing", PhD Thesis, Universitat Pompeu Fabra, Barcelona, Spain, 2004.

[amd67] Amdahl, Gene M.: "Validity of the single-processor approach to achieving large scale computing capabilities", in American Federation of Information Processing Societies (AFIPS) Conference Proceedings vol. 30, Atlantic City, N.J., Apr. 18-20, 1967, pp 483-485, retyped version by Guihai Chen.

[ansi31] ANSI S3.1-1999(R2003), Maximum Permissible Ambient Noise Levels for Audiometric Test Rooms, American National Standards Institute, New York, USA, 2003.

[ansi32] ANSI S3.2-1989(R1999), Method for Measuring the Intelligibility of Speech over Communications System, American National Standards Institute, New York, USA, 1999.

[ant04a] Antonacci, Fabio; Foco, Marco; Sarti, Augusto; Tubaro, Stefano: "Fast Modeling Of Acoustic Reflections And Diffraction In Complex Environments Using Visibility Diagrams", European Signal Processing Conference (EUSIPCO-2004), Vienna, Austria, September 2004.

[ant04b] Antonacci, Fabio; Foco, Marco; Sarti, Augusto; Tubaro, Stefano: "Accurate And Fast Audio-realistic Rendering Of Sounds In Virtual Environments", IEEE Multimedia Signal Processing Workshop (MMSP-04), Siena, Italy, September 2004, pp 271-274.

[ant04c] Antonacci, Fabio; Foco, Marco; Sarti, Augusto; Tubaro, Stefano: "Real Time Modeling of Acoustic Propagation in Complex Environments", Proc. of the 7th Int. Conference on Digital Audio Effects (DAFx'04), Naples, Italy, October 5-8, 2004.

[bar71] Barron, M.: "The subjective effects of first reflections in concert halls: the need for lateral reflections", J. of Sound and Vibration, 15:475-494, 1971.

[bat03] Battaglia, Peter W.; Jacobs, Robert A.; Aslin, Richard N.: "Bayesian Integration of Visual and Auditory Signals for Spatial Localization", J. Opt. Soc. Am., Vol. 20, No.7/July 2003, pp 1391-1397.

[bec95] Bech, Soren; Hansen, Villy: "Interaction Between Audio-Visual Factors in a Home Theater System: Experimental Results", AES Audio Engineering Society 99th Convention, Preprint 4096, New York, USA, 1995.

[bec06] Bech, Soren; Zacharov, Nick: "Perceptual Audio Evaluation - Theory, Method and Application", John Wiley & Sons Ltd., Chichester, West Sussex, England, 2006, ISBN 0-470-86923-2.

[beg94] Begault, Durand R.: "3-D Sound for Virtual Reality and Multimedia", AP Professional, Boston Mass., USA, ISBN 0-12-084735-3.

[beg96] Begault, Durand R.: "Audible and Inaudible Early Reflections: Thresholds for Auralization System Design", AES Audio Engineering Society 100th Convention, Preprint 4244, Copenhagen, Denmark, May 11-14, 1996.

[bek57] Békésy, G. von: "The Ear", in *Scientific American*, 197 (2), 1957, pp 66-78.

[bek60] Békésy, G. von: "Experiments in Hearing", McGraw Hill, New York, 1960.

[bel02] Beltrán, José R.; Beltrán, Fernando A.: "Matlab Implementation of Reverberation Algorithms", Journal of New Music Research, Vol. 31, Issue 2, June 2002, pp 153-161.

[ben01] Bencina, Ross; Burk, Phil: "PortAudio - an Open Source Cross Platform Audio API", Proc. of the 2001 International Computer Music Conference, Havana, Cuba, Sept. 2001, pp 263-266.

[ber88] Berkhout, A.J.: "A Holographic Approach to Acoustic Control", Journal of the Audio Engineering Society, vol. 36, December 1988, pp 977-995.

[ber96] Bernstein, Lynne E.; Benoît, Christian: "For Speech Perception by Humans or Machines, Three Senses are Better than One", Proc. Fourth International Conference on Spoken Language Processing ICSLP, Philadelphia, USA, October 3-6, 1996.

[ber99] Berg, Jan; Rumsey, Francis: "Identification of Perceived Spatial Attributes of Recordings by Repertory Grid Technique and Other Methods", AES 106th Convention, Munich, Germany, May 8-11, 1999, Preprint 4924.

[ber03] Berg, Jan; Rumsey, Francis: "Systematic Evaluation of Perceived Spatial Quality", Proceedings of the AES 24th International Conference, Banff, Alberta, Canada, June 26-28, 2003, pp 184-198.

[bla00] Blauert, Jens; Lehnert, H.; Sahrhage, J.; Strauss, H.: "An Interactive Virtual-Environment Generator for Psychoacoustic Research. I: Architecture and Implementation", Acustica - acta acustica, Vol. 86, 2000, pp 94-102.

[bla01] Blauert, J.: "Spatial Hearing: The Psychophysics of Human Sound Localization", MIT Press, Cambridge, MA, rev. 3rd ed., 2001.

[bor05] Bortz, Jürgen: "Statistik für Human- und Sozialwissenschaftler", (German language), 6th Edition, Springer Medizin Verlag, Heidelberg, 2005.

[bra03] Bradley, J.S.; Sato, H.; Picard, M.: "On the importance of early reflections for speech in rooms", Journal of the Acoustical Society of America, Vol. 113, No. 6, June 2003, pp 3233-3244.

[bra05] Braasch, Jonas: "Modelling of Binaural Hearing", in J. Blauert (Ed.), *Communication Acoustics*, Springer Verlag, Berlin Heidelberg, 2005, ISBN 3-540-22162-3.

[bro09] Brodmann, Korbinian: "Vergleichende Lokalisationslehre der Grosshirnrinde in ihren Prinzipien dargestellt auf Grund des Zellenbaues", (German language), Johann Ambrosius Barth Verlag, Leipzig, Germany, 1909.

[bru03] de Bruijn, Werner P.J.; Boone, Marinus M.: "Application of Wave Field Synthesis in Life-size Videoconferencing", AES 114th Convention, Amsterdam, The Netherlands, March 22-25, 2003, Preprint 5801.

[bue06] Bühl, Achim: "SPSS 14 - Einführung in die moderne Datenanalyse", (German language), 10th Edition, Pearson Studium, Munich, Germany, 2006, ISBN 3-827-37203-8.

[car96] Carey, James W.; Morgan, Mark; Oxtoby, Margaret J.: "Intercoder Agreement in Analysis of Responses to Open-Ended Interview Questions: Examples from Tuberculosis Research", Cultural Anthropology Methods Journal 8(3):1-5., 1996.

[cla76] Clark, James H., "Hierarchical Geometric Models for Visible Surface Algorithms", Communications of the ACM, Vol. 19, No. 10, Oct. 1976, pp 547-554.

[coe01] Coen, Michael H., "Issues in Intersensory Perception for Interactive Systems", Workshop on Developmental Embodied Cognition (DECO-2001), 23rd Annual Meeting of the Cognitive Science Society on July 31st in Edinburgh, Scotland, UK, July 31, 2001.

[coh97] Cohen, Jonathan D.; Perlstein, William M.; Braver, Todd S.; Nystrom, Leigh E.; Noll, Douglas C.; Jonides, John; Smith, Edward E.: "Temporal Dynamics of Brain Activation During a Working Memory Task", *Nature*, 386, 1997, pp 604-608.

[con00] Connor, S.: "Dumbstruck: A Cultural History of Ventriloquism", Oxford University Press, Oxford, 2000.

[coo04] Coolican, Hugh: "Research Methods and Statistics in Psychology", 4th Edition, Hodder Arnold, London, UK, 2004, ISBN 0-240-81258-3.

[coo07] personal email communication with Hugh Coolican, 2007-10-28 and 2007-10-30.

[cre78] Cremer, Lothar; Müller, Helmut A.: "Die wissenschaftlichen Grundlagen der Raumakustik, Band 1: Teil 1: Geometrische Raumakustik", (German language), 2nd Edition, S. Hirzel Verlag, Stuttgart, 1978, ISBN 3-7776-0315-5.

[cre06] Creswell, John W.; Plano Clark, Vicki L.: "Designing and Conducting Mixed Methods Research", Sage Publications Inc., Thousand Oaks, CA, USA, 2006, ISBN 1-412-92792-7.

[dan04] Dantele, Andreas; Reiter, Ulrich; Schwark, Mathias: "Audiovisual Virtual Environments: Enabling Realtime Rendering of Early Reflections by Scene Graph Simplification", AES 116th Convention, Berlin, Germany, May 8-11, 2004.

[dee95] Deering, Michael: "Geometry Compression", SIGGRAPH '95, 22nd ACM Conference on Computer Graphics and Interactive Techniques, Los Angeles, CA, August 1995, pp 13-20.

[dje00] Djelani, T.; Pörschmann, C.; Sahrhage, J.; Blauert, J.: "An Interactive Virtual-Environment Generator for Psychoacoustic Research II: Collection of Head-Related Impulse Responses and Evaluation of Auditory Localization", Acustica - acta acustica, Vol. 86, 2000, pp 1046-1053.

[ebu3276] EBU Technical Document Tech 3276 - Listening Conditions for the Assessment of Sound Programme Material: Monophonic and Two-Channel Stereophonic, 2nd Ed., European Broadcast Union, Geneva, Switzerland, 1998.

[ebu3276a] EBU Technical Document Tech 3276: Supplement 1 - Listening Conditions for the Assessment of Sound Programme Material: Multichannel Sound, European Broadcast Union, Geneva, Switzerland, 1999.

[ebu3286] EBU Technical Document Tech 3286 - Assessment Methods for the Subjective Evaluation of the Quality of Sound Programme Material - Music, European Broadcast Union, Geneva, Switzerland, 1997.

[ebu3286a] EBU Technical Document Tech 3286: Supplement 1 - Assessment Methods for the Subjective Evaluation of the Quality of Sound Programme Material - Multichannel, European Broadcast Union, Geneva, Switzerland, 2000.

[ell96] Ellis, Stephen R.: "Presence of Mind... a reaction to Thomas Sheridan's 'Musing on Telepresence' ", in *Presence*, 5, 1996, pp 247-259.

[far03] Farris, J. Shawn: "The Human Interaction Cycle: A Proposed and Tested Framework of Perception, Cognition, and Action on the Web", PhD Thesis, Kansas State University, USA, 2003.

[foc03] Foco, M.; Polotti, P. ; Sarti, A.; Tubaro, S.: "Sound Spatialization Based On Fast Beam Tracing In The Dual Space", Proc. of the 6th Int. Conference on Digital Audio Effects (DAFX-03), London, UK, September 2003, pp 198-202.

[fos91] Foster, Scott H.; Wenzel, Elizabeth M.: "Virtual acoustic environments: The Convolvotron", Computer Graphics, 25 (4), 386. [Demonstration system at the 1st annual *Tomorrow's Realities Gallery*, SIGGRAPH '91, 18th ACM Conference on Computer Graphics and Interactive Techniques, Las Vegas, NV, 1991, July 27-August 2.].

[fun98] Funkhouser, Thomas; Carlbom, Ingrid; Elko, Gary; Pingali, Gopal; Sondhi, Mohan; West, Jim: "A Beam Tracing Approach to Acoustic Modeling for Interactive Virtual Environments", ACM Computer Graphics, SIGGRAPH '98 Proceedings, Orlando, FL, USA, July 1998, pp 21-32.

[fun99] Funkhouser, Thomas; Min, Patrick; Carlbom, Ingrid: "Real-Time Acoustic Modeling for Distributed Virtual Environments", ACM Computer Graphics, SIGGRAPH '99 Proceedings, August 1999, pp 365-374.

[fun04] Funkhouser, Thomas; Tsingos, Nicolas; Carlbom, Ingrid; Elko, Gary; Sondhi, Mohan; West, James E.; Pingali, Gopal; Min, Patrick; Ngan, Addy: "A beam tracing method for interactive architectural acoustics", J. Acoust. Soc. Am. 115 (2), February 2004, pp 739-756.

[gar92] Gardner, W.G.: "A Realtime Multichannel Room Simulator", J. Acoust. Soc. Am., 92(4), pp 2395, and presented at the 124th Meeting of the Acoustical Society of America, New Orleans, USA, November 1992.

[gar98] Gardner, W.G.: Chapter 3. "Reverberation Algorithms". in: Kahrs, M. and Brandenburg, K. (Editors): *Applications of Signal Processing to Audio and Acoustics*. Kluwer Academic Publishers, Dordrecht, 1998.

[ger90] Gervautz, Michael; Purgathofer, Werner: "A simple method for color quantization: octree quantization", in *Graphics Gems*, ed. by Andrew S. Glassner, Academic Press, 1990, pp 287-293, ISBN 0122861663.

[gol02] Goldstein, E. Bruce: "Wahrnehmungspsychologie" (German language), 2nd edition, ed. by Manfred Ritter, Spektrum Akadem. Verlag, Berlin, 2002, ISBN 3-8274-1083-5.

[gow75] Gower, J. C.: "Generalized Procrustes Analysis", in *Psychometrika*, Vol. 40, No. 1, March 1975, pp 33-51.

[gri97] Griesinger, David: "Spatial Impression and Envelopment in Small Rooms", AES 103rd Convention, New York, New York, USA, September 26-29, 1997, Preprint 4638.

[gri98] Griesinger, David: "General Overview of Spatial Impression, Envelopment, Localization, and Externalization", Proceedings of the AES 15th International Conference: AUDIO, ACOUSTICS & SMALL SPACES, Copenhagen, Denmark, Oct. 31 - Nov. 2, 1998.

[gri99] Griesinger, David: "Objective Measures Of Spaciousness And Envelopment", Proceedings of the AES 16th International Conference on Spatial Sound Reproduction, Rovaniemi, Finland, April 10-12, 1999.

[gus88] Gustafson, John L.: "Reevaluating Amdahl's Law", Communications of the ACM, 31(5), 1988, pp 532-533.

[hai89] Haines, Eric: "Essential Ray Tracing Algorithms", in Andrew S. Glassner (Ed.), *An Introduction To Ray Tracing*, Academic Press Ltd, London, UK, 1989, pp 35-39, ISBN 0-12-286160-4. 1989.

[har97] Harris, Judith; Pryor, Jeffry; Adams, Sharon: "The Challenge of Intercoder Agreement in Qualitative Inquiry", University of Texas at Austin, TX, USA, 1997. 2007-11-17, http://emissary.wm.edu/templates/content/publications/intercoder-agreement.pdf

[hec82] Heckbert, Paul S.: "Color Image Quantization for Frame Buffer Display", Proc. SIGGRAPH '82, 9th ACM Conference on Computer Graphics and Interactive Techniques, Boston, MA, July 1982, pp 297-307.

[hei94] Heinz, Renate: "Entwicklung und Beurteilung von computergestützten Methoden zur binauralen Raumsimulation", dissertation thesis (German language), Shaker Verlag Aachen, 1984, ISBN 3-8265-0073-3.

[hha05] Holzhäuser, Stefan: "Realisierung einer Bedienoberfläche für die Durchführung von AV-Wahrnehmungstests", Diploma Thesis (German language), TU Ilmenau, Germany, 2005.

[hol97] Hollier, M.; Voelcker, R.: "Objective Performance Assessment: Video Quality as an Influence on Audio Perception", AES 103rd Convention, New York, New York, USA, September 26-29, 1997, Preprint 4590.

[hol97a] Hollier, M.; Voelcker, R.: "Towards a Multi-Modal Perceptual Model", BT Technol. J. Vol. 17 No 1, October 1997, pp 162-171.

[hol98] Hollier, Mike P.; Rimell, Andrew N.: "An Experimental Investigation into Multi-Modal Synchronization Sensitivity for Perceptual Model Development", AES 105th Convention, Sept. 1998, San Francisco, California USA, Preprint 4790.

[hol99] Hollier, Mike P.; Rimell, Andrew N.; Hands, David S.; Voelcker, Rupert M.: "Multi-modal Perception", BT Technol. J. Vol. 17 No 1, January 1999, pp 35-46.

[holz99] Holzem, Nicolas: "Implementing Reverberation Algorithms in Matlab", Engineering Studies Final Work, Universidad de Zaragoza, Spain / Université Libre de Bruxelles, Belgium, August 1999.

[how82] Howard, I. P.: Human visual orientation. Wiley, New York, 1982.

[hub87] Hubel, David H.; Wiesel, Torsten N.: "Die Verarbeitung visueller Information" (German language translation), in M. Ritter (Ed.), *Wahrnehmung und visuelles System* (German language), 2nd. ed., Spektrum Verlag, Heidelberg, 1987, ISBN 3-922508-36-7.

[hul04] Hulsebos, Edo Maria: "Auralization Using Wave Field Synthesis", Thesis, Technische Universiteit Delft, The Netherlands, October 2004.

[huo96] Huopaniemi, Jyri; Savioja, Lauri; Takala, Tapio: "DIVA Virtual Reality System", Proceedings of the International Conference on Auditory Display, ICAD'96, Palo Alto, CA, USA, November 1996.

[ias99] MIDI Manufacturers Association Inc.: 3D Audio Working Group, "IASIG: Interactive 3D Audio Rendering Guidelines Level 2.0", 1999. http://www.iasig.org/pubs/pubs.shtml

[iec60268] IEC 60268-13, Sound System Equipment - Part 13: Listening Tests on Loudspeakers, International Electrotechnical Commission, Geneva, Switzerland, 1998.

[ilm01] Ilmonen, Tommi: "Mustajuuri - an Application and Toolkit for Interactive Audio Processing", Proceedings of the 2001 International Conference on Auditory Display, Espoo, Finland, July 29-August 1, 2001.

[irt95] Akustische Information 1.11-1/1995, "Höchstzulässige Schalldruckpegel von Dauergeräuschen in Studios und Bearbeitungsräumen bei Hörfunk und Fernsehen", IRT publication, Munich 1995; http://www.irt.de/IRT/FuE/ak/PDF/AKI1-11.PDF

[itu100] Recommendation ITU-T J.100, *Tolerances for transmission time differences between vision and sound components of a television signal*, International Telecommunication Union, Geneva, Switzerland, 1990.

[itu102] Question ITU-R 102/6, *Methodologies for subjective assessment of audio and video quality*, International Telecommunication Union, Geneva, Switzerland, 1999.

[itu1116] Recommendation ITU-R BS.1116-1, *Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems*, International Telecommunication Union, Geneva, Switzerland, 1997.

[itu1284] Draft Revision of Recommendation ITU-R BS.1284, *General methods for the subjective assessment of sound quality*, Telecommunication Union, Geneva, Switzerland, 2003.

[itu1285] Recommendation ITU-R BS.1285, *Pre-selection methods for the subjective assessment of small impairments in audio systems*, International Telecommunication Union, Geneva, Switzerland, 1997.

[itu1286] Recommendation ITU-R BS.1286, *Methods for the subjective assessment of audio systems with accompanying picture*, International Telecommunication Union, Geneva, Switzerland, 1997.

[itu1359] Recommendation ITU-R BT.1359, *Relative timing of sound and vision for broadcasting*, International Telecommunication Union, Geneva, Switzerland, 1998.

[itu1534] Recommendation ITU-R BS.1534-1, *Method for the subjective assessment of intermediate quality levels of coding systems*, International Telecommunication Union, Geneva, Switzerland, 2003.

[itu500] Recommendation ITU-R BT.500-11, *Methodology for the subjective assessment of the quality of television pictures*, International Telecommunication Union, Geneva, Switzerland, 2002.

[itu775] Recommendation ITU-R BS.775-1, *Multichannel stereophonic sound system with and without accompanying picture*, International Telecommunication Union, Geneva, Switzerland, 1994.

[itu910] Recommendation ITU-T P.910, *Subjective video quality assessment methods for multimedia applications*, International Telecommunication Union, Geneva, Switzerland, 1999.

[itu911] Recommendation ITU-T P.911, *Subjective audiovisual quality assessment methods for multimedia applications*, International Telecommunication Union, Geneva, Switzerland, 1998, and *Corrigendum 1*, 1999.

[itu920] Recommendation ITU-T P.920, *Interactive test methods for audiovisual communications*, International Telecommunication Union, Geneva, Switzerland, 2000.

[jac91] Jack, Frances R.; Piggott, J.R.: "Free Choice Profiling in Consumer Research", in *Food Quality and Preference*, No. 3, 1991, pp 129-134.

[jot91] Jot, Jean-Marc; Chaigne, Antoine: "Digital Delay Networks for Designing Artificial Reverberators", AES 90th Convention, Paris, France, February 1991, Preprint 3030.

[jum07] Jumisko-Pyykkö, Satu; Häkkinen, Jukka; Nyman, Göte: "Experienced Quality Factors - Qualitative Evaluation Approach to Audiovisual Quality", Proc. of SPIE, Vol. 6507, Multimedia on Mobile Devices 2007, 65070M, Feb. 26, 2007.

[kal61] Kalman, R.E.; Bucy, R.S.: "New Results in Linear Filtering and Prediction Problems", J. Basic Eng. Ser., D. 83, 1961, pp 95-108.

[kas03] Kassier, Rafael; Zielinski, Slawomir; Rumsey, Francis: "Computer Games and Multichannel Audio Quality Part 2 - Evaluation of Time-Variant Audio Degradation under Divided and Undivided Attention", AES 115th Convention, New York, USA, October 2003, Preprint 5856.

[keb94] Kebeck, Günter: "Wahrnehmung: Theorien, Methoden und Forschungsergebnisse der Wahrnehmungspsychologie", (German language), Juventa Verlag, Weinheim, Germany, 1994, ISBN 3-7799-0316-4.

[ken95] Kendall, Gary S.: "The Decorrelation of Audio Signals and its Impact on Spatial Imagery", Computer Music Journal, Vol. 19, No. 4, 1995, pp 71-87.

[kle03] Klein, David J.; König, Peter; Körding, Konrad P.: "Sparse Spectrotemporal Coding of Sounds", EURASIP Journal of Applied Signal Processing, 2003:7, pp 659-667.

[kle93] Kleiner, Mendel; Dalenbäck, Bengt Inge; and Svensson, Peter: "Auralization - an Overview", Journal of the Audio Engineering Society, vol. 41(11), 1993, pp 861-875.

[kon02] Kono, Yoshinori; Hasegawa, Hiroshi; Ayama, Miyoshi; Kasuga, Masao; Matsumoto, Shuichi; Koike, Atsushi; Takagi, Koichi: "Visual Image Effects on Sound Localization in Peripheral Region Under Dynamic Multimedia Conditions", ITC-CSCC-2002, International Technical Conference On Circuits/Systems, Computers and Communications, Phuket, Thailand, July 16-19, 2002.

[kuh02] Kühhirt, Uwe; Drumm, Helge; Reiter, Ulrich; and Rittermann, Marco: "Application Systems for MPEG-4", in *Proceedings of the 2002 IEEE 6th International Symposium on Consumer Electronics (ISCE 2002)*, Erfurt, Germany, September 2002.

[kun99] Kunert, Joachim; Qannari, El Mostafa: "A Simple Alternative to Generalized Procrustes Analysis: Application to Sensory Profiling Data", in *Journal of Sensory Studies*, No. 2, 14/1999, pp 197-208.

[kus07] Kuschel, Martin; Freyberger, Franziska; Buss, Martin; Färber, Berthold: "A Presence Measure for Virtual Reality and Telepresence Based on Multimodal

Conflicts", PRESENCE2007, 10th Annual International Workshop on Presence, Barcelona, Spain, Oct. 25-27, 2007, pp 135-143.

[kut91] Kuttruff, Heinrich: "Room acoustics", 3rd Edition, Elsevier Applied Science London, 1991, ISBN 1-85166-576-5.

[lan01] Landragin, Frederic; Bellalem, Nadia; Romary, Laurent: "Visual Salience and Perceptual Grouping in Multimodal Interactivity", Proc. International Workshop on Information Presentation and Natural Multimodal Dialogue IPNMD, Verona, Italy, December 14-15, 2001.

[lar03] Larsson, Pontus; Västfjäll, Daniel; Kleiner, Mendel: "On the Quality of Experience: A Multi-Modal Approach to Perceptual Ego-Motion and Sensed Presence in Virtual Environments", in *Proceedings First ISCA ITRW on Auditory Quality of Systems AQS-2003*, Akademie Mont-Cenis, Germany, April 23-25, 2003, pp 97-100.

[lav95] Lavie, Nilli: "Perceptual Load as a Necessary Condition for Selective Attention", Journal of Experimental Psychology: Human Perception and Performance, Vol.21, No.3, 1995, pp 451-468.

[lav01] Lavie, Nilli: "Capacity Limits in Selective Attention: Behavioral Evidence and Implications for Neural Activity", in: Braun, J. & Koch C. (Eds.): *Visual Attention and Cortical Circuits*, MIT Press, Cambridge, MA, USA, 2001, pp 49-68.

[laz01] Lazzaro, John; Wawrzynek, John: " Compiling MPEG 4 structured audio into C", in Proceedings of Workshop and Exhibition on MPEG-4, San Jose, CA, USA, June 18-20, 2001, pp 5-8, ISBN 0-7803-7165-8.

[lee05] Lee, Kwan Min; Jin, S. A.; Park, N.; Kang, S.: "Effects of narrative on feelings of presence in computer/video games", Paper presented at the Annual Conference of the International Communication Association (ICA), New York, NY, USA, May 2005.

[lee07] Lee, Kwan Min; Jeong, Eui Jun; Park, Namkee; Ryu, Seoungho: "Effects of Networked Interactivity in Educational Games: Mediating Effects of Social Presence", PRESENCE2007, 10th Annual International Workshop on Presence, Barcelona, Spain, Oct. 25-27, 2007, pp 179-186.

[let01] Letz, Stéphane: "Porting PortAudio API on ASIO", GRAME Computer Music Research Lab Technical Note - 01-11-06, Lyon Cedex 01, France, Nov. 2001.

[lip99] Lipscomb, Scott D.: "Cross-Modal integration: Synchronization of Auditory and Visual Components in Simple and Complex Media", Proc. of Forum Acusticum, Berlin, Germany, 1999.

[lok01] Lokki, Tapio; Hiipakka, J.; Savioja, Lauri: "A framework for evaluating virtual acoustic environments", AES 110th Convention, Amsterdam, The Netherlands, May 12-15, 2001, Preprint 5137.

[lom97] Lombard, Matthew; Ditton, Theresa: "At the Heart of it All: The Concept of Presence", In *Journal of Computer-Mediated Communication*, 3, 1997.

[mac93] MacKenzie, I. Scott; Ware, Colin: "Lag as a Determinant of Human Performance in Interactive Systems", Proceedings of the ACM Conference on Human Factors in Computing Systems - INTERCHI '93, Amsterdam, The Netherlands, May 1993, pp 488-493.

[mah99] Mahalingam, Ganapathy: "A New Algorithm for the Simulation of Sound Propagation in Spatial Enclosures", IBPSA Building Simulation Conference and Exhibition 1999, Kyoto, Japan, September 13-15, 1999.

[mar00] Martinkauppi, S.; Rämä, P.; Aronen, H.J.; Korvenoja, A.; Carlson, S.: "Working Memory of Auditory Localization", Cerebral Cortex, Sep. 2000, 10: pp 889-898.

[mas98] Massaro, Dominic W.: "Illusions and Issues in Bimodal Speech Perception", Proceedings of Auditory Visual Speech Perception '98, Terrigal-Sydney, Australia, December 1998, pp 21-26.

[mcg76] McGurk, Harry; MacDonald, John: "Hearing Lips and Seeing Voices", *Nature*, 264, 1976, pp 746-748.

[mee02] Meehan, Michael; Insko, Brent; Whitton, Mary; Brooks, Frederick P., Jr.: "Physiological Measures of Presence in Stressful Virtual Environments", in *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, San Antonio, Texas, 2002, pp 645-652.

[mee03] Meehan, Michael; Razzaque, S.; Whitton, Mary C.; Brooks, Frederick P., Jr.: "Effect of latency on presence in stressful virtual environments", in *Virtual Reality*, Proceedings IEEE, 2003, pp 141-148.

[mei99] Meilgaard, Morten C.; Civille, Gail Vance; Carr, B. Thomas: "Sensory Evaluation Techniques", 3rd Edition, CRC Press, Boca Raton, FL, USA, 1999, ISBN 0-8493-0276-5.

[mel03] Melchior, Frank; Brix, Sandra; Sporer, Thomas; Röder, Thomas; Klehs, Beate: "Wave Field Synthesis in Combination with 2D Video Projection", Proceedings of the AES 24th International Conference on Multichannel Audio, Banff, Alberta, Canada, June 26-28, 2003, pp 30-40, ISBN 0-937803-50-2.

[mel06] Melchior, Frank; Fischer, Jens-Oliver; de Vries, Diemer: "Audiovisual Perception using Wave Field Synthesis in Combination with Augmented Reality Systems: Horizontal Positioning", Proceedings of the AES 28th International Conference, Pitea, Sweden, June 30 - July 2, 2006, pp 92-101, ISBN 0-937803-57-X.

[mey05] Meyers, Scott: "Effective C++", 3rd Edition, Addison-Wesley, New York, USA, 2005.

[mil02] Miller, Joel D.; Wenzel, Elizabeth M.: "Recent Developments in SLAB: a Software-Based System for Interactive Spatial Sound Synthesis", Proceedings of the 2002 International Conference on Auditory Display, Kyoto, Japan, July 2-5, 2002.

[mil03] Miller, J. D.; Anderson, M. R.; Wenzel, E. M.; McClain, B. U.: "Latency Measurement of a Real-Time Virtual Acoustic Environment Rendering System", Proceedings of the International Conference on Auditory Display, ICAD 2003, Boston, MA, USA, 2003.

[moo65] Moore, Gordon E.: "Cramming more components onto integrated circuits", Electronics, Vol. 38, Number 8, April 19, 1965.

[moo79] Moorer, James A.: "About this Reverberation Business", Computer Music Journal, Vol. 3, Nr. 2, 1979, pp 13-28.

[mur73] Murch, Gerald M.: "Visual and Auditory Perception", Bobbs-Merrill Co., Indianapolis, USA, 4th printing 1973, ISBN 0-672-60779-4.

[N6591] ISO/IEC 14496-11/2003 FDAM-3, N6591, MPEG-Meeting Redmond, USA, July 2004.

[nol02] Nolte, John: "The Human Brain - An Introduction to Its Functional Anatomy", 5th edition, Mosby Inc., St. Louis, 2002, ISBN 0-323-01320-1.

[nor05] Nordahl, Rolf: "Self-induced Footsteps Sounds in Virtual Reality: Latency, Recognition, Quality and Presence", PRESENCE 2005, 8th Annual International Workshop on Presence, London, UK, Sept. 21-23, 2005, pp 353-354.

[nov05] Novo, Pedro: "Auditory Virtual Environments", in J. Blauert (Ed.), *Communication Acoustics*, Springer Verlag, Berlin Heidelberg, 2005, ISBN 3-540-22162-3.

[owe05] Owen, A. M.; McMillan, K. M.; Laird, A. R.; Bullmore, E.: "N-Back Working Memory Paradigm: A Meta-Analysis of Normative Functional Neuroimaging Studies", Human Brain Mapping, 25, 2005, pp46-59.

[pak05] Pak, Richard: "An Investigation of Perceptual Load, Aging, and the Functional Field of View", PhD Thesis, School of Psychology, Georgia Institute of Technology, Atlanta, GA, USA, 2005.

[par07] Partzsch, Andreas: "Erweiterung der TANGA-Engine um Multithreading-Fähigkeit", Diploma Thesis (German language), TU Ilmenau, Germany, 2007.

[pas99] Pashler, Harold E.: "The Psychology of Attention", 1st paperback edition, The MIT Press, Cambridge, MA, USA, 1999, ISBN 0-262-66156-X.

[pel02] Pellegrini, Renato S.: "A Virtual Reference Listening Room as an Application of Auditory Virtual Environments", PhD Thesis, dissertation.de-Verlag Berlin, 2002, special edition of ISBN 3-89825-403-8.

[per04] Pereira, Fernando; Ebrahimi, Touradj: "The MPEG-4 Book", Prentice Hall, 2004, ISBN 0-13-061621-4.

[pot06] Potard, Guillaume: "3D-Audio Object Oriented Coding", PhD Thesis, University of Wollongong, NSW, Australia, September 2006.

[pul97] Pulkki, Ville: "Virtual Sound Source Positioning Using Vector Base Amplitude Panning", Journal of the Audio Engineering Society, vol. 45, no. 6, pp. 456-466, June 1997.

[pul01] Pulkki, Ville: "Spatial Sound Generation and Perception by Amplitude Panning Techniques", PhD Thesis, Helsinki University of Technology, Espoo, Finland, 2001.

[rei03] Reiter, Ulrich; Schuldt, Michael; Dantele, Andreas: "Determination of Sound Source Obstruction in Virtual Scenes", Proceedings of the AES 24th International Conference on Multichannel Audio, Banff, Alberta, Canada, June 26-28, 2003, pp 201-206, ISBN 0-937803-50-2.

[rei05] Reiter, Ulrich: "An Input Device for Subjective Assessments of Bimodal Audio Visual Perception", in *Proceedings of the 2005 IEEE 9th International Symposium on Consumer Electronics (ISCE 2005)*, Macau, SAR, June 14-16, 2005.

[rei05a] Reiter, Ulrich: "Audio Rendering System Design for an Object Oriented Audio Visual Human Perception Assessment Tool", Proc. of the 8th Int. Conference on Digital Audio Effects (DAFx'05), Madrid, Spain, September 20-22, 2005.

[rei06] Reiter, Ulrich; Großmann, Sebastian; Strohmeier, Dominik; Exner, Markus: "Observations on Bimodal Audio Visual Subjective Assessments", Proceedings of the 120th AES Convention, Paris, France, May 20-23, 2006, Preprint 6852.

[rei07] Reiter, Ulrich; Kühhirt, Uwe: "Object-Based A/V Application Systems: IAVAS I3D Status and Overview", in *Proceedings of the 2007 IEEE 11th International Symposium on Consumer Electronics (ISCE 2007)*, Dallas, TX, USA, June 20-23, 2007.

[rim98] Rimell, Andrew N.; Hollier, M.; Voelcker, R.: "The Influence of Cross-Modal Interaction on Audio-Visual Speech Quality Perception", AES 105th Convention, 1998, Preprint 4791.

[roe95] Roederer, Juan G.: "Physikalische und psychoakustische Grundlagen der Musik", 3rd Edition (German language), Springer Verlag Berlin, 2000, ISBN 3-540-61370-6

[rum02] Rumsey, Francis: "Spatial Quality Evaluation for Reproduced Sound: Terminology, Meaning, and a Scene-Based Paradigm", J. Audio Eng. Soc., Vol. 50, No. 9, 2002 September, pp 651-666.

[sav97] Savioja, Lauri; Huopaniemi, Jyri; Lokki, Tapio; Väänänen, Riitta: "Virtual Environment Simulation - Advances in the DIVA Project", Proc. Int. Conf. Auditory Display (ICAD'97), Palo Alto, California, USA, Nov. 3-5, 1997, pp 43-46.

[sav99a] Savioja, Lauri; Huopaniemi, Jyri; Lokki, Tapio; Väänänen, Riitta: "Creating Interactive Virtual Acoustic Environments", Journal of the Audio Engineering Society, Vol. 47, No. 9, pp 675-705, September 1999.

[sav99b] Savioja, Lauri: "Modeling Techniques for Virtual Acoustics", PhD Thesis, Helsinki University of Technology, 1999, ISBN 951-22-4765-8.

[sca05] Scavone, Gary P.; Cook, Perry R.: "RtMidi, RtAudio, and a Synthesis ToolKit (STK) Update", Proceedings of the 2005 International Computer Music Conference (ICMC), Barcelona, Spain, September 5-9, 2005.

[sch62] Schroeder, M.R.: "Natural sounding artificial reverberation", Journal of the Audio Engineering Society, vol. 10(3), 1962.

[sch70] Schroeder, M.R.: "Digital Simulation of Sound Transmission in Reverberant Spaces" (Part 1), Journ. Acoust. Soc. Amer., 47(2):424-431, 1970.

[sch96] Schmalstieg, D.: "LODESTAR: An Octree-Based Level of Detail Generator for VRML", Technical Report, Institute of Computer Graphics, Visualization and Animation Group, Vienna University of Technology, 1996.

[sch99a] Scheirer, Eric D.; Vercoe, Barry L.: "SAOL: The MPEG-4 Structured Audio Orchestra Language", Computer Music Journal, Vol. 23, No. 2, June 1999, pp 31-50.

[sch99] Scheirer, Eric D.; Väänänen, Riitta; Huopaniemi, Jyri: "AudioBIFS: Describing Audio Scenes with the MPEG-4 Multimedia Standard", IEEE Transactions on Multimedia, Vol. 1, No. 3, September 1999, pp 237-250.

[sch06] Schauer, Carsten: "Modellierung primärer multisensorischer Mechanismen der räumlichen Wahrnehmung", PhD Thesis (German language), TU Ilmenau, Germany, 2006.

[sch06a] Schröder, Dirk; Lentz, Tobias: "Real-Time Processing of Image Sources Using Binary Space Partitioning", Journal of the Audio Engineering Society, vol. 54(7/8), July/August 2006, pp 604-619.

[sha00] Shams, Ladan; Kamitani, Yukiyasu; Shimojo, Shinsuke: "What You See is What You Hear", Nature, Vol. 408, 14 December 2000, p 788.

[sha02] Shams, Ladan; Kamitani, Yukiyasu; Shimojo, Shinsuke: "Visual Illusion Induced by Sound", Cognitive Brain Research, Vol. 14, 2002, pp 147-152.

[she94] Sheridan, Thomas B.: "Further Musings on the Psychophysics of Presence", In *Presence*, 5/1994, pp 241-246.

[shi96] Shi, Yuan: "Reevaluating Amdahl's Law and Gustafson's Law", Computer and Information Sciences department, Temple University, Web link http://www.cis.temple.edu/~shi/docs/amdahl/amdahl.html, October 1996, last visited 2007-09-12.

[smi81] Smith, Dave; Wood, Chet: "The 'USI', or Universal Synthesizer Interface", AES 70th Convention, New York, NY, USA, October 30 - November 2, 1981, Preprint 1845.

[sne98] Snell, Richard S.; Lemp, Michael A.: "Clinical Anatomy of the Eye", 2nd Edition, Blackwell Science Inc., Malden, MA, USA, 1998, ISBN 0-632-04344-X.

[spe01] Spence, Charles; Nicholls, Michael E.R.; Driver, John: "The Cost of Expecting Events in the Wrong Sensory Modality", Perception & Psychophysics, 63(2), 2001, pp 330-336.

[spr06] Springer, Jan P.; Sladeczek, Christoph; Scheffler, Martin; Hochstrate, Jan; Melchior, Frank; Fröhlich, Bernd: "Combining Wave Field Synthesis and Multi-Viewer Stereo Displays", presented at the IEEE VR2006 Conference, Alexandria, VA, USA, March 25-29, 2006.

[ste46] Stevens, Stanley S.: "On the Theory of Scales of Measurement", Science, Vol. 103. no. 2684, June 1946, pp 677-680.

[ste92] Steuer, Jonathan: "Defining Virtual Reality: Dimensions Determining Telepresence", Journal of Communication, 42/4, 1992, pp 73-93.

[ste00] Steinman, Scott B.; Steinman, Barbara A.; Garcia, Ralph Philip: "Foundations of Binocular Vision - A Clinical Perspective", McGraw-Hill New York, 2000, ISBN 0-8385-2670-5.

[str98] Strauss, Anselm; Corbin, Juliet M.: "Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory", 2nd Edition, Sage Publications Ltd, Thousand Oaks, CA, USA, 1998, ISBN 0-803-95939-7.

[str07] Strohmeier, Dominik: "Wahrnehmungsuntersuchung von 2D vs. 3D Displays in A/V-Applikationen mittels einer kombinierten Analysemethodik", Diploma Thesis (German language), TU Ilmenau, Germany, 2007.

[sty97] Styles, Elizabeth A.: "The Psychology Of Attention", The Psychology Press, Hove, England, 1997.

[sux04] Su, Alvin W.Y.; Xiao, Yi-Song; Yeh, Jia-Lin; Wu, Jian-Lung: "Real-Time Internet MPEG-4 SA Player and the Streaming Engine", AES 116th Convention, Berlin, Germany, May 8-11, 2004.

[suz01] Suzuki, Kenji; Martens, William L.: "Subjective Evaluation of Room Geometry in Multichannel Spatial Sound Reproduction: Hearing Missing Walls in Simulated Reverberation", Proc. 11th International Conference on Artificial Reality and Telexistence ICAT2001, Tokyo, Japan, Dec. 5-7, 2001.

[ter98] Terhadt, Ernst: "Akustische Kommunikation" (German language), Springer Verlag Berlin, 1998, ISBN 3-540-63408-8.

[tho01] Thompson, S.; Shams, L.; Kamitani, Y.; Shimojo, S.: "Brain Mechanisms Underlying a Sound-Induced Visual Illusion", in *Society for Neuroscience Abstracts*, vol. 27, 2001, p 1342.

[tre80] Treisman, Anne M.; Gelade, Garry: "A Feature-Integration Theory of Attention", Cognitive Psychology, 12(1), January 1980, pp 97-136.

[uni06] EC Sixth Framework Programme, Priority IST-200202.3.1.8, Project title "A Distributed Interactive Audio-Visual Virtual Reality System", Deliverable D 7.4, Uni-Verse Sound Renderer, March 15th, 2006.

[vää97] Väänänen, R., Välimäki, V., Huopaniemi, J, and Karjalainen, M.: "Efficient and Parametric Reverberator for Room Acoustics Modeling", Int. Comp. Music Conf. (ICMC'97), Thessaloniki, Greece, 1997.

[ver99] Verschure, Paul F. M. J.; König, Peter: "On the Role of Biophysical Properties of Cortical Neurons in Binding and Segmentation of Visual Scenes", Neural Computation 11, 1999, pp. 1113-1138.

[vor88] Vorländer, Michael: "Die Genauigkeit von Berechnungen mit dem raumakustischen Schallteilchenmodell und ihre Abhängigkeit von der Rechenzeit", (German language), Acustica, vol. 66, 1988, pp. 90-96.

[vor89] Vorländer, Michael: "Untersuchungen zur Leistungsfähigkeit des raumakustischen Schallteilchenmodells", PhD Thesis (German language), Institut für Technische Akustik, RWTH Aachen, Germany, 1989.

[vos06] Voswinkel, Florian: "Implementierung und Untersuchung flächiger Schallquellen für Echtzeit-Audiorendering in einem MPEG-4 Player", Diploma Thesis (German language), TU Ilmenau, Germany, 2006.

[vrc06] VRCA - Virtual Reality Center Aachen, Jahresbericht 2005/2006 (German language), RWTH Aachen, Germany, October 2006.

[wad01] Wade, Nicholas J.; Swanston, Michael T.: "Visual Perception: An Introduction", 2nd edition, Psychology Press, Hove, East Sussex, UK, 2001, ISBN 1-84169-204-2.

[wag03] Wager, Tor D.; Phan, Luan K.; Liberzon, Israel; and Taylor, Stephan F.: "Valence, gender, and lateralization of functional brain anatomy in emotion: a meta-analysis of findings from neuroimaging", NeuroImage 19, 2003, pp 513-531.

[wei07] Weitzel, Mandy: "Einfluss von Interaktion auf die Qualitätswahrnehmung von audiovisuellen Szenen", Diploma Thesis (German language), TU Ilmenau, Germany, 2007.

[wei07a] Weitzel, Mandy: unpublished test report, protocols and annotations, TU Ilmenau, Germany, 2007.

[wen98] Wenzel, Elizabeth M.: "The Impact of System Latency on Dynamic Performance in Virtual Acoustic Environments", Proc. of the 15th International Congress on Acoustics and 135th Meeting of the Acoustical Society of America, Seattle, WA, USA, 1998, pp 2405-2406.

[wen99] Wenzel, Elizabeth M.: "Effect of Increasing System Latency on Localization of Virtual Sounds", Proc. of the AES 16th International Conference on Spatial Sound Reproduction, Rovaniemi, Finland, 1999, pp 42-50.

[wen00] Wenzel, Elizabeth M.; Miller, Joel D.; Abel, Jonathan S.: "A software-based system for interactive spatial sound synthesis", Proceedings of the International Conference on Auditory Display, ICAD 2000, Atlanta, GA, April 2000, pp 151-156.

[wen01] Wenzel, Elizabeth M.: "Effect of Increasing System Latency on Localization of Virtual Sounds with Short and Long Duration", Proc. of the 2001 International Conference on Auditory Display, Espoo, Finland, July 29 - August 1, 2001.

[wer23] Wertheimer, Max: "Untersuchungen zur Lehre von der Gestalt II", in *Psychologische Forschung*, 4, 1923, pp 301-350.

[wit98] Witmer, Bob G.; Singer, Michael J.: "Measuring Presence in Virtual Environments: A Presence Questionnaire", In *Presence*, 7/1998, pp 225-240.

[wos95] Woszczyk, Wieslaw; Bech, Soren; Hansen, Villy: "Interactions Between Audio-Visual Factors in a Home Theater System: Definition of Subjective Attributes", AES Audio Engineering Society 99th Convention, Preprint 4133, New York, USA, 1995.

[you97] Young, Paul A.; Young, Paul H.: "Basic Clinical Neuroanatomy", Williams & Wilkins, Baltimore, 1997, ISBN 0-683-09351-7.

[zie03] Zielinski, Slawomir; Rumsey, Francis; Bech, Soren; de Bruyn, Bart; Kassier, Rafael: "Computer Games and Multichannel Audio Quality - the Effect of Division of Attention Between Auditory and Visual Modalities", Proc. AES 24th International Conference on Multichannel Audio, Banff, Alberta, Canada, June 2003

[zoz98] Zollner, Manfred; Zwicker, Eberhard: "Elektroakustik", 3rd Edition (German language), 1st corr. Reprint, Springer Verlag, Berlin, 1998, ISBN 3-540-64665-5.

[zwi99] Zwicker, Eberhard; Fastl, Hugo: "Psychoacoustics - Facts and Models", 2nd updt. ed., Springer Verlag, Berlin, 1999, ISBN 3-540-65063-6.

Sources on the World Wide Web:

[w-3dc] The 3DConnexion SpaceNavigator 3d mouse product homepage, 2007-11-21, http://www.3dconnexion.com/3dmouse/spacenavigator.php

[w-ada] Acoustic Design Ahnert, 2007-10-16, http://www.ada-acousticdesign.de/english/software/index_swr.html

[w-als] Advanced Linux Sound Architecture, 2007-06-09, http://www.alsa-project.org/

[w-amei] Association of Musical Electronics Industry, 2007-07-29, http://www.amei.or.jp/

[w-asi] Steinberg Media Technologies AG, "ASIO 2.2 Audio Streaming Input Output Development Kit, Documentation Release 2", 2007-06-09, http://www.steinberg.net

[w-boo] boost C++ Libraries, 2007-07-12, http://www.boost.org/

[w-cat] Computer Aided Theater Technique - CATT-Acoustic v8.0, 2007-10-16, http://www.catt.se/CATT-Acoustic.htm

[w-cla] CLAM - C++ Library for Audio and Music, 2007-02-11, http://mtg.upf.edu/clam/

[w-col] Colvin, Greg; Dawes, Beman; Dimov, Peter; Adler, Darin: BOOST.ORG (Ed.): "Boost.SmartPointer Documentation", 2005, 2007-06-13, http://www.boost.org/libs/smart_ptr/smart_ptr.htm

[w-div] DIVA - Auralization homepage, 2007-01-27, http://www.tml.tkk.fi/Research/DIVA/past/acoustics.html

[w-eve] The EVE - Acoustics and Audio Webpage, 2007-01-27, http://eve.hut.fi/aural.html

[w-ffm] Fast Forward MPEG - The FFmpeg Homepage, 2007-11-21, http://ffmpeg.mplayerhq.hu/

[w-gpac] GPAC Project on Advanced Content Homepage, 2007-12-12, http://gpac.sourceforge.net

[w-gra] Online Version of Henry Gray's *Anatomy of the Human Body*, 20th edition, 1918/2000, 2007-03-25, http://www.bartleby.com/107/

[w-hyp] Hyperphysics Web Site, Georgia State University, 2007-06-13, http://hyperphysics.phy-astr.gsu.edu/hbase/sound/hearcon.html

[w-iem] Zmoelnig, Johannes W.; Ritsch, Winfried; Sontacchi, Alois: "The IEM-Cube", preliminary paper about the IEM-Cude, 2007-10-10, http://iem.at/~zmoelnig/publications/cube_en.pdf

[w-ietf] The Internet Engineering Task Force Tools Page, Common Format and MIME Type for Comma-Separated Values (CSV) Files (RFC 4180), 2007-08-10, http://tools.ietf.org/html/rfc4180

[w-int] The Interactivity Consultants Homepage, 2007-11-20, http://www.interactivityconsultants.com/pages/resources/interactivity_definition_and_resources.htm

[w-ios] IOSONO - The Art of Natural Sound Homepage, 2007-12-12, http://www.iosono-sound.com/index.html

[w-joy] The Joytech Europe Homepage, Dancing Mate product information, 2007-11-21, http://www.joytech.net/8/products/ps2/1/84/Dancing-Mate-1-Player.htm

[w-lak] Lake Technologies homepage, 2007-02-09, http://www.lake.com.au/Lake_Huron.htm

[w-lgpl] GNU Lesser General Public License, 2007-07-30, http://www.gnu.org/licenses/lgpl.html

[w-m4if] MPEG Industry Forum, MPEG-4 FAQ, 2007-07-14, http://www.m4if.org/resources/mpeg4userfaq.php

[w-mew07] Merriam-Webster's Online Dictionary, 2007-11-20, http://www.m-w.com/dictionary/interactive

[w-mio] Midibox Hardware Platform and MIOS MIdi Operating System for Microchip PIC micro-controller devices, by Thorsten Klose, 2007-02-22, http://www.ucapps.de/

[w-mma] MIDI Manufacturers Association, 2007-07-29, http://www.midi.org/about-midi/specshome.shtml

[w-msd] Microsoft Developer Network, MSDN Library, DirectX SDK, 2007-06-09, http://msdn2.microsoft.com/en-us/library/bb219818.aspx

[w-ode] http://www.odeon.dk

[w-pol] The FASTRAK Motion Tracking System, 2007-10-16, http://www.polhemus.com/?page=Motion_Fastrak

[w-por] PortAudio: an Open-Source Cross-Platform Audio API., 2007-10-16, http://www.portaudio.com.

[w-r] The R Project for Statistical Computing, 2007-08-07, http://www.r-project.org/

[w-rtm] RtMidi, a cross-platform API for real time MIDI input / output, 2007-08-04, http://www.music.mcgill.ca/~gary/rtmidi/

[w-son] Sonic Emotion Homepage, 2007-12-12, http://www.sonicemotion.com/

[w-spss] SPSS Statistical Tools Website, 2007-08-07, http://www.spss.com/scientific_research/index.htm

[w-sta] The Stanford Bunny, The Stanford 3D Scanning Repository, 2007-11-28, http://www-graphics.stanford.edu/data/3Dscanrep/; pictures taken and modified from http://www.cs.umd.edu/∼djacobs/CMSC427/Modeling.pdf

[w-wii] The Nintendo Wii Remote Controller product information page, 2007-11-21, http://wii.nintendo.com/controller.jsp

[w-wrap] Wrapper Library for Windows MIDI API, by Leslie Sanford, 2007-07-30, http://www.codeproject.com/audio/MIDIWrapper.asp

[w-xml] The W3C Extensible Markup Language (XML) Website, 2007-08-09, http://www.w3.org/XML/

Part VI

Appendix

A Additional Assessment Documentation

This section gives additional documentation related to the assessments performed. It reproduces the instructions presented to the test subjects before the assessments (and sometimes between trials, see the respective section on the assessment itself). As all participants were native German speakers, the instructions were given in German language.

A.1 Ass.: Optimum Number of Loudspeakers (8.7)

Wahrnehmungstest Lautsprecheraufstellungen bei Bewegtbild

Sehr geehrte Testteilnehmer,

im Folgenden werden Ihnen 12 jeweils zweiminütige Teststücke vorgespielt. Diese bestehen aus einem bewegten Videobild (vorgegebene, einfache Bewegungspfade durch einen virtuellen Raum) und einer zugehörigen Mehrkanal-Audioaufnahme. Die Teststücke unterscheiden sich zum einen durch unterschiedliche Nachhall-Qualitäten (diese sind **nicht** zu beurteilen), zum anderen durch unterschiedliche Bewegungspfade im virtuellen Raum.

Mit Hilfe der vor Ihnen stehenden Bedieneinheit sollen Sie pro Teststück („Trial“) drei angebotene Lautsprecheraufstellungen für die Tonwiedergabe bewerten, und zwar im Vergleich zu einer Referenz. Dabei geht es um die Richtung, aus der die Schallquelle zu hören ist. **Wie gut ist die Lokalisierbarkeit der Schallquelle insgesamt? Wie gut ist die Position der Schallquelle wahrnehmbar?** Beurteilen Sie dabei **nicht** die Qualität des Audiosignals oder des Nachhalls, sondern die Lokalisierbarkeit!

Ihre Bewertung können Sie über die Schieberegler stufenlos abgeben.

Zur Bedienung der Faderbox:

- **PLAY** startet die Bild- und Tonwiedergabe mit einer systembedingten Verzögerung von bis zu 5 Sekunden
- **STOP** stoppt die Wiedergabe
- **NEXT** springt zum nächsten Teststück („Trial“), wenn Sie Ihre Bewertungen für alle Items abgegeben haben

- **REFERENZ** wählt das Referenz-Item an
- **ITEM 2**, **ITEM 3**, **ITEM 4** wählen jeweils eine andere Lautsprecheraufstellung („Item“) an

Über das Display der Bedieneinheit können Sie sehen, an welcher Stelle im Test Sie sich gerade befinden („Trial x – Item y“). Ebenso zeigen LEDs über den Schiebereglern an, welches Item gerade zu hören ist.

Wenn ein Teststück nach 2 Minuten beendet ist, so verstummt die Tonwiedergabe, das Bild des virtuellen Raumes verschwindet und der Projektor zeigt verschwommen eine Programmoberfläche an.
Wollen Sie das **Teststück erneut hören**, so halten Sie es mit **STOP an und starten es danach mit PLAY wieder** von vorn.
Haben Sie **alle Bewertungen abgegeben** und wollen zum nächsten Teststück springen, so drücken Sie **NEXT**.

Vielen Dank!

A.2 Ass.: Number of Internal Workchannels for Perceptual Approach (8.8)

TESTANLEITUNG

Medienprojekt
Echtzeitberechnung der Raumakustik mittels MPEG-4 /
Perceptual Approach für interaktive AV-Anwendunassysteme

Herzlich willkommen zu diesem bimodalen Test, der im Rahmen unseres Medienprojektes stattfindet. Wir danken Ihnen vorab für Ihrer Teilnahme.

Hintergrund
Im IAVAS MPEG-4 Player des Instituts für Medientechnik existieren zwei Möglichkeiten die Akustik im virtuellen Raum zu simulieren. Der Physical Approach beschreibt die Raumakustik auf Basis geometrischer und physikalischer Parameter. Der Perceptual Approach beruht hingegen auf einer wahrnehmungsbezogenen Beschreibung der Akustik.
Ziel dieses Tests ist es die von uns implementierten Modelle des Perceptual Approach hinsichtlich ihrer Klangqualität zu untersuchen. Hierbei soll nicht die Übereinstimmung der Hallsimulation mit der visuellen Szene im Vordergrund stehen, sondern konkret auf folgendes Kriterium eingegangen werden:

„Beurteilen Sie die klangliche Qualität der Hallsimulation!“

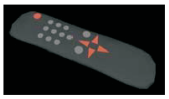
Der Test wird mit Hilfe eines AB-Paarvergleiches durchgeführt. Dabei werden immer zwei Modelle gegenübergestellt und anschließend mit dem Längsfader bewertet.

Testszenario
Da es sich um einen bimodalen Test handelt, werden Sie nicht nur Hören, sondern auch Sehen.
Der Test findet in drei virtuellen Räumen (Wohnzimmer, Seminarraum, Sporthalle) unterschiedlicher Größe statt. In jedem Raum werden Sie ein Musikstück und eine Sprachaufnahme beurteilen. Zusätzlich ist ihre Interaktion gefragt. In den Räumen befindet sich jeweils ein Gegenstand den es einzusammeln gilt.

TESTANLEITUNG

Medienprojekt
Echtzeitberechnung der Raumakustik mittels MPEG-4 /
Perzeptual Approach für interaktive AV-Anwendunassysteme

Testablauf
Zur Durchführung des Wahrnehmungstests stehen Ihnen eine Maus und eine Faderbox als Eingabegeräte zur Verfügung.
Nach Betätigung der „weiter" Taste auf der Faderbox befinden Sie sich im ersten Raum. Zunächst wird Modell A gespielt. Es folgen eine kurze Pause und Modell B. Das gespielte Modell (A oder B) kann zur Orientierung am Bildschirm im virtuellen Raum abgelesen werden.
Sie navigieren mit der Maus durch den Raum mit dem Ziel die folgenden Gegenstände einzusammeln:

im Wohnzimmer die Fernbedienung 

im Seminarraum das Matheheft 

in der Sporthalle den Fussball 

Anschließend geben Sie über den Fader auf der Faderbox Ihre Bewertung zur Hallsimulation ab. Für ein erneutes Hören des Paares drücken Sie „Wiederholung" und bewerten Sie danach.
Starten Sie nun mit „weiter" den nächsten Paarvergleich. In jedem Raum werden Ihnen insgesamt 8 Testpaare vorgespielt.

Trainingsphase
Für ein besseres Verständnis erfolgt vorab ein kurzes Training in einem der Räume. Bei Fragen stehen wir Ihnen zur Verfügung.

Ein angenehmes Hörvergnügen wünschen
Mandy und Andreas

A.3 Ass.: Influence of Interaction on Perceived Quality (8.9)

Before starting the assessment session, subjects were asked to fill in a pre-test questionnaire. This questionnaire is shown in the following:

Alter: _____ Jahre

Geschlecht: [ ] weiblich
[ ] männlich

Treffen die folgenden Sätze auf Sie zu?
Bitte kreuzen Sie die am besten passende Antwort an.

	Nein	Ja
Ich studiere oder arbeite im weiteren Bereich der Computer-Wissenschaften.		
In Studium, Beruf oder Hobby beschäftige ich mich häufig mit Video.		
Ich spiele regelmäßig ein Musikinstrument.		
Ich spiele nicht regelmäßig Computerspiele.		
In Studium, Beruf oder Hobby beschäftige ich mich nicht mit Audio.		
Ich nehme regelmäßig als Testperson an Audioqualitätstests teil.		
Ich habe keinerlei Hörbeeinträchtigung.		
Ich habe Erfahrung im bewussten Hören von Mehrkanalton (Surround).		

Wie gut stimmen Sie mit folgenden Aussagen überein?
Bitte kreuzen Sie die am besten passende Aussage an.

	Stimme überhaupt nicht zu	Stimme nicht zu	Weiß nicht	Stimme zu	Stimme voll ung ganz zu
1. Ich erwarte eine gute Klangqualität im Hörlabor.					
2. Üblicherweise bin ich in meinem Freundeskreis einer der ersten, der neue interessante Produkte kauft.					
3. Ich habe Interesse daran, neu auf dem Markt befindliche technische Produkte, die für mich interessant sind, zu erwerben.					
4. Im Vergleich zu meinen Freunden besitze ich relativ wenige technische Geräte.					
5. Üblicherweise bin ich in meinem Freundeskreis einer der letzten, die über technische Neuheiten Bescheid wissen.					
6. Ich kaufe keine technischen Geräte, bevor ich sie nicht gründlich ausprobiert habe.					
7. Ich kaufe technische Geräte früher als andere Personen in meinem Bekanntenkreis.					

Ich stimme zu, dass im Anschluß an den Test ein kurzes Interview mit mir auf Tonband aufgezeichnet wird. Das Tonbandmaterial wird anonym ausgewertet und dient ausschließlich wissenschaftlichen Zwecken.

_____ _____

Datum Unterschrift

Before the assessment, subjects were presented with an anchoring session. For each of the three consecutive assessment sessions, separate instructions were provided.

Anchoring

Unter „Anchoring“ versteht man die Vorstellung der Extremwerte der im weiteren Test angebotenen Qualitätsstufen.

Im Folgenden werden Ihnen zweimal drei Beispielszenen vorgespielt, deren (unterschiedliche) Gesamtqualität den Extremwerten entspricht. Es wird zunächst das eine Extrem der Qualitätsskala, dann das andere, und schließlich wieder das eine (erste) Beispiel demonstriert.

1) Musikbeispiel, Bewegungspfad 1

2) Sprachbeispiel, Bewegungspfad 2

Die Abfolge der insgesamt sechs Beispielszenen erfolgt automatisch, jeweils unterbrochen von einer kleinen Pause, in der eine graue Leinwand zu sehen sein wird.

Training 1

Unter „Training“ versteht man die Vorbereitung des Testteilnehmers auf die bevorstehende Aufgabe.

Sie werden sich in den 12 nun folgenden Szenen automatisch durch eine Sporthalle bewegen. Jede Szene ist ca. 30 Sekunden lang.

Aufgabe:

Ihre Aufgabe besteht darin, die subjektive Gesamtqualität der jeweiligen Szene zu beurteilen.

Dazu sollen Sie mit Hilfe des Schiebereglers auf der vor Ihnen stehenden Faderbox nach dem Ende jeder Szene deren Gesamtqualität beurteilen.

Die Skaleneinteilung beschreibt die Gesamtqualität von 0 - 100:

0	=	niedrigste Gesamtqualität
100	=	höchste Gesamtqualität

Haben Sie den Schieberegler entsprechend Ihrer Bewertung positioniert, so drücken Sie bitte einmal die Taste *Bewertung*. Die Leinwand wird grau, und nach kurzer Zeit beginnt automatisch die nächste Szene.

Training 2

Unter „Training“ versteht man die Vorbereitung des Testteilnehmers auf die bevorstehende Aufgabe.

Sie werden sich in den 12 nun folgenden Szenen automatisch durch eine Sporthalle bewegen. Jede Szene ist ca. 30 Sekunden lang.

Aufgabe:

Ihre Aufgabe während des Versuchsablaufs besteht darin, die Taste *Ball* zu drücken, sobald ein Ball neu auf der Bildfläche erscheint. Am Ende einer jeden Szene sollen Sie die subjektive Gesamtqualität der jeweiligen Szene beurteilen.

Dazu sollen Sie mit Hilfe des Schiebereglers auf der vor Ihnen stehenden Faderbox nach dem Ende jeder Szene deren Gesamtqualität beurteilen.

Die Skaleneinteilung beschreibt die Gesamtqualität von 0 - 100:

0	=	niedrigste Gesamtqualität
100	=	höchste Gesamtqualität

Haben Sie den Schieberegler entsprechend Ihrer Bewertung positioniert, so drücken Sie bitte einmal die Taste *Bewertung*. Die Leinwand wird grau, und nach kurzer Zeit beginnt automatisch die nächste Szene.

Training 3

Unter „Training“ versteht man die Vorbereitung des Testteilnehmers auf die bevorstehende Aufgabe.

Sie werden sich in den 12 nun folgenden Szenen frei mit Hilfe der Computermaus durch diese Sporthalle bewegen. Für eine Vorwärtsbewegung halten Sie die Maustaste gedrückt und bewegen die Maus nach vorne. Rückwärtsbewegungen erfolgen analog. Für Drehungen nach links oder rechts bewegen Sie die Maus bei gedrückter linker Maustaste nach links bzw. rechts. Jede Szene ist ca. 30 Sekunden lang.

Aufgabe:

Ihre Aufgabe während des Versuchsablaufs besteht darin, die Bälle einzusammeln, indem Sie sich ihnen so weit wie möglich annähern. Am Ende einer jeden Szene sollen Sie die subjektive Gesamtqualität der jeweiligen Szene beurteilen.

Dazu sollen Sie mit Hilfe des Schiebereglers auf der vor Ihnen stehenden Faderbox nach dem Ende jeder Szene deren Gesamtqualität beurteilen.

Die Skaleneinteilung beschreibt die Gesamtqualität von 0 - 100:

0	=	niedrigste Gesamtqualität
100	=	höchste Gesamtqualität

Haben Sie den Schieberegler entsprechend Ihrer Bewertung positioniert, so drücken Sie bitte einmal die Taste *Bewertung*. Die Leinwand wird grau, und nach kurzer Zeit beginnt automatisch die nächste Szene.

After the interviews, subjects were asked to fill in a post-test questionnaire[1].

3. Kannten Sie das präsentierte Material schon vor dem Test?
Bitte kreuzen Sie die am besten passende Aussage an.

	Ja	Nein
Musikstück	[ ]	[ ]
Sprachbeispiel	[ ]	[ ]
Visuelle Szene	[ ]	[ ]

4. Hatten Sie den Eindruck, dass einige der zu beurteilenden Sequenzen von höherer Qualität als andere waren?
Bitte kreuzen Sie die am besten passende Aussage an.

[ ] Ja, ich denke, dass einige Sequenzen eine höhere Qualität als andere hatten.

Bitte kreuzen Sie die am besten passende Aussage an, um die Sequenzen in eine Rangfolge zu bringen.

	schlechteste	mittlere	beste
Sequenzen:			
Hören und zusehen	[ ]	[ ]	[ ]
Hören und Taste drücken	[ ]	[ ]	[ ]
Hören und Ball einsammeln	[ ]	[ ]	[ ]

[ ] Nein, ich denke, dass alle Sequenzen die gleiche Audioqualität hatten.

5. Hatten Sie den Eindruck, dass die Beurteilung der Sequenzen unterschiedlich schwierig war?
Bitte kreuzen Sie die am besten passende Aussage an.

[ ] Ja, ich denke dass einige Sequenzen einfacher zu beurteilen waren als andere.

Bitte kreuzen Sie die am besten passende Aussage an, um die Sequenzen in eine Rangfolge zu bringen.

	am leichtesten	mittel	am schwierigsten
Sequenz:			
Hören und Taste drücken	[ ]	[ ]	[ ]
Hören und zusehen	[ ]	[ ]	[ ]
Hören und Ball einsammeln	[ ]	[ ]	[ ]

[ ] Nein, ich denke, dass alle Sequenzen gleich leicht oder schwer zu beurteilen waren.

[1] Questions 1 and 2 were part of the pre-test questionnaire.

6. Hatten Sie den Eindruck, dass eines der Tonbeispiele (Sprache oder Musik) leichter zu beurteilen war?
Bitte kreuzen Sie die am besten passende Aussage an.

[ ] Ja, ich denke, dass eines der Tonbeispiele leichter zu beurteilen war.
Bitte kreuzen Sie die am besten passende Aussage an, um die Tonbeispiele in eine Rangfolge zu bringen.

	am leichtesten	am schwierigsten
Tonbeispiel:		
Musikstück	[ ]	[ ]
Sprachbeispiel	[ ]	[ ]

[ ] Nein, ich denke, dass alle Tonbeispiele gleich leicht oder schwer zu beurteilen waren.

7. Hatten Sie den Eindruck, dass eines der Tonbeispiele eine höhere Qualität als das andere hatte?
Bitte kreuzen Sie die am besten passende Aussage an.

[ ] Ja, ich denke, dass eines der Tonbeispiele eine höhere Qualität hatte.
Bitte kreuzen Sie die am besten passende Aussage an, um die Tonbeispiele in eine Qualitätsrangfolge zu bringen.

	beste	schlechteste
Tonbeispiel:		
Musikstück	[ ]	[ ]
Sprachbeispiel	[ ]	[ ]

[ ] Nein, ich denke, dass alle Tonbeispiel die gleiche Qualität hatten.

8. Bitte notieren Sie hier generelle Anmerkungen zum Test!

A.4 Ass.: Influence of a Working Memory Task on Perceived Quality (8.10)

Testanleitung - Anchoring

Herzlich willkommen zu diesem bimodalen Wahrnehmungstest und vielen Dank für Ihre Teilnahme. Der Test findet in einem virtuellen Foyer statt, in dem Sie die Länge der Nachhallzeit verschiedener auditiver Sequenzen im Vergleich zu einer Referenz bewerten sollen.

Der Versuch ist in drei Versuchsteile untergliedert, zu denen Sie jeweils eine Testanleitung erhalten und ein kurzes Training durchlaufen.

Zunächst werden Ihnen in einem Anchoring die Referenz und die Extremwerte der im weiteren Test angebotenen Nachhallzeiten präsentiert. Diese werden in der folgenden Reihenfolge abgespielt:

- Referenz
- erstes Extrem (Untergrenze der Bewertungsskala - 0, kürzeste Nachhallzeit)
- zweites Extrem (Obergrenze der Bewertungsskala - 100, längste Nachhallzeit)

Drücken Sie bitte immer die Anchoring Taste um die nächste auditive Sequenz zu hören. Anschließend wird diese Folge ein zweites Mal wiederholt.

Testanleitung - Navigation

Im folgenden Versuchsteil werden Ihnen insgesamt fünf Testpaare im virtuellen Foyer präsentiert. Zunächst wird die Referenz gespielt. Es folgen eine kurze Pause und die zu bewertende auditive Sequenz. Vergleichen Sie bitte die empfundene Länge der Nachhallzeit der zweiten auditiven Sequenz mit der zuvor gehörten Referenz.

Die Audiosignale bestehen aus einer Folge von Zahlen, wobei nach jeweils 2 Sekunden eine neue Zahl abgespielt wird. Während der Wiedergabe nutzen Sie bitte die Maus um durch das Foyer zu navigieren, halten Sie dafür die linke Maustaste gedrückt und bewegen Sie die Maus in die gewünschte Richtung.

Nach der Wiedergabe eines Testpaares geben sie über den Schieberegler auf der vor Ihnen stehenden Faderbox ihre Bewertung ab. Auf einer Skala von 0 (kürzeste Nachhallzeit) bis 100 (längste Nachhallzeit) können Sie die Länge der Nachhallzeit abschätzen. Nach der Positionierung des Schiebereglers drücken Sie bitte die Taste weiter. Es folgt ein grauer Bildschirm und automatisch die Wiedergabe des nächsten Testpaares.

Testanleitung - 1back

Im folgenden Versuchsteil werden Ihnen insgesamt fünf Testpaare im virtuellen Foyer präsentiert. Zunächst wird die Referenz gespielt. Es folgen eine kurze Pause und die zu bewertende auditive Sequenz. Vergleichen Sie bitte die empfundene Länge der Nachhallzeit der zweiten auditiven Sequenz mit der zuvor gehörten Referenz.

Die Audiosignale bestehen aus einer Folge von Zahlen, wobei nach jeweils 2 Sekunden eine neue Zahl abgespielt wird. Während der Wiedergabe konzentrieren Sie sich auf eine 1-back Aufgabe. Immer wenn Sie eine Zahl hören, vergleichen Sie diese mit der letzten Zahl. Stimmen die Zahlen überein, drücken Sie die Taste "ja", bei keiner Übereinstimmung die Taste "nein".

Beispiel:	Zahlenfolge:	1	2	2	3
	Taste:		nein	ja	nein

Zusätzlich nutzen Sie bitte die Maus um durch das Foyer zu navigieren, halten Sie dafür die linke Maustaste gedrückt und bewegen Sie die Maus in die gewünschte Richtung.

Nach der Wiedergabe eines Testpaares geben sie über den Schieberegler auf der vor Ihnen stehenden Faderbox ihre Bewertung ab. Auf einer Skala von 0 (kürzeste Nachhallzeit) bis 100 (längste Nachhallzeit) können Sie die Länge der Nachhallzeit abschätzen. Nach der Positionierung des Schiebereglers drücken Sie bitte die Taste weiter. Es folgt ein grauer Bildschirm und automatisch die Wiedergabe des nächsten Testpaares.

Testanleitung - 2back

Im folgenden Versuchsteil werden Ihnen insgesamt fünf Testpaare im virtuellen Foyer präsentiert. Zunächst wird die Referenz gespielt. Es folgen eine kurze Pause und die zu bewertende auditive Sequenz. Vergleichen Sie bitte die empfundene Länge der Nachhallzeit der zweiten auditiven Sequenz mit der zuvor gehörten Referenz.

Die Audiosignale bestehen aus einer Folge von Zahlen, wobei nach jeweils 2 Sekunden eine neue Zahl abgespielt wird. Während der Wiedergabe konzentrieren Sie sich auf eine 2-back Aufgabe. Immer wenn Sie eine Zahl hören, vergleichen Sie diese mit der vorletzte Zahl. Stimmen die Zahlen überein, drücken Sie die Taste "ja", bei keiner Übereinstimmung die Taste "nein".

Beispiel:	Zahlenfolge:	1	2	3	2	3
	Taste:			nein	ja	ja

Zusätzlich nutzen Sie bitte die Maus um durch das Foyer zu navigieren, halten Sie dafür die linke Maustaste gedrückt und bewegen Sie die Maus in die gewünschte Richtung.

Nach der Wiedergabe eines Testpaares geben sie über den Schieberegler auf der vor Ihnen stehenden Faderbox ihre Bewertung ab. Auf einer Skala von 0 (kürzeste Nachhallzeit) bis 100 (längste Nachhallzeit) können Sie die Länge der Nachhallzeit abschätzen. Nach der Positionierung des Schiebereglers drücken Sie bitte die Taste weiter. Es folgt ein grauer Bildschirm und automatisch die Wiedergabe des nächsten Testpaares.

A.5 Ass.: Influence of a Working Memory Task on Perceived Quality, II (8.11)

Testanleitung - Überblick

Herzlich Willkommen zu diesem bimodalen Wahrnehmungstest und vielen Dank für Ihre Teilnahme. Der Test findet in einem virtuellen Foyer statt, in dem Sie die Länge der Nachhallzeit von auditiven Sequenzen bewerten sollen, während Sie sich auf ein Zahlenspiel und / oder auf die Navigation im Foyer konzentrieren.

Der Versuch ist in drei Teile untergliedert, zu denen Sie jeweils eine Testanleitung erhalten und ein kurzes Training durchlaufen. Die Versuchsteile variieren in der Art des Spiels, die Bewertung der Nachhallzeit erfolgt jedoch analog.

In den Versuchsteilen werden Ihnen acht Zahlenfolgen mit jeweils einer Länge von 30 Sekunden präsentiert, wobei immer nach zwei Sekunden die nächste Zahl abgespielt wird. Innerhalb einer Zahlenfolge wird die Länge der Nachhallzeit einmalig verändert. Die empfundene Länge dieser modifizierten Nachhallzeit im Vergleich zur vorherigen (Referenz) beurteilen Sie nach der Wiedergabe der Zahlenfolge. Verwenden Sie dazu den Schieberegler mit einer 5-stufigen Skala auf der vor Ihnen stehenden Faderbox. Nach der Positionierung des Schiebereglers drücken Sie bitte die Taste weiter. Es folgt ein grauer Bildschirm und automatisch die Wiedergabe der nächsten Zahlenfolge.

Vorab werden Sie sich mit den im späteren Test angebotenen Nachhallzeiten und der Bewertungsskala vertraut machen. Dafür befinden sich auf der Faderbox fünf Tasten, die entsprechend der Skalenstufen beschriftet sind. Bei der Taste Referenz handelt es sich um die Mittelstellung der Skala. Bei Tastendruck wird eine Zahl mit der jeweiligen Nachallzeit abgespielt.

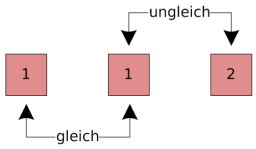
Testanleitung - Aufgabe 1

Während der Wiedergabe der auditiven Sequenz nutzen Sie bitte die Maus, um durch das Foyer zu navigieren. Halten Sie dafür die linke Maustaste gedrückt und bewegen Sie die Maus in die gewünschte Richtung.

Konzentrieren Sie sich im Laufe des Versuchsteils auf die Navigation!

Testanleitung - Aufgabe 2

Während der Wiedergabe der auditiven Sequenz, konzentrieren Sie sich auf das folgende Zahlenspiel: Immer wenn Sie eine Zahl hören, vergleichen Sie diese mit der vorangegangenen Zahl. Stimmen die Zahlen überein, drücken Sie die Taste "gleich", bei keiner Übereinstimmung die Taste "ungleich". Die nachfolgende Grafik verdeutlicht das Prinzip.

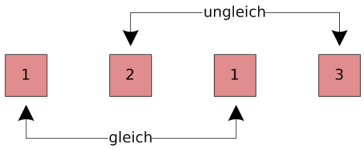


Zusätzlich nutzen Sie bitte die Maus, um durch das Foyer zu navigieren. Halten Sie dafür die linke Maustaste gedrückt und bewegen Sie die Maus in die gewünschte Richtung.

Konzentrieren Sie sich im Laufe des Versuchsteils auf das Zahlenspiel und die Navigation! Denn für jeden richtigen Tastendruck erhalten Sie einen Punkt. Die erreichte Punktzahl wird am Ende jeder Zahlenfolge im Display der Faderbox angezeigt.

Testanleitung - Aufgabe 3

Während der Wiedergabe der auditiven Sequenz, konzentrieren Sie sich auf das folgende Zahlenspiel: Immer wenn Sie eine Zahl hören, vergleichen Sie diese mit der vor-vorangegangenen Zahl. Stimmen die Zahlen überein, drücken Sie die Taste "gleich", bei keiner Übereinstimmung die Taste "ungleich". Die nachfolgende Grafik verdeutlicht das Prinzip.



Zusätzlich nutzen Sie bitte die Maus, um durch das Foyer zu navigieren. Halten Sie dafür die linke Maustaste gedrückt und bewegen Sie die Maus in die gewünschte Richtung.

Konzentrieren Sie sich im Laufe des Versuchsteils auf das Zahlenspiel und die Navigation! Denn für jeden richtigen Tastendruck erhalten Sie einen Punkt. Die erreichte Punktzahl wird am Ende jeder Zahlenfolge im Display der Faderbox angezeigt.

A.6 Ass.: Influence of Visual Interaction on Perceived Audio Quality (8.12)

Testanleitung - Überblick

In diesem Test haben Sie die Aufgabe in einem dreidimensionalen Spiel die Qualitätsänderung eines Audiosignals zu bewerten. Nachdem Sie das Spiel mit der "Start" Taste auf der vor Ihnen stehenden Faderbox gestartet haben, beginnt das Abspielen des Audiosignals.

Innerhalb des Spiels wird die Qualität des Audiosignals einmalig verändert. Im Anschluss an einen Spieldurchlauf beurteilen Sie, wie Sie die qualitative Änderung wahrgenommen haben. Dazu verwenden Sie den Schieberegler auf der Faderbox. Nach der Positionierung des Schiebereglers betätigen Sie die Taste "Bewertung". Es folgt ein grauer Bildschirm und automatisch das nächste Spiel.

Testanleitung - Anchoring

Vorab werden Ihnen nun die im späteren Test angebotenen Extremwerte der Qualitätsstufen (die Ober- und Untergrenze der Bewertungsskala) präsentiert. Dafür befinden sich auf der Faderbox zwei Tasten, die entsprechend der Skalenstufen "nicht wahrnehmbar" und "sehr störend" beschriftet sind. Nach dem Öffnen der Szene können Sie durch Betätigen der Taste "nicht wahrnehmbar" das Referenz-Audiosignal von einer Minute Länge starten. Während der Wiedergabe kann über die Taste "sehr störend" zur niedrigen Qualitätsstufe umgeschaltet werden. Ein Zurückschalten zur Referenz ist durch wiederholtes Drücken der Taste "nicht wahrnehmbar" möglich.

Testanleitung - Passiv

In diesem Versuchsteil besteht ihr Aufgabe darin, dem Spiel passiv zu folgen. Es werden Ihnen 16 Spielszenen mit einer Länge von 30 Sekunden präsentiert. Konzentrieren Sie sich im Laufe des Spiels auf das wiedergegebene Audiosignal. Am Ende des Spiels bewerten Sie bitte, wie Sie die qualitativen Änderung des Audiosignals wahrgenommen haben. Zunächst werden Sie in einem Training das Spiel einmal durchlaufen.

Testanleitung - Aktiv

In diesem Versuchsteil besteht Ihre Aufgabe darin, in einem Spiel ausgewählte Gegenstände einzusammeln und anderen Gegenständen auszuweichen. Es werden Ihnen 16 Spielszenen mit einer Länge von 30 Sekunden präsentiert.

Konzentrieren Sie sich im Spiel darauf, eine möglichst hohe Punktzahl zu erreichen, indem Sie die Donuts einsammeln und den Schneebällen ausweichen. Um im Spiel nach links und rechts zu navigieren nutzen Sie bitte die Pfeiltasten der Tastatur. Für jeden eingesammelten Donut erhalten Sie einen Punkt. Bei der Kollision mit einem Schneeball wird ein Punkt abgezogen. Die erreichte Punktzahl wird Ihnen im Display der Faderbox und in der Szene angezeigt. Am Ende des Spiels bewerten Sie bitte, wie Sie die qualitative Änderung des Audiosignals wahrgenommen haben. Zunächst werden Sie in einem Training das Spiel einmal durchlaufen.