ISBN 978-82-326-1476-9 (printed ver.)
ISBN 978-82-326-1477-6 (electronic ver.)
ISSN 1503-8181

Doctoral theses at NTNU, 2016:69

Adam D. Borowiak

QUALITY EVALUATION OF LONG DURATION AUDIOVISUAL CONTENT

Adam D. Borowiak

Doctoral thesis

NTNU
Norges teknisk-naturvitenskapelige universitet
Thesis for the Degree of
Philosophiae Doctor
Faculty of Information Technology,
Mathematics and Electrical Engineering
Department of Electronics and
Telecommunications

Doctoral theses at NTNU, 2016:69

NTNU

NTNU
Norwegian University of
Science and Technology

NTNU
Norwegian University of
Science and Technology

Adam D. Borowiak

QUALITY EVALUATION OF LONG DURATION AUDIOVISUAL CONTENT

Thesis for the Degree of Philosophiae Doctor

Trondheim, April 2016

Norwegian University of Science and Technology
Faculty of Information Technology,
Mathematics and Electrical Engineering
Department of Electronics and Telecommunications

NTNU
Norwegian University of
Science and Technology

NTNU
Norwegian University of Science and Technology

Thesis for the Degree of Philosophiae Doctor

Faculty of Information Technology, Mathematics and Electrical Engineering
Department of Electronics and Telecommunications

ISBN 978-82-326-1476-9 (printed ver.)
ISBN 978-82-326-1477-6 (electronic ver.)
ISSN 1503-8181

Doctoral theses at NTNU, 2016:69

Printed by NTNU Grafisk senter

Preface

This thesis is submitted in partial fulfillment of the requirements for the degree of Philosophiae Doctor (PhD) at the Norwegian University of Science and Technology (NTNU). This study was performed within the PERCEVAL project (Perceptual and Cognitive Quality Evaluation Techniques for Audiovisual Systems), funded by The Research Council of Norway under project number 193034/S10. From January 2010 to December 2012 the work was carried out at the Centre for Quantifiable Quality of Service in Communication Systems (Q2S), which was a Norwegian Center of Excellence, appointed by the Research Council of Norway and funded by the Research Council, NTNU and UNINETT. The last part of the work (from January 2013 to November 2015) has been done at Department of Electronics and Telecommunications. Professor U. Peter Svensson has been the main supervisor of this work, and Dr. Ulrich Reiter has been the co-supervisor.

Acknowledgements

With the deepest gratitude I wish to thank my two thesis advisors, prof. U. Peter Svensson and dr. Ulrich Reiter. Their support, encouragement and insightful discussions have been invaluable during my work. Without their patience, feedback and wisdom, this work would not be possible. I would also like to acknowledge and gratitude to all the former colleagues who worked at Q2S for creating a very enjoyable working environment. I would like to thank my parents for allowing me to realize my own potential and for being very supportive during my entire education process. Finally, I'd like to thank my wonderful wife Anna and my two sons Bartek and Filip for all their love, understanding and patience.

Abstract

Subjective quality assessment of multi-modal services depends on a number of external factors that affect the final judgment, e.g. user expectations, user fatigue, room environment or methodology used in the evaluation process. In order to obtain as accurate as possible measurement of the perceived quality, an experimenter should carefully consider all factors contributing to the overall experience. Particulary important is to choose the right measurement method for the purpose of a specific task. In spite of the fact that a number of standardized test procedures for the quality assessment exist it is not always possible to find the one which suits a certain research purpose. In such case, the development of new assessment techniques is usually necessary, which then needs to be followed by appropriate testing and validation procedures. The lack of an appropriate methodology for instantaneous measurement of user's audio-visual quality expectations/preferences over extended periods of time, as well as the scarce attention devoted to the topic of temporal development of Quality of Experience, were the main driving factors for this work.

This dissertation, composed of five papers, helps to understand the underlying attributes of perceived quality and user cognitive processes used in evaluation of long duration audiovisual content. The work described here is twofold: firstly, a novel methodology for continuous quality evaluation is proposed, and secondly, using the method, the effect of the time dimension on user's behavioral reaction to the experienced quality is investigated. The momentary-based approach described in this work reflects instantaneous measures of users' quality judgements. Such measures allow capturing time varying changes of system characteristics and help to contribute to the holistic vision of quality of experience. The results obtained from several experiments described in this work reveal the importance of content duration in the process of quality assessment and its impact on user's quality requirements. The knowledge gained from those studies can be directly applied to the quality assurance models of multimedia content providers and may serve as a valuable source of information for objective quality metrics development.

Contents

List of Figures

List of Tables

Glossary

acceptable quality level	The highest distortion level of a sample that can still be considered satisfactory.
acceptance	A person's assent to the reality of a situation, recognizing a process or condition without attempting to change it, protest.
adjustment device	A device used to adjust the quality of a presented stimulus.
anchor	In the context of this work an 'anchor' means a stimulus with a reference quality.
artifacts	Errors in the decompressed signal that may result when compressing of a digital signal, e.g. when a high compression ratio is used.
attention	The ability to concentrate on a task/stimuli at a given time.
automatic degradation procedure	A process of gradual quality decrease introduced by a system.
bit-rate	The rate at which the bit stream is delivered from the channel to the input of a decoder.
device sensitivity	A characteristic of a device which determines the minimum usable input or the least input which produces an output which satisfies certain specified requirements.
expectations	Feelings or beliefs based on user's previous experiences about how successful, good, etc. a product or service will be.
frame-rate	The number of unique frames (i.e. total frames - repeated frames) per second.
impairment discrimination	The ability to see the difference between two stimuli with different levels of impairments.
involvement	An unobservable state of motivation, arousal or interest toward a recreational activity or associated service.

just noticeable difference	The detection threshold. A value of the smallest perceptible change in the physical intensity of a stimulus.
mean opinion score	The average subjective quality judgment assigned by a panel of viewers (or listeners) to a processed video (or audio).
modality	A particular form of the sensory perception.
perceived quality	The user's perception of the overall quality or superiority of a product or service with respect to its intended purpose, relative to alternatives.
preferences	The evaluative judgment in the sense of liking or disliking an object or service which can be notably modified by decision-making processes, such as choices.
quality adjustment	A correction or modification of a perceived quality to reflect actual conditions or preferences. The process of selection of a quality level fulfilling internal user's preferences, by means of discrimination between the neighboring quality levels.
quality evaluation	The process of determination of quantitative or qualitative value of a product or service. User's opinion of a product/service's ability to fulfill his or her expectations.
quality judgment	The process of forming an opinion or evaluation of perceived stimulus characteristics based on comparison with an internal reference. It's an active process which encompasses different levels of human information processing and which might combine information from various modalities.
quality level	An auditory/visual stimulus quality defined by specific characteristics (e.g. bit-rate).
quality requirements	Characteristics that determine whether a service meets user's expectations.
quantization parameter	A variable used by the decoding process for scaling of transform coefficient levels.
response time	A time period between the time when a sending of response request is triggered and the time when its response is received by the response confirmation role object.

satisfaction	Fulfillment of one's wishes, expectations or needs.
semantic designator	The term pertaining to the relationships between a symbol and what it represents; a type of verbal equivalent of a given symbol.
semantic structure	A meaningful structure with an explicit beginning and end.
sensation	A mental process (as seeing, hearing, or smelling) resulting from the immediate external stimulation of a sense organ, often as distinguished from a conscious awareness of the sensory process.
subject	A test person evaluating the stimuli in a listening or viewing test.
time slot	A time section, time interval between two neighboring events (e.g. succeeding quality degradations procedures).
unweighted sound level	A linear (unweighted) sound pressure level. Sometimes written as dBL or dB(lin).
variable bit-rate	A type of encoding algorithm which can dynamically switch encoding bit-rate based on the properties of the signal.

List of Papers

Publications Included in the Thesis

The following papers are included in part II of this thesis.

- PAPER A:
  Borowiak, Adam; Reiter, Ulrich; Svensson, U. Peter: Quality Evaluation of Long Duration Audiovisual Content. Proc. of *The 9th Annual IEEE Consumer Communications and Networking Conference (CCNC). Special Session on Quality of Experience (QoE) for Multimedia Communications*, pp. 353–357, Las Vegas, 2012.

- PAPER B:
  Borowiak, Adam; Reiter, Ulrich: Quality Evaluation of Long Duration AV Content — An Extended Analysis Using a Novel Assessment Methodology. *Multimedia Tools and Applications Journal*, volume 74(2), pp. 367-380, 2015.

- PAPER C:
  Borowiak, Adam; Reiter, Ulrich; Svensson, U. Peter: Audio Quality Requirements and Comparison of Multimodal vs. Unimodal Perception of Impairments for Long Duration Content. *Journal of Signal Processing Systems*, volume 74(1), pp. 79-89, 2013.

- PAPER D:
  Borowiak, Adam; Reiter, Ulrich: Long Duration Audiovisual Content:Impact of Content Type and Impairment Appearance on User Quality Expectations Over Time. Proc. of *The 5th International Workshop on Quality of Multimedia Experience (QoMEX)*, pp. 200–205, Klagenfurt, 2013.

- PAPER E:
  Borowiak, Adam; Reiter, Ulrich; Svensson, U. Peter: Momentary Quality of Experience: Users' Audio Quality Preferences Measured Under Different Presentation Conditions. *Journal of The Audio Engineering Society*, volume 62(4), pp. 235-243, 2014.

Other publications by the author

- Borowiak, Adam; Reiter, Ulrich; Svensson, U. Peter: Evaluation of audio quality requirements over extended periods of time using long duration audiovisual content. *Advances in Multimedia Information Processing. PCM, Lecture Notes in Computer Science*, volume 7674, pp. 10–20, 2012.

- Borowiak, Adam; Reiter, Ulrich; Tomic, Oliver: Measuring the quality of long duration AV content. Analysis of test subject/time interval dependencies. *EuroITV – Adjunct Proceedings*, pp. 266–269, 2012.

- Weiss, Benjamin; Guse, Dennis; Möller, Sebastian; Raake, Alexander; Borowiak, Adam; Reiter, Ulrich: Temporal Development of Quality of Experience. *Quality of Experience: Advanced Concepts, Applications and Methods*. Eds. S. Möller and A. Raake. Springer, 2014.

Part I

THESIS INTRODUCTION

1. General Introduction

Today's multi-media services are mostly digital in their nature and are delivered to the end-user through different transmission schemes (e.g. internet, traditional broadcast or mobile networks). Together with the technological progress and according to users expectations, the technical quality of produced multi-media content successively increases. However, before the produced content reaches the end-user, its quality is usually affected by many factors (e.g. encoder settings, temporal variability of transmission characteristics, technical constraints of devices used to reproduce the content, etc.). In order to provide the required level of consumer satisfaction in an efficient manner it is desirable to have objective measures which are able to accurately predict the quality of presented material. Unfortunately, the existing objective metrics neither fully reflect the human perception system nor provide sufficient information on how humans interpret and quantify the quality. Consequently, the results obtained using such measures are not precise. Therefore, evaluations with human assessors are still considered the most accurate and trustworthy measurement techniques [145].

The subjective quality opinions incorporate much more than just the assessment of technical parameters of presented stimulus. There are other factors like e.g. assessors' previous experiences, preferences, involvement in the task etc., which might affect the final judgement [123]. It's difficult to include such cognitive mechanisms in the numerical model and hence obtain an ideal correlation with results of subjective tests. However, in case of mono-modal stimulus, some of the existing quality metrics perform quite well (e.g. EET_ PEAQ for audio [145], NTIA VQM for video [143]) but only for specific applications and content types. A bigger problem appears when multi-modal material is considered. This seems to be directly related to the way humans process multi-modal stimuli. Namely, different kinds of sensory information are processed by different areas of the human brain. The brain's modular structure allows for mapping some aspect of external stimuli across many parts of the brain's surface. The different modules influence each other and are mutually dependent [38, 25]. The process is complex and thus difficult to mimic by numerical models. Therefore, the use of subjective quality evaluation methods is invaluable as they provide the most realistic measures of perceived/experienced quality and help elucidate the relationship between different modalities. Moreover, results obtained during subjective studies are cornerstones for the development of objective quality metrics and their validations.

Currently, many standardized methodologies for quality assessment of audio [60, 53, 54, 55, 58, 63] and video [64, 59, 70, 74, 61, 62] are available for experimenters. However, very few methods exist for assessing the complex effects of human audiovisual quality perception [57, 71, 73, 72].

Moreover, those techniques are not adequate for long-term evaluation of user sensations in real-life application scenarios, in which quality may fluctuate over time, and where cognitive and perceptual aspects are of paramount importance. Also, low-attention studies in which focus is on the presented material rather than on the assessment task are not covered by those techniques.

A lack of appropriate methodologies, limited literature available and little interest devoted to the aforementioned problems became the main driving factors for this research work.

This thesis tackles the problem of quality assessment of long duration audiovisual content from a different perspective compared to existing studies. The test methodology developed for the purpose of this dissertation is based on a quality adjustment approach. In contrast to traditional quality assessment methods, in which stimulus quality is judged using numerical/descriptive scales, the approach presented here allows assessors to actively control the quality according to their own requirements in case of perceived quality degradation. The proposed method has been employed to investigate the effect of the time dimension on users' audiovisual quality expectations/preferences. The detailed description of the proposed methodology (Paper A) as well as results of several subjective studies performed with usage of the method (Papers B-E) are described in the second part of this dissertation.

The thesis introduction is structured as follows. Section 2 focuses on long-duration audiovisual content and its coding and transmission schemes. Section 3 provides information about the Quality of Experience concept and factors influencing audiovisual quality perception. A state-of-the-art survey on existing test-approaches for assessment of audiovisual long duration content is presented in section 4. An overview of the thesis, including research questions, summaries of the articles included in the thesis and conclusions can be found in section 5.

2. Long Duration Audiovisual Media

2.1. Scope and Application

Audiovisual (AV) stimuli, perceived by both hearing and vision, are a foundation of today's multimedia services. Apart from purely entertaining, advertising and information purposes, AV content has become a powerful tool and means of communication for a majority of our population. Moreover, the audiovisual material plays a more and more important role in many public sectors like e.g. education and health. Application areas of AV material range from gaming, broadcasting of standard-/high-definition TV content and videoconferencing over mobile TV up to very high quality applications such as digital cinema/ large screen digital imagery or professional, high resolution, digital video recording [134].

The technological progress and success of digital media standards, such as those developed by the International Telecommunication Union (ITU) and the International Organization for Standardization (ISO) in cooperation with the International Electrotechnical Commission (IEC), has lead to that the AV content has become easy to produce, process and distribute. As a result, a huge amount of AV material is currently available in digital form, including professionally produced content (e.g. TV programmes, movies), as well as personal videos shared online. This growing quantity of AV resources is being driven by a long list of content providers, including: cable/internet television, news agencies, movie studios, teaching institutions, digitally focused professional and prosumer producers, amateurs (user generated content), etc. All those content providers generate and provide the audiovisual material via different communication channels.

By using various distribution channels, e.g. internet, digital television, mobile networks, etc., users have increased access to those resources. The widespread availability/popularity of audiovisual content became a driving factor for many studies in which researchers have tried to e.g. categorize multi-modal material [2, 122, 124, 136] or measure its quality [142, 81, 130], Paper D.

The AV content can be classified with respect to different attributes, e.g. modality, motion/colour intensity as well as based on its duration [95, 100, 52]. Most of the attributes have been thoroughly investigated and well defined. One of the exceptions, however, is the content's duration and its effect on user experience.

The terms *long duration/long-form* and *short duration/short-form* are typically used to categorize AV content with respect to its length. However, these terms are not consistently defined by any of the standards developing organizations. One of the available definitions proposed by the Interactive Advertising Bureau (IAB) says: *"Whether professionally produced or user generated, long-form video content*

always has a content arc with a beginning, middle, and end which in its entirety typically lasts longer than 10 minutes" [24]. However, it should be mentioned here that this definition was proposed in their 2009 white paper and since then a lot has changed. These days, the length of time that constitutes a long duration AV content may be anywhere between 10 min and an hour, possibly even longer (i.e. programmes/films rather than short video clips). On YouTube, for example, long duration videos are defined as those with a length greater than 20 min [1].

In terms of quantity, today, shorter-form AV material still represents the majority of the content available. This is mainly due to the large number of short internet videos like seen on e.g. YouTube. However, at the same time, users become more comfortable watching feature-length movies and TV content on mobile devices like laptops, tablets and mobile phones. Such change in users' behaviour, together with the digital and internet driven evolution of video and TV services, cause that long duration AV material successively reaches more and more people.

According to [111], long-form content dominates all device viewing. Together with the rise in accessibility and quality of produced AV content, the overall usage (time spent watching, frequency) has also increased. This tendency can be clearly seen i.a. in reports from Oolaya [110, 111]. The latter shows that Connected TV viewers spent 80% of their time watching videos 10 min or longer, tablet viewers 68% and even those viewing on their mobiles spent 48% of their time watching long-form video.

All the above emphasize the importance of long duration AV content in our daily life and therefore its attributes and potential impact of its length on Quality of Experience should be further investigated.

2.2. Coding and Transmission Basics

The past decades have brought an enormous evolution in the field of audiovisual coding and transmission techniques. The rapid development and popularization of the Internet has enabled a possibility to transfer multimedia content between any two points in the world in real-time. As a consequence, a huge portion of today's traffic over communication networks is occupied by multimedia applications and services. In spite of a continuous increase of the available network bandwidth, handling raw high definition (HD) or higher resolution audiovisual files in real time is still not possible in a majority of current commercial networks. This is due to insufficient throughput of transmission channels which may be affected by e.g. behavior of multimedia consumers, available processing power of the system components and limitations of underlying analog physical medium [35]. In addition, raw video files require massive storage capacity which may be very costly (e.g. an hour long HD video (1080p) with 12bit color depth, RGB 4:4:4 color space and a frame-rate of 60fps, produces of around 2TB of data to be stored!). In order to meet the network and storage requirements, a raw AV material is usually

[1] Ref. youtube.com; on that website one can use different filters when searching for a specific content type. One of the filters relates to duration.

processed using data coding/compression techniques, which exploit redundancy in spatial, temporal and frequency domain, resulting in much lower bit-rate and data volume. Many compression standards have been developed and standardized so far, mainly by the ITU and ISO/IEC bodies. Among the most known ones are H.261, H.263, MPEG-1, MPEG-2, MPEG-4, H.264/MPEG-4 AVC (Advanced Video Coding), and H.265/MPEG-HEVC (High Efficiency Video Coding). The two latter standards, namely H.264/MPEG-4 AVC and the most recent H.265/MPEG-HEVC were developed by Joint Video Team (JVT)- a partnership of ITU-T VCEG (Video Coding Experts Group) with ISO/IEC MPEG (Moving Picture Experts Group [2]). The scope of this cooperation was to develop a standard capable of providing broadcast video quality at very low bit-rates and which could be applied to a wide variety of applications on a wide variety of networks and systems [76].

All the above-mentioned standards follow the same basic scheme for video coding, which is composed of the following phases: the spatial, the temporal, the transform, the quantization and the entropy coding phase [1].

The spatial phase identifies and leverages similarities between pixels within a single frame which leads to a reduction of spatial redundancy. In the temporal stage similarities between successive video frames are exploited by coding their differences. Consequently, the output parameters of the temporal and spatial stages are further transformed and quantized. The final step is an entropy coding using the Run Length Encoding and the Huffman coding algorithms. Those algorithms remove the statistical redundancy in the data, producing an even more compressed video stream. An example of MPEG4/H264 encoder structure is presented in Fig. 2.1. More technical details of the two most recent MPEG coding standards can be found in [139] and [133].

In case of audio coding, several standards should be recalled here, namely, MPEG-1 (Layer III), MPEG-2 (Part 3 and Part 7) and MPEG-4 (Part 3). The MPEG-1 Layer 3 (MP3) still is one of the most popular standards around in spite of the fact that it was first published over 20 years ago. The MP3 encoding process can be divided into several steps. In the first step, the input audio signal passes through a filter bank that converts the signal from time domain to frequency domain. Simultaneously, it passes through a psychoacoustic model that utilizes the concept of auditory masking to determine which are the sonically important parts of the waveform that is being encoded. The bit allocation block minimizes the quantization noise to inaudible levels. Finally, the bit stream formatting block accumulates all the information and processes it into a coded bitstream. The algorithm for MP3 compression can be seen in Fig. 2.2.

A similar coding scheme is followed in MPEG-2 Part 3 which is backwards compatible with MPEG-1 and which supports higher quality multichannel audio. MPEG-2 Part-7/Advanced Audio Coding (AAC) was designed to be an improvement over the MP3 providing i.a. more sample frequencies, more channels, better perceptual quality at low bit-rates, higher coding efficiency and accuracy, etc. The

[2]The MPEG is a working group that was formed by ISO and IEC to develop standards for audio and video compression and transmission.

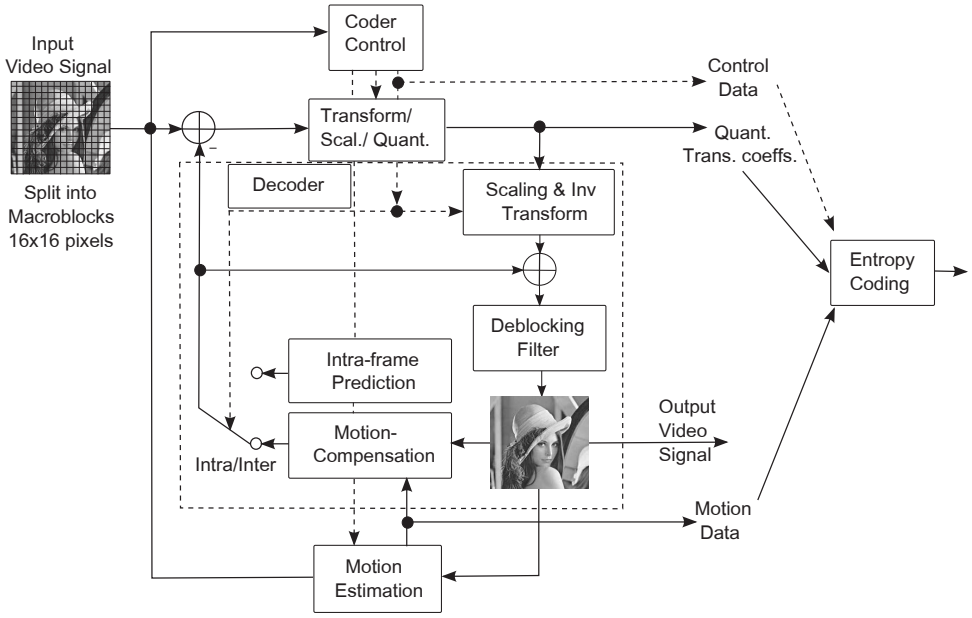


Figure 2.1.: Basic coding structure of H.264/AVC. Adopted from [140].

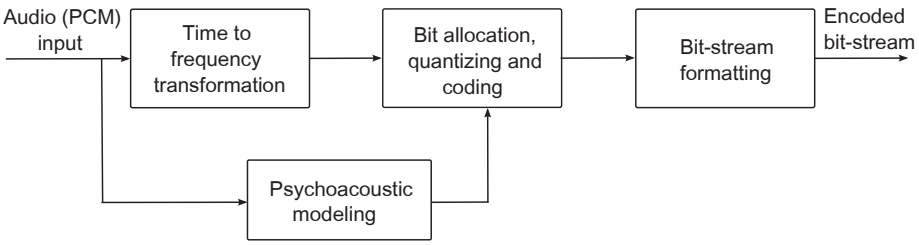


Figure 2.2.: Basic MPEG Audio encoder. Reproduced from [96].

AAC is modular and based on set of modules and profiles. The MPEG-4 adopted the MPEG-2 AAC coder including several extensions to it (a perceptual noise substitution tool, a long-term prediction, Transform-Domain Weighted Interleave Vector Quantization and Bit Sliced Arithmetic Coding) [108].

The data compression (source encoding) of multimedia sources is a first major step of multimedia networking. Once the compression is done, the bitstreams are packetized and sent over the Internet. This enables the receiver to decode and playback the parts of the bit stream that are already received. During this process, the service provider-centric Quality of Service (QoS) [66] issues (like e.g. jitter, packet loss, packet delay) may appear.

Existing standards demonstrate significant improvement in audiovisual compression and transmission capabilities and simplify the process of content distribution. As a result the end-user experience is enhanced through better audiovisual quality at reduced file size, causing long-form AV content growth.

3. Quality of Experience

3.1. Definition and Background

Quality assessment of mono- and especially multi-modal services has received increased interest during the past decades. Researchers from around the world have carried out a number of studies in which audiovisual quality was measured. In the beginning, the focus was rather on technology-centric approaches based on QoS parameters, like e.g. spatial resolution, color depth, number of channels or frame-rate. Only more recently has this trend been directed toward a user-centered approach. The audiovisual quality started to be viewed as an entity consisting of technical and environmental factors as well as a person's experiences, expectations, etc. The Quality of Experience (QoE) concept has arisen as a consequence of this.

In contrast to the QoS approach, QoE focuses on the entire service experience which includes the complete end-to-end system effects (client, terminal, network, services infrastructure, etc.) and which may be influenced by content type, context and user expectations [67]. Moreover, QoE involves peoples' aesthetic and even hedonic needs [98, 120]. A graphical representation of the basic QoE model can be seen in Fig. 3.1.

As the QoE is a very broad term, many different definitions have been proposed to explain it. The ITU defines the QoE as *"the overall acceptability of an application or service as perceived subjectively by the end-user"* [67].

More broadly, Raake et al. [116] have summarized QoE as *"the degree of delight or annoyance of a person whose experiencing involves an application, service, or system. It results from the person's evaluation of the fulfillment of his or her ex-*

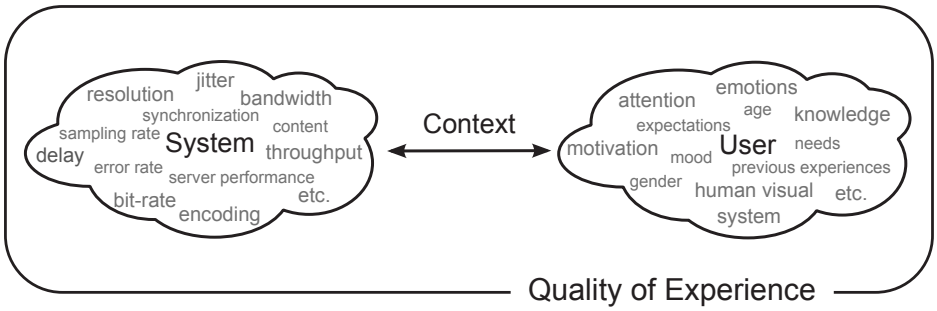


Figure 3.1.: A basic model of Quality of Experience.

pectations and needs with respect to the utility and / or enjoyment in the light of the person's context, personality and current state."

Many researchers consider QoE as an extension of the QoS concept by mapping the end user's perception to QoS parameters. A clear focus on such an approach can be found e.g. in [36, 121, 26]. In those studies authors try to predict QoE from a given set of QoS parameters which then are linked to QoE by means of various mathematical models. However, due to the random component of human behaviour it's difficult to assure a relatively good accuracy of such a concept [120]. As the non-technical aspects, like motivations, attention, mood, etc. (cf. section 3.2) proved to be influential in the process of quality judgments [124], the aforementioned approach may not be sufficient in capturing the QoE. Many studies have been performed under the QoE framework and a number of QoE related publications can be found in the literature [4, 3, 51, 21, 147]. Nevertheless, the QoE-based quality evaluation is still a difficult task. A task which depends on human cognition and perception related processes which, as for now, are hard to quantify and measure [22]. Moreover, it depends on many external factors which may heavily affect the user's experience. A current trend of studies on quality is to optimize and measure factors influencing QoE in various environments/contexts and with as little negative perceptual effects as possible.

3.2. QoE - Influencing Factors of Long Duration AV Media

As mentioned earlier (see previous section), an audiovisual QoE is dependent on a number of factors that relate directly to our perceptual and cognitive processes as well as to technical and environmental aspects. Some of them have physical properties which are easy to identify and measure (e.g. content and system characteristics) while others are user and/or situation dependent (e.g. a viewer's individual interest in the content, or the social context) which make them more difficult to describe or quantify. QoE is a multi-aspect quality, resulting from the synergy of those multiple influencing factors. Most of those factors are interdependent and therefore influence QoE in a complex way.

In [123] Reiter et al. divided the QoE influencing factors into three main categories: system related (e.g. signal and network variables, devices), human related (e.g. age/gender, interest, affect/mood, personality) and context related (e.g. social motivation, physical environment, economic conditions). All the three categories can influence each other in the process of QoE forming and may differ per service.

Following [123], Human Influencing Factors (HIF) can be defined as *"any variant or invariant property or characteristic of a human user. The characteristic can describe the demographic and socio-economic background, the physical and mental constitution, or the user's emotional state"*. HIF highly depend on features of an individual (e.g. experience, expectations, mood, motivation, age/gender, etc.) which in many cases are difficult to extract and measure during QoE tests. Except

for factors that are easy to establish like e.g. age or gender, HIF have often been overlooked in the quality research studies. Only more recently, an effort has been made by researchers to investigate them more systematically bringing more and more interesting insights into the field of QoE [105, 137, 109, 144, 94, 19].

In contrast to HIF, the System Influencing Factors (SIF) are more 'tangible' as they refer to *"properties and characteristics that determine the technically produced quality of an application or service"* [79]. SIF relate directly to the multimedia capturing, configuration, transmission and reproduction chain. Each of those processes determine the presence of specific attributes which have an impact on the final QoE. The multimedia capturing process addresses several influencing factors like e.g. content type, duration, texture or color depth [89]. Influencing factors of media configuration include encoding, frame-rate, resolution, audio-video synchronization, sampling-rate [147]. The data transmission process is determined by factors such as bandwidth, throughput, delay, jitter, error rate, packet losses [104, 36]. The final link in the AV media chain is highly related to specifications of an audio/video reproduction device (e.g. screen resolution, display size, luminance, audio loudness and clarity, etc.) but also relates to provider specification and capabilities [123].

The last category, Context Influencing Factors (CIF), include all situational factors that can be used to *"describe the user's environment in terms of physical, temporal, social, economic, task, and technical characteristics"* [79]. CIF can be divided into characteristics of the environment (e.g. lighting conditions, background noise level) and the way the service is offered (e.g. purchase and/or account conditions).

An overview of the factors that may have effect on the QoE in the multi-media context can be found in Table 3.1. All the aforementioned influencing factors of AV media can have a great impact on quality perception and therefore require special attention in audiovisual quality studies.

Table 3.1.: Influencing Factors of AV Media.

Human Influencing Factors (HIF)	System Influencing Factors (SIF)	Context Influencing Factors (CIF)
Factors at low-level cognitive processing: gender, the user's visual and auditory acuity, age, personality traits, lower-order emotions, user's mood, attention level, motivation, etc. **Factors at high-level cognitive processing**: knowledge, education background, attitudes and values, emotions, expectations, needs, previous experiences, socio-economic situation, etc.	**Content related**: color depth, temporal and spatial requirements, 2D/3D, texture, etc. **Media related**: sampling rate, encoding, resolution, synchronization, etc. **Network related**: delay, bandwidth, throughput, jitter, error rate, etc. **Device related**: personalization, interoperability, server performance and availability, security, privacy, etc.	**Physical context**: characteristics of location and space **Temporal context**: frequency of use (of the system/service), time of day, duration, etc. **Economic context**: brand of service/system, cost, subscription plan, etc. **Technical and information context**: availability of other networks than the one currently used, existing interconnectivity of devices over NFC or Bluetooth, availability of an app instead of the currently used browser-based solution of a service, etc. **Task context**: multitasking situation **Social context**: other people present or even involved in the task

4. Audiovisual Quality Perception and Assessment

4.1. Multi-Modal Quality Perception

Generic relationships between QoE and its influencing factors provide fundamental insights into the nature of user quality perception. Perception is the process of recognizing, organizing and interpreting sensory stimuli [103]. It is determined by an interaction between bottom-up processing (also called sensory-based processing), which starts with stimulation of the receptors, and top-down processing (also called knowledge-based processing), which brings the observer's knowledge into play [106, 38]. In the first step of the perception process, the low-level sensory information is transformed to higher-level information (e.g. shapes are extracted for object recognition). Once incoming information has been initially coded by the sensory systems, a person's concept, expectations (knowledge), and selective mechanisms (attention) are applied to the information [38].

The incoming multi-modal information is not processed independently [86], therefore the audiovisual quality perceived by humans is not a simple sum of individual perceptual channels [44]. Moreover, information which emerges from two or more sensory modalities cannot be obtained from each of the modalities separately [86]. An integration of different sensory inputs into one unified experience (multisensory integration) contributes to richer experiences [33, 50], enhances human's cognitive abilities [127] and assists in temporal and spatial judgements [16]. For example, Santangelo et al. [128] proved that the human attention is more rapidly and more precisely directed towards the source when the source stimulates two or more modalities. Nevertheless, the multisensory integration may also lead to wrong conclusions in case mismatched auditory and visual stimuli are merged together. A classical example of such an observation is the McGurk effect [102]. In order to better understand the multisensory integration process, the interrelationship between sensory modalities needs to be thoroughly investigated.

As yet, the mutual influence between the auditory and visual stimuli has been confirmed in many subjective experiments. An overview of various types of interaction between those two modalities can be found e.g. in [49, 131, 146, 83, 112, 15]. The quality judgement of one modality is often affected by the presence of an other modality in case a combined audiovisual stimulus is introduced to a subject [11, 87]. Most of the available studies indicate that video quality has more influence on perceived audio quality than vice versa as witnessed in [11, 77]. However, the reverse situation can also be found in the literature. Hands et al. [44] showed that in

case of "talking head" type of content audio quality is more important than video quality, as audio conveys most of the information in such a scenario. Such content dependency has also been reported in [82]. It seems that one modality appears to be more important than the other when different content [44], context [80, 14] and task [141] are considered.

In addition to the above findings, many studies show that the relative timing of audio and video can also affect the quality perception. A number of investigations related to audio-video synchronization requirements (e.g. lip sync timing) have been performed so far [135, 12, 30, 31]. Results of these studies show that the proper temporal synchronization of sound and vision is crucial in the process of quality perception. According to ITU-R BT.1359 the threshold of acceptability for audio leading video is about 90 ms and in reverse situation about 185 ms on average. Improper synchronization of audiovisual stimuli can distract the viewer from presented material and may reduce the clarity of the intended message [118]. Furthermore, in [28] authors claim that with increased asynchrony the perceived quality degrades rapidly.

All the mentioned studies indicate that the perceived quality of audio and video stimuli in a multi-modal presentation should not be treated as separate events nor evaluated as such. Therefore, it's especially important to better understand the underlying mechanisms responsible for merging of stimuli perceived by different senses as well as each of the steps in the complex process of human perception. A special attention may also need to be directed towards the development of novel subjective assessment techniques and objective quality metrics able to integrate factors influencing the quality perception and QoE in general.

4.2. State of the Art - Subjective and Objective Quality Metrics for Audiovisual Material

The necessity to consider all the involved modalities described above is reflected in studies of Quality of Experience. As mentioned earlier, Quality of Experience (QoE) focuses on the entire service experience which can be influenced by human and contextual factors (ref. section 3.2). Such factors have a very subjective nature and therefore, the QoE seems to be extremely difficult to measure without involvement of human assessors. However, subjective testing usually requires special assessment facilities to produce reliable and reproducible test results. Moreover, the assessment process is time-consuming and expensive. Therefore, there is a need for computation based (objective) methods able to assess the QoE with reasonable accuracy. Nevertheless, those methods need to rely on results of perception-based (subjective) studies. The two measure categories are presented in the following sections with the focus on audiovisual material.

4.2.1. Objective Metrics for Audiovisual Material

Objective quality metrics can be classified into three main categories according to the availability of the undistorted, reference signal: full-reference (FR), reduced-reference (RR), and no-reference (NR) [125, 4]. FR metrics compare a reference signal against a distorted one in order to compute the quality difference between the two. FR algorithms are usually the most accurate and relatively straightforward which contributes to their widespread use. However, in many real-life applications (e.g. videoconferencing, IPTV, etc.) FR models cannot be used as the reference signal is simply not available for comparison. In such cases, NR metrics are usually employed. The NR methods are an absolute measurement of characteristics and features in an impaired signal and are often focused on a specific degradation type (e.g. blurring, blockiness) and the analysis of coding parameter settings. Due to the lack of a reference signal, they may be less accurate than other approaches, but are more efficient to compute. RR algorithms, instead of a full reference, use quality features extracted from the reference and distorted signals. Those features are then compared in order to generate a single quality score. The RR models are usually adopted in cases the full referenced signal cannot be used (e.g. in a transmission with a limited bandwidth).

The aforementioned computational models automatically predict the perceptual quality using mathematical operations. Such operations are often made with usage of a model of the human visual system (HVS) and the auditory system.

Most of the existing objective quality metrics focus only on a single modality, either audio or video, and do not consider the strong mutual influence of the two in the process of quality evaluation. In case of multi-modal stimulus, our brain uses multiple sources of sensory information derived from several different modalities. The multi-modal perception is not a simple linear combination of single modalities' perceptions, as discussed in Section 4.1. Most of the research indicates that at some point in the perceptual processing, all these different sources of information integrate to form a coherent and robust percept (perceptual fusion) [34]. During this process, one modality can modify and complement the perception derived from another modality [83, 129, 13]. In case of multi-modal quality assessment such a cross-modal effect of multisensory integration may heavily influence (positively or negatively) the QoE.

So far, several fusion approaches have been proposed for the purpose of audiovisual quality measurements. Most of them rely on results of subjective experiments in which the relationship between audio quality, video quality and audiovisual quality was studied [11, 85]. The basic model which can be found in the literature implies equal importance of audio and video quality in the overall audiovisual quality:

$$AVQ = \alpha + \beta(AQ \times VQ) \tag{4.1}$$

where AQ is audio MOS, VQ is video MOS, and AVQ is audiovisual MOS. Bellcore (Bell Communications Research) came to such conclusion in their early 90's studies [6, 7, 8]. Later research by Institute of Telecommunication Sciences (ITS) and Pinson et al. [113] have confirmed that a multiplicative model (see Eq. 4.1)

fairly accurately predicts audiovisual quality; but only if the audio quality spans the same range as the video quality.

Kitawaki et al. [85], however, emphasize the importance of the mutual influence of audio and video information in the process of construction of audiovisual quality model. This has been supported in other studies [44, 101] in which models based on a linear fusion of audio and video quality (with fusion parameters chosen empirically) were tested. An example of such a fusion model is

$$AVQ = \alpha_0 + \alpha_1 AQ + \alpha_2 VQ + \alpha_3 (AQ \times VQ) \tag{4.2}$$

The parameter α_0 improves the fit in terms of residual between the perceived and predicted quality and is not relevant to the correlation between the two. The fusion parameters $\alpha_0, \alpha_1, \alpha_2, \alpha_3$ may vary depending on experimental scenario and are usually optimized over one data set. Therefore, they are not applicable in other cases. Garcia et al. [37] indicated that this model cannot fully capture the impact of impairments (mainly audio related) on integral audiovisual quality. Moreover, the models described by Eq. 4.1 and Eq. 4.2 do not include other factors that may heavily influence the final audiovisual quality, like e.g. synchronicity.

In order to improve the aforementioned models Heyashi et al. [46] proposed an approach in which the audio-video synchronization has been included in the fusion model.

$$MMQ = \beta_0 + \beta_1 AVQ + \beta_2 DQ + \beta_3 (AVQ \times DQ) \tag{4.3}$$

In this equation the MMQ stands for multimedia quality, AVQ for audiovisual quality and DQ for quality degradation which is calculated based on audiovisual delay.

Other fusion models can be found in the literature [101, 43, 97, 10], however, most follow the aforementioned approaches not exhibiting a significant breakthrough in this field [145]. It may be due to the very subjective factors which cannot be easily integrated in such numerical models, like e.g. semantic importance of audiovisual material, assessors' previous experiences, attention of subjects or usage context. All this makes the development process of audiovisual quality metrics quite complicated.

In order to help with the development and implementation of the quality metrics/models for multimedia applications/services, the ITU has issued two recommendations: P.931 [75] and J.148 [68]. The first one specifies the parameters and measurement methods to assess relative synchronization between media channels, delay and frame-rate. The latter details the requirements for the development of an objective audiovisual perceptual quality model. The basic components of such a model can be seen in Fig. 4.1.

Although some of the requirements of the multimedia perceptual model have been described in those recommendations there still is a long way to obtain a reliable metric which is able to measure the audio-visual quality automatically. It is mainly due to many unknowns related to human perceptual processes. Therefore, further research needs to be done to better understand how humans perceive the combined

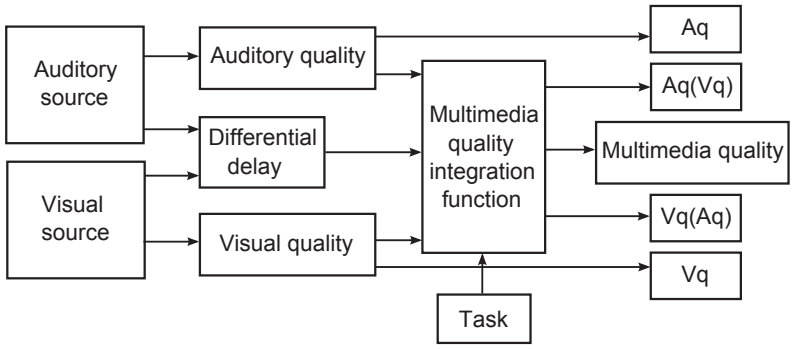


Figure 4.1.: Basic components of a multimedia model. Aq: Objective measurement of audio quality, Vq: Objective measurement of video quality, Aq(Vq): Objective measurement of audio quality, accounting for the influence of video quality, Vq(Aq): Objective measurement of video quality, accounting for the influence of audio quality. Adopted from ITU-R J.148.

audio-visual stimuli and at what stage the fusion process, which forms a single overall quality, appears.

So far, the existing objective metrics have been found to be sufficient only for some of the encoding parameters and for predetermined character of sequences. Therefore, in all the cases in which the reliability and precision of the obtained results count, e.g. for evaluation of new encoding algorithms, or selection of compression parameters, subjective methods are applied.

4.2.2. Subjective Methods for Audiovisual Material

Many ideas for conducting subjective quality studies for audio and video material exist. As results obtained during such tests need to be comparable between different laboratories, the International Telecommunications Union (ITU) has issued a number of recommendations regarding the implementation of tests with human assessors. However, the ITU provides only very few recommendations which are relevant to quality assessment of audiovisual material. Some of them were designed for evaluation of one modality in the presence of an accompanying signal from other modality while others were issued directly for the purpose of audiovisual material. Among the first group are ITU-R BT.500 [64] and ITU-T P.910 [70]. The ITU-R BT.500 has its roots in broadcasting and is dedicated to evaluation of television picture quality in the audiovisual context. The recommendation provides information on important aspects related to experimental design, such as standard viewing conditions, general test methods, rating scales, criteria for selection of observers, etc. The ITU-T P.910 also provides suggestions for the experimental design, however, it is intended for evaluation of one-way overall video quality for multimedia applications such as e.g. telemedical applications, videoconferencing, storage and retrieval

applications. Other relevant recommendations are: ITU-R BS.775-3 [60] (*Multichannel stereophonic sound system with and without accompanying picture*) and ITU-R BS.1286 [56] (*Methods for subjective assessment of audio systems with accompanying picture*). The second group consists of ITU-T recommendations P.911 [71] and P.920 [73]. The ITU-T P.911 is similar to ITU-T P.910 but applies to audiovisual (instead of visual only) subjective assessment in a non-interactive context whereas ITU-T P.920 is intended to define interactive evaluation methods. Another relevant recommendation is ITU-R BT.1359 [57] which deals with relative timing of sound and vision for broadcasting. All the above recommendations were initially designed (and later updated accordingly) for fixed audio/video services transmitted through a reliable link to a static display device (e.g. a TV set) located in a relatively peaceful/non-distracting environment. However, the recently emerged paradigms of Internet video and distribution quality television do not fit into the aforementioned concept. Therefore, ITU has issued a set of regulations suitable for the new services and summarized them in ITU-T recommendation P.913 [72].

A summary of ITU recommendations related to quality assessment of audiovisual content can be found in Fig. 4.2.

All the above recommendations provide many different test methodologies for quality assessment. Those techniques are often classified into two main categories, namely, Single Stimulus (SS) methods and Double-Stimulus (DS) methods. The Single Stimulus methods are often preferred by researchers over other methodologies, as they are well defined and straightforward to implement [117]. An example of such methods is the Absolute Category Rating (ACR) [70] or Single Stimulus Continuous Quality Evaluation (SSCQE) method [71, 42]. In the ACR method, each sequence is presented to the assessors once only and after the presentation its quality is rated on an ACR scale. The ACR scale is evaluated based on numbers that are assigned to the word descriptors, where ”bad” corresponds to 1 and ”excellent” corresponds to 5. The average numerical score over all participants and for each test condition, forms the so called mean opinion score (MOS). The SSCQE, however, allows participants to continuously rate longer sequences using a slider device. Samples are taken in regular intervals, resulting in a quality curve over time rather than a single quality rating. The presentation structure of ACR and SSCQS method can be seen in Figs. 4.3 and 4.4, respectively.

In the DS methods, the test-sequences are presented in pairs: the reference and the impaired one. The main representatives of this category are Double Stimulus Impairment Scale (DSIS) [71], Double Stimulus Continuous Quality Scale (DSCQS) [64] and PC (Pair Comparison) [70]. In the DSIS method the viewer always sees a reference sequence first, then the same sequence impaired (see Fig. 4.5).

Subjects are asked to assess the quality of the second stimulus in relation to the reference using a so-called impairment scale (from ”impairments are imperceptible” to ”impairments are very annoying”). The DSIS method is well suited for evaluation of transmission errors which result in clearly visible impairments. As opposed to DSIS, in the DSCQS method the reference and the impaired sequence are presented to the assessor twice in an alternating fashion. In the series of trials, the position of the reference is changed randomly. Subjects are asked to assess the overall quality

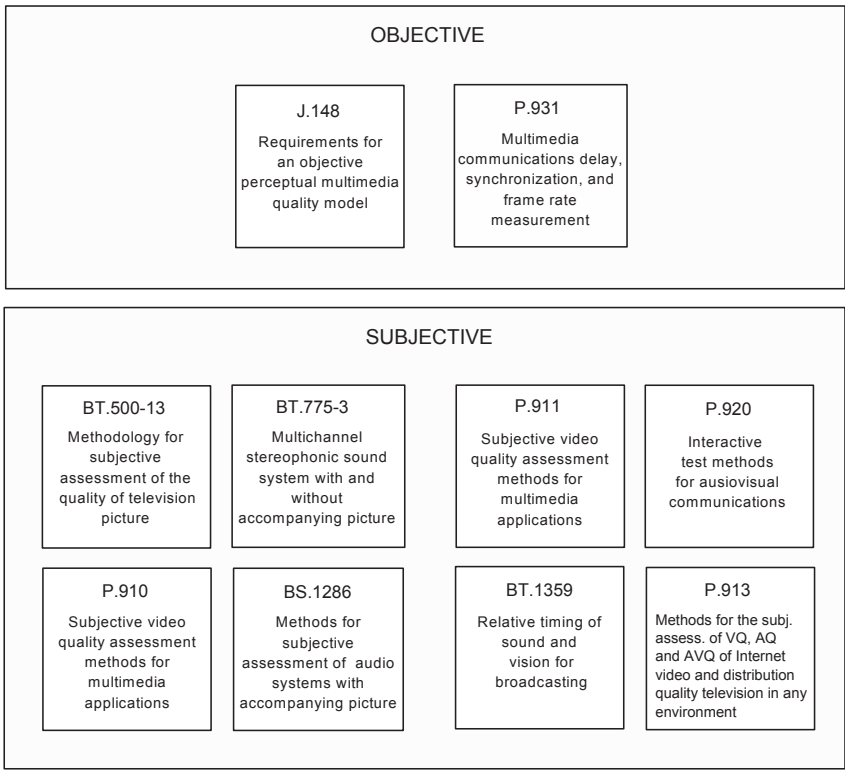


Figure 4.2.: ITU Recommendations for objective and subjective audiovisual quality assessment.

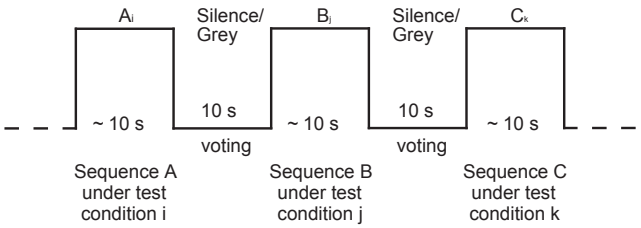


Figure 4.3.: ACR methodology presentation concept. Adopted from ITU-R BT.500-13.

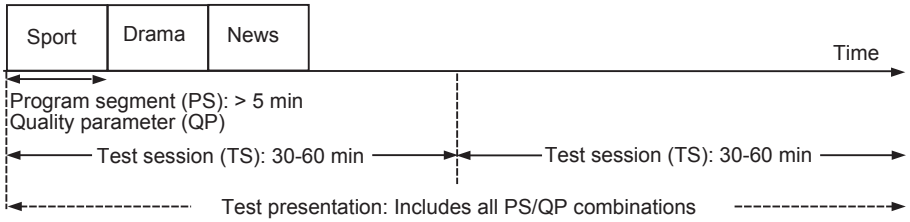
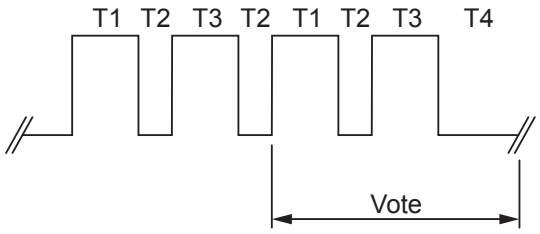


Figure 4.4.: SSCQE methodology presentation concept. Adopted from NTT website.



Phases of presentation:

T1 =	10 s	Test sequence A
T2 =	3 s	Mid-grey produced by a video level of around 200 mV
T3 =	10 s	Test sequence B
T4 =	5-11 s	Mid-grey

Figure 4.5.: DSCQS and DSIS methodology presentation concept. Adopted from ITU-R BT.500-13.

of each of the two sequences using a continuous quality scale divided into five equal lengths, ranging from "bad" to "excellent". The scales are presented in pairs to accommodate the double presentation of each test sequence. Moreover, the scales are continuous to avoid quantizing errors. The analysis of such collected data is based on the difference in rating between each of the pairs which is calculated using an equivalent numerical scale ranging from 0 to 100. The DSCQS is quite sensitive to small differences in quality and therefore is often use when the quality of the impaired and the reference sequence is similar. An alternative method derived from DSCQS is SAMVIQ (Subjective Assessment Methodology for Video Quality) proposed by EBU (European Broadcasting Union) Project Group B/VIM (Video In Multimedia) [29, 92]. In principle, the SAMVIQ has been designed for assessing the perceptual video quality of a multimedia service. In the SAMVIQ method, video sequences are presented in multi-stimulus configuration, so that the assessors can choose the order of tests and correct their votes if necessary. The impaired sequences can be directly compared among themselves and against the reference at any point of time of the test. An assessor has full control over the grading process and can adapt it according to his/her personal wishes (can start/stop the presentation, modify the grades, repeat playout, etc.) [92]. For the PC method the clips from the same scene but different conditions are paired and participants make the preference judgement for each pair. The PC method allows very fine discrimination between the test sequences. The main issue related to the PC method is an exponential growth of the number of pairs with the number of factors and factor levels being investigated. To overcome this limitation Eichhorn et al. [32] proposed a variation of the original method called Randomized Pair Comparison in which a small subset of pairs is randomly selected creating a unique experiment session for each assessor. Such a solution, in contrast to full factorial designs, allows for more realistic assumptions about the time and effort assessors have to spend on a study.

Some of the methods may have fewer context related effects, which are unwanted test biases. The choice of a test-methodology depends upon objectives of the study, resources available and possible other constraints (e.g. time).

4.2.3. Beyond Traditional Subjective Test Methodologies for Audiovisual Content

Apart from the standardized procedures described in the previous sections a number of alternative subjective methods for assessment of audiovisual content exist. A huge part of such methods follow the main principles of the standardized methodologies but entirely unique techniques have also been proposed.

Staelens et al. [130] proposed a method which uses a feature length movie (recorded on a DVD disc) with impairments (blockiness and frame freezes) introduced beforehand. Such a prepared DVD together with a questionnaire in a sealed envelope is given to the assessors. Their task is to watch the movie at home under the same conditions as they usually watch television. Immediately after the movie the assessors answer questions related to perceived quality degradations and provide an overall quality judgement using an ACR scale. Similar procedure is followed

by another group of subjects but in a laboratory setup. Staelens found that the relative impact of impairment types changed with the setting. While watching the movie, the subjects were more tolerant of impairments that did not interrupt the flow of the movie. The Staelens experiments indicate that the traditional subjective testing methods may not accurately predict the quality perceived by end-users.

A different approach has been suggested by Strohmeier [132]. His mixed methods research approach for audiovisual quality, called Open Profiling of Quality (OPQ), creates a link between qualitative and quantitative studies. The methodology consists of three sessions. In the first session an ACR test is performed by participants in which they rate the integral quality of presented stimuli. The second session is devoted to extraction of quality features that participants used to evaluate overall quality in the first session. Features which are not unique or cannot be defined by subjects are excluded from further procedure. As a result, a list of attributes is created and written on an assessment card. To each of the attributes a continuous rating scale is assigned, with a "min" (minimum sensation) and a "max" (maximum sensation) label at the ends of the scale. In the last session, each participant again rates the stimuli using all scales on his/her scoring card. This way new knowledge related to quality perception of a individual can be gained.

A new method for immersive audiovisual subjective testing has been proposed by Pinson et al. [114] and later followed in [126]. The main difference compared to the traditional quality testing is in the way the test conditions (Hypothetical Reference Circuits, HRCs) are applied to source sequences(SRCs). In typical tests following ITU recommendations, a low number of SRCs s is treated with a similar (or larger) number of HRCs h, resulting in $h \times s$ processed sequences (PVSs). A subject task is to rate all PVSs, which leads to repetition of content. In the proposed immersive design, such repetition does not exist. According to the authors, without a repetition of content, subjects will remain immersed in the viewing/listening task instead of focusing on the "technical quality" — which is expected to occur in case of multiple views of the same sequence. Moreover, subjects may get bored of the repetition.

More recently a focus has been directed to usage of EEG and other sensory based devices in the quality assessment studies. Arndt et al. [9] conducted two experiments in which long duration content (40 min and 60 min respectively) was evaluated by participants. In addition electroencephalography (EEG) has been employed to quantify users quality perception. During the presentation of the quality-wise manipulated test sequences, an EEG signal and other physiological measures were recorded. Authors show that in both presented experiments, the QoE reported by test subjects is represented in the EEG data. More specifically, the EEG recordings indicate a change in the cognitive state of the subject during the exposure to low-quality compared to high-quality sequences. Usage of EEG can be helpful for quantifying perceptual quality changes as well as for broadening our knowledge on human perception and quality judgment.

As witnessed in the above studies there is a continuous need for new quality assessment methodologies able to increase our knowledge in the field of quality perception and evaluation.

4.3. State of the Art - Subjective Quality Assessment of Long Duration Content

In order to quantify the quality of user experience many different test-methodologies have been developed, as described in the previous section. However, a majority of those assessment techniques has been designed for very short excerpts of content with duration restricted to 10-15 s and with constant quality throughout the entire stimulus length. Such an approach is not necessarily representative of a typical media consumption situation.

Treating the QoE as a static event contradicts the real-life scenarios in which content of extended duration is used and where quality variations might appear over time. Using such methods, the influence of time-varying and scene-dependent effects of impairments on temporal development of QoE cannot be investigated. By eliminating perceptual, affective and cognitive factors from the process of quality assessment, inappropriate results may be obtained and in consequence - false conclusions derived.

Although this seems to be an important issue, relatively little attention has been devoted to this problem. This lack is directly reflected by the limited number of studies in which long duration audio-visual stimuli have been used as well as by the limited number of standardized methodologies developed specifically for such purpose [1].

Another important issue is the 'long duration' term itself. As mentioned earlier, it's difficult to find a clear guidance on how long should the content be in order to be considered as such. Among the so called 'long duration studies', only very few are designed for stimuli with lengths exceeding several minutes [90, 91] and only more recently have studies with over 30-min long content been performed [130, 18, 17, 9], Paper B. Moreover, in many cases, methods designed for short duration content are employed in studies where e.g. full length movies are used as a test-material. Neglecting differences between those stimulus types may lead to the mentioned contradictions among the specific studies.

In general, three commonly used approaches for subjective quality assessment of long duration content can be identified:

- overall quality judgement provided after the entire stimulus occurrence
- quality assessment over fixed time-window sizes
- momentary quality evaluation of time varying system characteristics

A more detailed description of those approaches is provided in the following.

[1] To be more precise, only one such method exists - Single Stimulus Continuous Quality Evaluation (SSCQE) methodology specified in ITU-R Rec. BT.500 (ref. section 4.2.2)

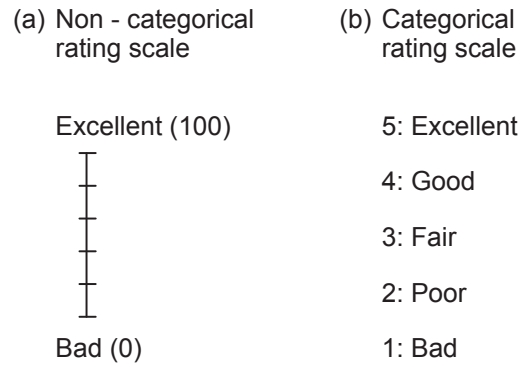


Figure 4.6.: Examples of rating scales used in quality assessment studies. Adopted from ITU-R BT.500-11.

4.3.1. Overall Quality Judgement

A single, retrospective appraisal of the whole experience (often supplemented by different survey techniques such as interviews, focus groups and questionnaires) is the most common approach in audio-visual quality measurement studies [3]. In such case, the assessment procedure takes place directly after the stimulus occurrence and the quality judgement is provided on a pre-defined quality scale (either non-categorical or numerical) as defined in i.a. [70, 71, 64] (see Fig. 4.6).

This approach has been mainly employed in studies with stimulus length up to 2-3 min (e.g. [83, 90, 91]) but also in studies in which full length movies are evaluated [130]. In both cases the same standardized assessment methods (e.g. ACR) [70, 71] are employed for the purpose of votes acquisition (see section 4.2).

The cumulative experience provided after the end of a given service is usually treated as an estimation/average representation of the perceived quality [138]. However, such quality ratings do not necessarily seem appropriate in case of long duration stimuli. This is due to the fact that humans are more likely to make an overall quality judgment based on the first, or the most recent experiences which are assumed to be most influential in the decision-making process [45]. The judgements provided directly after the end of presented material are subject to i.e. *primacy* or *recency effect* [47, 78, 93] and therefore may be different from the QoE at specific time periods of an episode. The *forgiveness effects* of the observers have been investigated in many studies by inserting artifacts at different time positions within sequences of varying lengths and by collecting an overall quality judgement after each episode. The results revealed that the reporting time and the human memory process (beyond 10-15 s time slots) play extremely important roles [5].

All the above suggest that in case of long duration content the cumulative experience may lead to conclusions which are not necessarily reflecting reality. In real-life scenarios, the presented quality is judged at a given point/period of time and based on the impairment intensity, as witnessed in [45] and [48]. This means

that in case of the perceived quality being poor over e.g. 3-5 min there may be a risk that the offered service will be rejected by the viewer before its end-point. An example of such behaviour can be found in a study by De Pessemier et al. [27] in which quality of experience of a commercial voice-over-IP service was evaluated. The authors showed that voice calls that received a low quality rating from the users had a significantly shorter duration in comparison to those of which quality was rated higher.

Of course, such a situation does not happen in a controlled laboratory setup where test subjects are usually expected to watch/listen the entire presented material and provide the average rating right after. As shown in [42] for video quality and in [39] for speech quality, such averaged results are often too optimistic and do not reflect the user's actual quality needs and expectations. Moreover, this way the behavioral responses to momentary quality changes cannot be investigated.

4.3.2. Quality assessment over fixed window sizes

In some quality assessment studies, a longer content is divided into short pieces (e.g. 10 s long time-windows) and then quality evaluation is performed for each segment after its occurrence, or continuously, during the episode presentation. For this purpose, standardized techniques designed for short duration sequences (up to 16 s long) are used (e.g. Absolute Category Rating (ACR) [70], Double Stimulus Continuous Quality Scale (DSCQS) [64], Paired Comparison (PC) [70], etc.) (see section 4.2).

By quality assessment of short time-windows a better approximation of how the QoE develops over time can be obtained. Moreover, this way problems related to the *primacy* and *recency* effects (see above) can be significantly reduced which makes the evaluation procedure more accurate compared to the after-entire-stimulus judgment mentioned previously.

Unfortunately, the *time-window* approach is lengthy in nature and therefore impractical in real life application scenarios. It requires a very long time to perform quality assessment tests, especially when contents of extended duration and hence with large numbers of segments are considered. Moreover, such an approach is very distractive and attention demanding which makes it inadequate for real life applications (e.g. watching television in a living room environment).

Many variations of the *time-windows* based approach can be found in the literature, e.g. [119, 40]. As an example, the study by Gutierrez et al. [40] can be recalled here. The method described in their work requires a quality judgment by the users after impaired segments of fixed duration. The segments are reproduced in a continuous manner (no breaks in between). The quality evaluation is made during the appearance of segments with no degradations in contrast to the gray segments used for this purpose in the ACR method. This way, according to the authors statement, the continuity of the sequence is preserved making the viewing conditions more realistic. This is a sort of quasi-momentary approach and it can be considered a step toward a momentary quality assessment (see section 4.3.3).

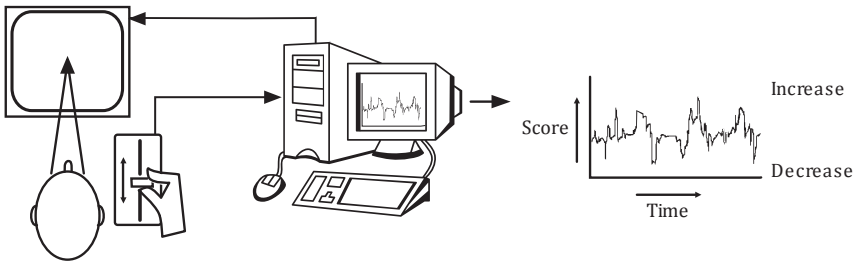


Figure 4.7.: Operational principles of the SSCQE method.

4.3.3. Momentary Quality Evaluation

In order to assess dynamically varying quality, an appropriate methodology needs to be employed (see section 4.2). What is needed is a methodology which allows for instantaneous quality evaluation at the time of impairment appearance and which can be used in cases when reference stimulus is not available (e.g. home viewing scenario). Moreover, it has to be as little invasive as possible in order to direct user attention to the presented material rather than the evaluation task. This way, not only the basic quality of audiovisual material can be assessed but also the fidelity of the information transmitted. As a result, a much more realistic approximation of perceived quality could be obtained in comparison to the after-stimulus appearance judgement. It's a challenging task, however, to develop a method which incorporates all the above requirements and which is suitable for audiovisual content of extended duration.

An example of a method designed for instantaneous quality judgement is the Single Stimulus Continuous Quality Evaluation (SSCQE) developed by the RACE MOSAIC project [41] and adopted as part of ITU-R recommendation BT.500-7 [65]. The SSCQE allows for continuous acquisition of votes (samples recorded twice a second) by means of a slider mechanism (score recording device) with an associated quality scale. The slider has to meet the following conditions: fixed or desk-mounted position, slider mechanism without any sprung position, linear range of travel of 10 cm. The operational principles of the SSCQE methodology can be seen in Fig. 4.7.

Such a setup mimics the objective evaluation approach based on acquisition of specific data from a real-time system. Therefore, it may be helpful in defining a link between subjective and objective evaluation of picture quality. A corresponding methodology, with the same principles as SSCQE, for continuous assessment of speech quality is described in ITU-T Rec. P.880 [69]. At present, the slider based methods are the only internationally accepted and recommended methodologies for instantaneous quality judgement.

Although the SSCQE allows for continuous quality assessment of long duration stimulus (up to 30 min), it's not free from disadvantages. It has been reported that the continuous operation of the slider can be too demanding for the test participants

and hence can influence the ratings negatively [45]. In addition, such activity may divert the user's attention from the process of quality assessment [140] and cause the differences in participants' reaction time to quality changes reducing the accuracy of the method [115]. Consequently, the resulting assessment values must be suitably compensated. Moreover, in order to obtain relatively stable results, the assessors need to be well-trained before inputting their assessment scores. All the mentioned drawbacks undermine the validity of the SSCQE for real life applications and services. Therefore, development and validation of alternative methodologies is needed.

Recently, several variations of the SSCQE method have been proposed with the slider device being replaced by e.g. a steering wheel [96] and a glove [20] or in case of 3D video quality assessment by a tablet [84]. Also, a comparison between several devices (mouse, throttle, sliding bar and joystick) which were used for quality rating of time-varying sequences has been reported in [107]. Except for the different rating instruments all those methods share the same methodological approach (same procedures and type of rating scale as SSCQE, often with reduced resolution). Thus they do not offer any new insights into the momentary quality assessment. Moreover, the performance improvement of the mentioned methods over the SSCQE has not been proven [107] and an effect of stimulus duration on users' fatigue related to usage of new rating devices has not been verified.

Generally, very few corresponding research activities devoted to the topic of continuous quality evaluation can be found, resulting in a small number of related publications. However, some interesting research findings have been claimed with respect to the momentary quality assessment. In [88] it has been concluded that subjects react almost immediately when a change from good to poor quality occurs while in the reverse situation the adaptation process is much slower. This asymmetry in tracking the quality has been confirmed in [23] where changes in momentary speech quality were evaluated. These aspects of reactions to quality changes are underpinning the suggested new methodology as presented in the next chapter.

5. A New Methodology for Momentary Quality Assessment

5.1. Scope of the Research

In Chapter 4, the various standardized methods for the subjective evaluation of audiovisual media were presented. It was pointed out that there is a lack of methods able to provide realistic approximation of perceived quality in case of multi-modal content of extended duration. Therefore, the main topic of this dissertation is the development of a methodology which allows low-attention, continuous quality assessment of long duration, audiovisual material.

A set of research questions have been formulated as part of this development:

- Do the quality expectations of human spectators change over time and with increased involvement in the content?
- When do users become aware of quality degradation under the assumption that no direct reference exists?
- Is the quality level at which the degradation is noticed similar to the quality level chosen by the user?
- Do users' quality expectations/requirements develop differently over extended periods of time for different types of content?
- Are viewers' quality requirements independent of the magnitude and appearance of quality impairments introduced?
- Does coherence/continuity of presented material influence quality preferences?

All these questions have been thoroughly examined using a novel subjective methodology that was developed especially for this purpose. Using an innovative approach for measuring users' audiovisual quality preferences/expectations, several milestones in the field of multi-modal quality assessment and human perception have been achieved (cf. section 'Overview of the Contributions by This Research'). Although the proposed test-methodology is applicable to many audiovisual applications, this work focuses only on one main scenario, namely, home cinema setup with HDTV projection. Other scenarios are considered to be included in future work. The thesis consists of two conference papers and three journal articles. A short summary and a specific input into the field of quality assessment is provided for each paper in the following section.

5.2. Overview of the Contributions by This Research

This work has been conducted in order to explore behavioural aspects related to quality perception of long duration audiovisual material. The research questions described in the previous section turned out to be difficult to answer using existing quality assessment methods. Limitations of the standardized techniques, as described in section 4.3, forced the development of a new methodology allowing to investigate dependencies between perceived quality and impairment variations over extended periods of time. The method has been proposed and described in detail in **Paper A**. The new technique addresses the problem of quality assessment in the context of a user's quality preferences/expectation rather than the ordinary quality judgment. Instead of measuring the quality by using traditional mean opinion score (MOS) based approaches, the method allows participants to select the most appreciated quality themselves. In case quality degradation occurs (as introduced by the experiment-running system), subjects have the possibility to adjust the quality to a desired level by using a rotary controller (e.g. a knob). During the assessment task, automatic quality alterations are introduced at random or periodically and stepwise (e.g. the degradation procedure begins every third minute and subsequently, the quality level decreases every 10 s). The participant's response (movement of the controller) to a quality change stops the automatic degradation procedure and provides him/her with full control over the quality adjustment. There are no physical limits in the rotation mechanism and the maximum quality can be surpassed if not recognized by the user, causing gradual decrease in quality. Rotating the device further clockwise introduces a gradual decrease in quality. This is a reversible process, which means the user can return to the reference quality by rotating the device in the opposite direction again. This is in a way a penalty introduced when the maximum quality level is being surpassed. The adjustment process is based on perception of quality changes solely (no tactile feedback from the device). The scale assigned to the assessment instrument, in fact, is a direct representation of the quality levels used in the test, and translation of the perceived quality into a numerical score or position of the rating device is not required. The operational principles of the method are visualized in Fig. 5.1.

In this case the adjustment procedure is represented by a simple switching between different quality levels prepared beforehand. However, it could be a real-time process with usage of e.g. the scalable video coding (SVC). The method allows to collect information about users' quality preferences in a non-intrusive way which makes it suitable for real life applications in which content of extended duration (e.g. TV programmes, movies, etc.) is used. Furthermore, the method seems to be very intuitive and easy to understand and therefore does not require thorough training of test subjects. Consequently, the time needed to conduct an experiment can be significantly reduced which is beneficial in case of long testing procedures.

The process behind the adjustment procedure can be directly related to the pairwise comparison based methods which are seen as the most accurate testing procedures [99]. This means that participants mark their quality preferences by in-

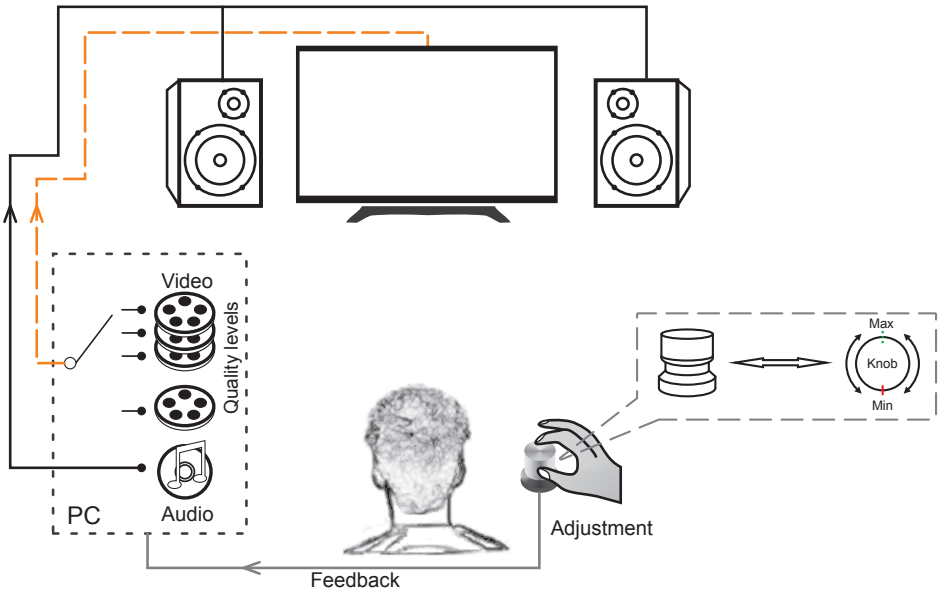


Figure 5.1.: Experimental setup of the test. 'Min' and 'Max' on this drawing are for explanatory purposes only and are not visible for the assessor.

stantaneously selecting the more appreciated quality level out of the two neighboring ones. As a result, a direct answer to which quality level is prefered/acceptable to a subject at a given point in time can be obtained. It is clearly an advantage when compared with other methods, wherein an acceptability threshold is hard to be determined (e.g. 'fair' or '3' on the rating scale doesn't tell us whether the quality is acceptable or not).

In addition to the extended description of the principles of the new method, the **Paper A** also contains preliminary results of a study conducted using this technique.

Several studies have been carried out using the new method in which the impact of the time dimension on users' Quality of Experience has been investigated. Those studies delivered large data sets, which needed to be appropriately processed in order to derive patterns and gain knowledge about the investigated phenomenon. Therefore, a guideline on how to prepare, analyze and interpret the data collected using the novel methodology was presented in **Paper B**. This article is a direct extension of **Paper A** and describes results of two studies in which an over 30-min long audiovisual sequence was evaluated. In the first study, the impairments were introduced only to the video domain whereas in the second study video and audio domains were distorted simultaneously. In both cases, it turned out that test-subjects are consistent with respect to their quality preferences over the stimulus' duration. It seems that the time dimension does not play a major role when it comes

to users' quality expectations and reactions to quality changes. More interestingly, it was found that there is a significant and relatively constant difference between the quality level at which participants notice a quality degradation and the quality level set by them during the adjustment procedure. In addition, it has been found that in case when both modalities are distorted simultaneously, the process of impairment discrimination and quality adjustment is easier for participants (there is smaller variation in the data) and that the quality levels set by them are higher on average.

A continuation of the previous work has been described in **Paper C**, which includes two different studies. Firstly, an influence of audio artifacts related to different compression rates on participants' reactions to quality changes was investigated. For this purpose a 32-min long audiovisual clip with impairments introduced into the audio domain solely was evaluated. Secondly, cross-modal effects between the visual and auditory modalities were examined. The overall goal of the study was to investigate possible dissimilarities in reactions of participants to quality changes of audiovisual presentation in case of audio only, video only, and simultaneous audio and video distortions. Both studies were carried out using the method described earlier in this section. The obtained results show that no matter which modality is impaired (audio only, video only or audio and video simultaneously), users' quality expectations are rather constant over the stimulus duration. However, significant differences between each of the above cases have been found, i.e. subjects reacted faster to quality changes and preferred higher quality levels when impairments were introduced in both modalities simultaneously. In addition, the suitability of our method for quality assessment of long duration audio streams (with accompanying video) has been demonstrated.

The next study, presented in **Paper D**, investigated the impact of content type on user reactions to quality changes. In the previous articles it was shown that the time dimension does not affect assessors' quality expectations and reactions to quality changes. However, the stimuli used in all of the previous studies were of similar nature and therefore further investigation with use of different types of content was needed. Through the study, indications were found which suggest that content type may have a significant effect on quality acceptance level over extended periods of time. However, this was proven only for the action-movie type content with a very high spatial and temporal activity throughout the entire duration of the stimulus. This phenomenon could be explained to some extent by the fact that high motion scenes are more affected by transmission artifacts (e.g. related to bandwidth limitations) than low motion scenes [45]. Nevertheless, the finding only gave an indication that such phenomenon might occur. Therefore, further investigation is necessary in this case. In addition, it was found that for long duration content, viewers' quality requirements are independent of magnitude and appearance (gradual increase vs. spontaneous leap) of quality impairments introduced. Moreover, it turned out that results obtained for one of the content types were congruous with results of a study described in **Paper B** in which similar experimental material was used. This demonstrates the repeatability of results obtained using the proposed method, thus making the method more desirable for purposes of quality assessment of long duration content than other traditional methods.

The last paper included in the thesis (**Paper E**) investigated the effect of content coherence/continuity as well as the influence of video stream on audio quality preferences. Obtained results show that audio quality expectations do not change significantly over the extension of a clip span, no matter if the accompanying visual stimulus is displayed or not. However, the perception of the auditory stimulus appears to be significantly influenced by the presence of video. It was observed that the users' audio quality preferences were higher when audio was played without the presence of an accompanying visual stimulus. This was true no matter if the presentation of a stimulus started with the highest or the lowest audio quality. Moreover, it also held for different ways of introducing the audio quality degradations (stepwise over time or by a large drop). Furthermore, results showed that the audio quality preferences were lower when the entire presented material was played in a continuous manner than when the same clip was cut into segments which were reproduced in a random order with short pauses in between. This may mean that subjective studies that use short audio stimuli extracted from long-duration content generate results that exaggerate the actual quality needs of consumers.

5.3. Conclusions

This thesis is devoted to the topic of temporal development of Quality of Experience and contributes to the field of QoE in two ways. The major contribution is the novel subjective assessment methodology developed in order to investigate the influence of stimulus duration on quality perception of audiovisual content.

The method was proposed as a result of lack of alternative methodologies that allow to answer research questions introduced in this work. Compared to the conventional techniques for continuous quality evaluation (e.g. SSCQE) the proposed quality adjustment approach allows for gathering the data with little psycho-physical demands and without the need to translate the perceived quality into a numerical value or its semantic designator.

Although new findings are possible with the method, it should not be treated as a direct replacement of the existing ITU-R rating-scales-based methodologies, but rather as a new source of information with respect to the user's cognitive experiences. The new technique cannot provide absolute quality scores like the SSCQE method, hence the direct comparison between the two is not straightforward and requires a thorough investigation. This is a topic for further research.

The other contribution is the outcome of several studies conducted using the new method. Multiple findings related to users' QoE are presented in this dissertation (Part II. Included Papers). The most important, however, can be summarized as follows:

- The time dimension does not affect user consistency with respect to quality selection. The only exception that was found was related to a very high motion content with intense action scenes throughout its entire duration. However, whether this is a general trend or not has not been verified yet.

- Impairments introduced in both modalities simultaneously cause faster reaction to quality changes, which make the process of quality discrimination much easier for participants, which in turn leads to the selection of, or requirement for, higher quality levels.

- The sensitivity to quality variations highly depends on the awareness of the process of quality changes, and is higher when the subjects are in charge of the quality adjustment, than when the process is controlled externally. These findings hold, no matter which way the degradations appear in the presented material; whether stepwise over time or immediately, in one move.

- The continuity of the stimulus affect the perception of audio quality changes. It seems that the audio quality preferences are lower when the entire presented material, with a semantic structure preserved, is evaluated than when it is cut into smaller pieces and played randomly. This may suggest that results of a typical subjective study in which short audio stimuli are used are exaggerated.

As mentioned earlier, the validity of the mentioned findings have been verified for the home cinema scenario (with HDTV projection) only and therefore a general validity of the results needs to be checked for other usage scenarios (e.g. Scalable Video Coding) and in different contexts, e.g. mobile context. For the latter, some adjustments in the proposed methodology may be required due to specific technical requirements of the adjustment device. A solution in such a case could be a small controller with a Bluetooth interface attached to the back of a mobile device (e.g. smartphone, tablet) or application-based touch adjustment (rotary moves on the surface of a touch screen). This could be the next step in the validation process of the obtained results and to confirm the suitability of the proposed methodology in different contexts.

Summarizing, the thesis explores the field of the momentary QoE in an innovative way providing results valuable for current and future research. The outcome of this research helps us to better understand perception of time-varying audiovisual quality in more realistic applications. The findings mentioned above together with the novel methodology extend the existing literature and provide new insights into the field of quality assessment.

References

[1] *Encyclopedia of Information Science and Technology, Second Edition*. Edited by Khosrow-Pour, M., 2008.

[2] *TV content analysis: Techniques and applications*. Y. Kompatsiaris, B. Merialdo and S. Lian (eds.). CRC Press, 2012.

[3] *Quality of Experience. Advanced Concepts, Applications and Methods*. S. Möller and A. Raake (eds.). Springer, 2014.

[4] *Multimedia Quality of Experience (QoE): Current Status and Future Requirements*. C.W. Chen, P. Chatzimisios, T. Dagiuklas and L. Atzori (eds.). John Wiley & Sons, 2016.

[5] T. Alpert and J. P. Evian. Subjective quality evaluation - the SSCQE and DSCQE methodologies. In *Ebu Technical Review*, pages 12–20, 1997.

[6] ANSI-Accredited Committee T1 Contribution. Report on an experimental combined audio/video subjective test method. Bellcore, T1A1.5/93-104, Red Bank, New Jersey, 1993.

[7] ANSI-Accredited Committee T1 Contribution. Report on extension of combined audio/video quality model. Bellcore, T1A1.5/94-141, Red Bank, New Jersey, 1993.

[8] ANSI-Accredited Committee T1 Contribution. Combined A/V model with multiple audio and video impairments. Bellcore, T1A1.5/94-124, Red Bank, New Jersey, 1995.

[9] S. Arndt, J. N. Antons, R. Schleicher, and S. Möller. Using electroencephalography to analyze sleepiness due to low-quality audiovisual stimuli. *Signal Processing: Image Communication*, 42:120–129, 2016.

[10] H. Becerra Martinez and M. C. Q. Farias. A no-reference audio-visual video quality metric. In *Proc. of the 22nd European Signal Processing Conference (EUSIPCO)*, pages 2125–2129, Lisbon, 2014.

[11] J. G. Beerends and F. E. De Caluwe. The influence of video quality on perceived audio quality and vice versa. *Journal of the Audio Engineering Society*, 47(5):355–362, 1999.

[12] D. M. Behne, M. Alm, A. Berg, T. Engel, C. Foyn, C. Johnsen, T. Srigaran, and A. Torsdottir. Effects of musical experience on perception of audiovisual synchrony for speech and music. *Journal of the Acoustical Society of America*, 19(1):1–6, 2013.

[13] B. Belmudez. *Audiovisual Quality Assessment and Prediction for Videotelephony*. Springer, 2015.

[14] B. Belmudez, S. Möller, B. Lewcio, A. Raake, and A. Mehmood. Audio and video channel impact on perceived audio-visual quality in different interactive contexts. In *Proc. of IEEE International Workshop on Multimedia Signal Processing (MMSP '09)*, pages 1–5, Rio De Janeiro, 2009.

[15] J. Beyer, R. Varbelow, J. N. Antons, and S. Möller. Using electroencephalography and subjective self-assessment to measure the influence of quality variations in cloud gaming. In *Proc. of the 7th International Workshop on Quality of Multimedia Experience (QoMEX 15)*, pages 1–6, Pylos-Nestoras, 2015.

[16] N. Bolognini, F. Frassinetti, A. Serino, and E. L'adavas. "Acoustical vision" of below threshold stimuli: Interaction among spatially converging audiovisual inputs. *Experimental Brain Research*, 160(3):273–282, 2005.

[17] A. Borowiak, U. Reiter, and U. P. Svensson. Evaluation of audio quality requirements over extended periods of time using long duration audiovisual content. In *Advances in Multimedia Information Processing. PCM*, volume 7674 of *Lecture Notes in Computer Science*, pages 10–20, 2012.

[18] A. Borowiak, U. Reiter, and O. Tomic. Measuring the quality of long duration AV content. Analysis of test subject/time interval dependencies. In *EuroITV - Adjunct Proceedings*, pages 266–269, Berlin, 2012.

[19] C. C. Bracken. Presence and image quality: The case of high-definition television. *Media Psychology*, 7(2):191—-205, 2005.

[20] S. Buchinger, W. Robitza, M. Nezveda, M. Sack, P. Hummelbrunner, and H. Hlavacs. Slider or glove? Proposing an alternative quality rating methodology. In *Proc. of the 5th International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, Scottsdale, Arizona, 2010.

[21] K. T. Chen, C. C. Wu, Y. C. Chang, and C. L. Lei. A crowdsourceable QoE evaluation framework for multimedia content. In *Proc. of the 17th ACM international conference on Multimedia*, pages 491–500, Beijing, 2009.

[22] Y. Chen, K. Wu, and Q. Zhang. From QoS to QoE: A tutorial on video quality assessment. *Perception*, 17(2):1126–1165, 2014.

[23] A. Clark. Modeling the effect of burst packet loss and recency on subjective voice quality. In *Proc. of the Internet Telephony Workshop (IPtel 2001)*, pages 123–127, New York, 2001.

[24] I. D. V. Committee. Long form video overview, 2009. http://www.iab.net/media/file/long-form-video-final.pdf.

[25] S. Coren, L. M. Ward, and J. T. Enns. *Sensation and Perception (6th Edition)*. John Wiley & Sons, 2004.

[26] E. Danish, A. Fernando, M. Alreshoodi, and J. Woods. A hybrid prediction model for video quality by QoS/QoE mapping in wireless streaming. In *Proc. of 2015 IEEE International Conference on Communication Workshop (ICCW)*, pages 1723 – 1728, London, 2015.

[27] D. De Pessemier, I. Stevens, L. De Marez, L. Martens, and W. Joseph. Analysis of the quality of experience of a commercial voice-over-IP service. *Multimedia Tools and Applications*, 74(15):5873–5895, 2015.

[28] N. F. Dixon and L. Spitz. The detection of auditory visual desynchrony. *Perception*, 9(6):719–721, 1980.

[29] EBU BPN 056. SAMVIQ – Subjective Assessment Methodology for Video Quality. Report by EBU Project Group B/VIM (Video In Multimedia), 2003.

[30] R. Eg and D. M. Behne. Perceived synchrony for realistic and dynamic audiovisual events. *Frontiers in Psychology*, 6(736):1–12, 2015.

[31] R. Eg, C. Griwodz, P. Halvorsen, and D. M. Behne. Audiovisual robustness: exploring perceptual tolerance to asynchrony and quality distortion. *Multimedia Tools and Applications*, 74(2):345–365, 2015.

[32] A. Eichhorn, P. Ni, and R. Eg. Randomised Pair Comparison - An economic and robust method for audiovisual quality assessment. In *Proc. of the 20th International Workshop on Network and Operating Systems Support for Digital Audio and Video (NOSSDAV 2010)*, pages 63–68, Amsterdam, 2010.

[33] M. Eldridge and E. Saltzman. Seeing what you hear: Visual feedback improves pitch recognition. *European Journal of Cognitive Psychology*, 22(7):1078–1091, 2010.

[34] M. O. Ernst and H. H. Bülthoff. Merging the senses into a robust percept. *Trends in Cognitive Sciences*, 8(4):162–169, 2004.

[35] R. A. Farrugia and C. J. Debono. *Multimedia Networking and Coding*. 2012.

[36] M. Fiedler, T. Hossfeld, and P. Tran-Gia. A generic quantitative relationship between quality of experience and quality of service. *IEEE Network Magazine*, 24(2):36–41, 2010.

[37] M. N. Garcia, R. Schleicher, and A. Raake. Impairment-factor-based audiovisual quality model for IPTV: influence of video resolution, degradation type, and content type. *EURASIP Journal on Image and Video Processing*, 2011(1):1–14, 2011.

[38] E. B. Goldstein. *Sensation and perception, 9th edition*. Cangage Learning, 2013.

[39] P. Gray, R. Massara, and M. Hollier. An experimental investigation of the accumulation of perceived error in time-varying speech distortions. In *AES Preprints: AES 103rd Convention*, number 4588, 1997.

[40] J. Gutierrez, P. Perez, F. Jaureguizar, J. Cabrera, and N. Garcia. Subjective evaluation of transmission errors in IPTV and 3DTV. In *Proc. of Visual Communications and Image Processing*, pages 1–4, Tainan, 2011.

[41] R. Hamberg and H. de Ridder. Continuous assessment of perceptual image quality. *Journal of the Optical Society of America*, 12(12):2573–2577, 1995.

[42] R. Hamberg and H. de Ridder. Time-varying image quality: Modeling the relation between instantaneous and overall quality. *SMPTE Motion Image Journal*, 108(11):802—811, 1999.

[43] X. Han, Y. Wei, and X. Xie. An audiovisual objective quality model based on bp neutral network. In *Proc. of 2011 International Conference on Multimedia Technology (ICMT)*, pages 5277–5288, Hangzhou, 2011.

[44] D. S. Hands. A basic multimedia quality model. *IEE Transactions on Multimedia*, 6(6):806—816, 2004.

[45] D. S. Hands and S. E. Avons. Recency and duration neglect in subjective assessment of television picture quality. *Journal of Aplied Cognitive Psychology*, 15(6):639—657, 2001.

[46] T. Hayashi, K. Yamagishi, T. Tominaga, and A. Takahashi. Multimedia quality integration function for videophone services. In *Proc. of Global Telecommunications Conference, GLOBECOM 07. IEEE*, pages 2735–2739, Washington, DC, 2007.

[47] R. M. Hogarth and H. J. Einhorn. Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology*, 24(1):1–50, 1992.

[48] M. P. Hollier, A. N. Rimell, D. S. Hands, and R. M. Voelcker. Multi-modal perception. *BT Technology Journal*, 17(1):35–46, 1999.

[49] M. P. Hollier and R. Voelcker. Objective performance assessment: Video quality as an influence on audio perception. In *Presented at the 103rd AES Convention*, New York, 1997.

[50] N. P. Holmes and C. Spence. Multisensory integration: Space, time and superadditivity. *Current Biology*, 15(18):762–764, 2005.

[51] T. Hosfeld, P. E. Heegaard, and M. Varela. QoE beyond the MOS: Added value using quantiles and distributions. In *Proc. of the 7th International Workshop on Quality of Multimedia Experience (QoMEX 15)*, pages 1–6, Pylos-Nestoras, 2015.

[52] B. Ionescu, K. Seyerlehner, K. Rasche, C. Vertan, and P. Lambert. Content-based video description for automatic video genre categorization. In *Proc. of the 18th International Conference on MultiMedia Modelling (MMM-2012)*, volume 7131, pages 51–62, Klagenfurt, 2012.

[53] ITU-R Rec. BS.1116. Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems. Int. Telecomm. Union, Geneva, 1997.

[54] ITU-R Rec. BS.1284-1. General methods for the subjective assessment of sound quality. Int. Telecomm. Union, Geneva, 2003.

[55] ITU-R Rec. BS.1285. Pre-selection methods for the subjective assessment of small impairments in audio systems. Int. Telecomm. Union, Geneva, 1997.

[56] ITU-R Rec. BS.1286. Methods for the subjective assessment of audio systems with accompanying picture. Int. Telecomm. Union, Geneva, 1997.

[57] ITU-R Rec. BS.1359. Relative timing of sound and vision for broadcasting. Int. Telecomm. Union, Geneva, 2001.

[58] ITU-R Rec. BS.1534-1. Method for the subjective assessment of intermediate quality level of coding systems. Int. Telecomm. Union, Geneva, 2001-2003.

[59] ITU-R Rec. BS.710-4. Subjective assessment methods for image quality in high-definition television. Int. Telecomm. Union, Geneva, 1998.

[60] ITU-R Rec. BS.775-3. Multichannel stereophonic sound system with and without accompanying picture. Int. Telecomm. Union, Geneva, 2012.

[61] ITU-R Rec. BT.1129-2. Subjective assessment of standard definition digital television (SDTV) systems. Int. Telecomm. Union, Geneva, 1998.

[62] ITU-R Rec. BT.1788. Methodology for the subjective assessment of video quality in multimedia applications. Int. Telecomm. Union, Geneva, 2007.

[63] ITU-R Rec. BTS.1679-1. Subjective assessment of the quality of audio in large screen digital imagery applications intended for presentation in a theatrical environment. Int. Telecomm. Union, Geneva, 2015.

[64] ITU-T Rec. BT.500-13. Methodology for the subjective assessment of the quality of television pictures. Int. Telecomm. Union, Geneva, 2012.

[65] ITU-T Rec. BT.500-7. Methodology for the subjective assessment of the quality of television pictures. Int. Telecomm. Union, Geneva, 1996.

[66] ITU-T Rec. E.800. Definitions of terms related to quality of service. Int. Telecomm. Union, Geneva, 2008.

[67] ITU-T Rec. G.1080. Quality of experience requirements for IPTV services. Int. Telecomm. Union, Geneva, 2008.

[68] ITU-T Rec. J.148. Requirements for an objective perceptual multimedia quality model. Int. Telecomm. Union, Geneva, 2003.

[69] ITU-T Rec. P.800. Methods for subjective determination of transmission quality. Int. Telecomm. Union, Geneva, 1996.

[70] ITU-T Rec. P.910. Subjective video quality assessment methods for multimedia applications. Int. Telecomm. Union, Geneva, 2008.

[71] ITU-T Rec. P.911. Subjective audiovisual quality assessment methods for multimedia applications. Int. Telecomm. Union, Geneva, 1998.

[72] ITU-T Rec. P.913. Methods for the subjective assessment of video quality, audio quality and audiovisual quality of internet video and distribution quality television in any environment. Int. Telecomm. Union, Geneva, 2014.

[73] ITU-T Rec. P.920. Interactive test methods for audiovisual communications. Int. Telecomm. Union, Geneva, 1996.

[74] ITU-T Rec. P.930. Principles of a reference impairment system for video. Int. Telecomm. Union, Geneva, 1996.

[75] ITU-T Rec. P.931. Multimedia communications delay, synchronization and frame rate measurement. Int. Telecomm. Union, Geneva, 1998.

[76] Joint Video Group. Terms of reference for Joint Video Team (JVT) activities, 2008.

[77] C. Jones and D. J. Atkinson. Development of opinion-based audiovisual quality models for desktop video-teleconferencing. In *Proc. of the 6th International Workshop on Quality of Services (IWQoS 98)*, pages 196–203, Napa Valley, CA, 1998.

[78] E. E. Jones and G. R. Goethals. Order effects in impresion formation: Attribution context and the nature of the entity. *Attribution: Perceiving the causes of behavior*, pages 27–46, 1972.

[79] S. Jumisko-Pyykkö. User-centered quality of experience and its evaluation methods for mobile television. *Doctoral dissertation, Tampere University of Technology*, 2011.

[80] S. Jumisko-Pyykkö and J. Häkkinen. Evaluation of subjective video quality on mobile devices. In *Proc. of ACM Multimedia*, volume 6507, pages 535–538, Singapore, 2005.

[81] S. Jumisko-Pyykkö, J. Häkkinen, and G. Nyman. Experienced quality factors: qualitative evaluation approach to audiovisual quality. In *Proc. of SPIE Multimedia on Mobile Devices*, volume 6507, page 12, 2007.

[82] S. Jumisko-Pyykkö, M. V. Vinod Kumar, and J. Korhonen. Unacceptability of instantaneous errors in mobile television: From annoying audio to video. In *Proc. of 8th International Conference on Human Computer Interaction with Mobile Devices and Services (MobileHCI 2006)*, pages 1–8, Espoo, 2006.

[83] R. Kassier, S. Zielinski, and F. Rumsey. Computer games and multichannel audio quality part II – evaluation of time-variant audio degradations under divided and undivided attention. In *Presented at the 115th AES Convention*, New York, 2003.

[84] T. Kim, J. Kang, S. Lee, and A. C. Bovik. Multimodal interactive continuous scoring of subjective 3D video quality of experience. *IEEE Transactions on Multimedia*, 16(2):387–402, 2014.

[85] N. Kitawaki, Y. Arayama, and T. Yamada. Multimedia opinion model based on media interaction of audio-visual communications. In *Proc. of the 4th International Conference on Measurement of Speech and Audio Quality in Networks*, pages 5–10, Prague, 2005.

[86] A. Kohlrausch and S. van de Par. Auditory-visual interaction: From fundamental research in cognitive psychology to (possible) applications. *SPIE Human Vision and Electronic Imaging IV*, 3644:34–44, 1999.

[87] A. Kohlrausch and S. van de Par. *Audio—Visual Interaction in the Context of Multi-Media Applications. In book: Communication Acoustics*. J. Blauert (ed.), Springer, 2005.

[88] A. Kokotopoulos. Subjective assessment of multimedia systems for distance learning. In *Proc. of European Conference on Multimedia Applications, Services and Techniques (ECMAST 97)*, volume 1242, pages 395–408, Milan, 1997.

[89] J. Korhonen, U. Reiter, and J. You. Subjective comparison of temporal and quality scalability. In *Proc. of the 3rd International Workshop on Quality of Multimedia Experience (QoMEX 11)*, pages 161–166, Mechelen, 2011.

[90] P. Kortum and M. A. Sullivan. Content is king: The effect of content on the perception of video quality. In *Proc. of the Human Factors and Ergonomics Society 48th Annual Meeting*, volume 48, pages 1910–1914, Santa Monica, 2004.

[91] P. Kortum and M. A. Sullivan. The effect of content desirability on subjective video quality ratings. *The Journal of the Human Factors and Ergonomics Society*, 52(1):105–118, 2010.

[92] F. Kozamernik, V. Steinman, P. Sunna, and E. Wyckens. A new EBU methodology for video quality evaluations in multimedia. *Motion Imaging Journal, SMPTE*, 114(4):152–116, 2015. First time published in 2005.

[93] C. Li. Primacy effect or recency effect? A long-term memory test of the 2006 super bowl commercials. In *Proc. of Proceedings of the 2007 Academy of Marketing Science (AMS) Annual Conference*, pages 1–4, 2014.

[94] C. Li and T. Chen. Aesthetic visual quality assessment of paintings. *IEEE Journal of Selected Topics in Signal Processing*, 3(2):236—252, 2009.

[95] Y. Li, S. Narayanan, and C. C. J. Kuo. Content-based movie analysis and indexing based on audiovisual cues. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(8):1073—-1085, 2004.

[96] T. Liu, G. Cash, N. Narvekar, and J. Bloom. Continuous mobile video subjective quality assessment using gaming steering wheel. In *Proc. of the 6th International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, page 6, Scottsdale, Arizona, 2012.

[97] T. Mäki, D. Kukolj, D. Dordevic, and M. Varela. A reduced-reference parametric model for audiovisual quality of IPTV services. In *Proc. of the 5th International Workshop on Quality of Multimedia Experience (QoMEX 13)*, pages 6–11, Klagenfurt, 2013.

[98] W. A. C. Mansilla. Quality of aesthetic experience and implicit modulating factors. *Doctoral dissertation, Norwegian University of Technology*, 2013.

[99] R. K. Mantiuk, A. Tomaszewska, and R. Mantiuk. Comparison of four subjective methods for image quality assessment. In *Computer Graphics Forum*, volume 31, pages 2478–2491, Amsterdam, 2012.

[100] N. Matos and F. Pereira. Using MPEG-7 for generic audiovisual content automatic summarization. In *Proc. of the 9th International Workshop on Image Analysis for Multimedia Interactive Services*, pages 41–45, Klagenfurt, 2008.

[101] M. A. McFarland, M. H. Pinson, C. Ford, A. A. Webster, W. J. Ingram, S. Haines, and K. Anderson. NTIA Tech. Memo TM-10-472: Relating audio and video quality using CIF video. ITS, NTIA, U.S. Department of Commerce, Boulder, 2010. Available: http://www.its.bldrdoc.gov/pub/ ntia-rpt/10-472/.

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

[103] D. G. Myers. *Psychology, 9th Edition*. Worth Publisher, 2010.

[104] K. Nahrstedt and R. Steinmetz. Resource management in networked multimedia systems. *IEEE Computer*, 28(5):52–63, 1995.

[105] A. B. Nauman, I. Wechsung, and J. Hurtienne. Multimodal interaction: A suitable strategy for including older users? *Interacting with Computers*, 22(6):465—474, 2010.

[106] U. Neisser. *Cognition and Reality: Principles and Implications of Cognitive Psychology*. W. H. Freeman, 1976.

[107] O. Nemethova, M. Ries, A. Dantcheva, and S. Fikar. Test equipment of time-variant subjective perceptual video quality in mobile terminals. In *Proc. of International Conference on Human Computer Interaction*, Phoenix, 2005.

[108] P. Noll. *MPEG digital audio coding standards. In book: The Digital Signal Processing Handbook*. IEEE Press/CRC Press, 1998.

[109] H. L. O'Brien and E. G. Toms. What is user engagement? A conceptual framework for defining user engagement with technology. *Journal of the American Society for Information Science and Technology*, 59(6):938—955, 2008.

[110] Ooyala. Global Video Index Q1 2014, 2014. http://go.ooyala.com/rs/OOYALA/images/Ooyala-Global-Video-Index-Q1-2014.pdf.

[111] Ooyala. Global Video Index Q3 2014, 2014. http://go.ooyala.com/rs/OOYALA/images/Ooyala-Global-Video-Index-Q3-2014.pdf.

[112] A. Ostaszewska-Liżewska, R. Kłoda, and S. Żebrowska Łucyk. Multimodal perception in subjective quality evaluation of compressed video. *Advanced Mechatronics Solutions*, 393(1):569–574, 2015.

[113] M. H. Pinson, W. Ingram, and A. Webster. Audiovisual quality components: An analysis. *IEEE Signal Processing Magazine*, 28(6):66–67, 2011.

[114] M. H. Pinson, M. Sullivan, and A. A. Catellier. A new method for immersive audiovisual subjective testing. In *Proc. of 8th International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM 2014)*, page 6, Chandler, AZ, 2014.

[115] M. H. Pinson and S. Wolf. Comparing subjective video quality testing methodologies. In *Visual Communications and Image Processing. Proceedings of the SPIE*, volume 5150, pages 573–582, 2003.

[116] A. Raake and S. Möller. *Quality and Quality of Experience. In book: Quality of Experience: Advanced Concepts, Applications and Methods*. S. Möller and A. Raake (eds.). Springer, 2014.

[117] J. Redi, H. Liu, H. Alers, R. Zunino, and I. Heynderickx. Comparing subjective image quality measurement methods for the creation of public databases. In *Proc. SPIE 7529, Image Quality and System Performance VII*, 2010.

[118] B. Reeves and C. Nass. *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge University Press, 1996.

[119] A. Rehman and Z. Wang. Perceptual experience of time-varying video quality. In *Proc. of the 5th International Workshop on Quality of Multimedia Experience (QoMEX 13)*, pages 218–223, Klagenfurt, 2013.

[120] K. Rehman Laghari, N. Crespi, B. Molina, and C. E. Palau. QoE aware service delivery in distributed environment. In *Proc. of IEEE Conference on Advanced Information Networking and Applications (WAINA)*, pages 837–842, 2011.

[121] P. Reichl, S. Egger, R. Schatz, and A. D'Alconzo. The logarithmic nature of QoE and the role of the Weber-Fechner Law in QoE assessment. In *Proc. of 2010 IEEE International Conference on Communications (ICC)*, pages 1–5, Cape Town, 2010.

[122] U. Reiter. Towards a classification of audiovisual media content. In *Proc. of the 129th Convention of the Audio Engineering Society*, 2010.

[123] U. Reiter, K. Brunnström, K. De Moor, M. C. Larabi, M. Pereira, A. Pinheiro, Y. You, and A. Zgank. *Factors Influencing Quality of Experience. In book: Quality of Experience: Advanced Concepts, Applications and Methods*. S. Möller and A. Raake (eds.). Springer, 2014.

[124] U. Reiter and K. De Moor. Content categorization based on implicit and explicit user feedback: Combining self-reports with EEG emotional state analysis. In *Proc. of the 4th International Workshop on Quality of Multimedia Experience (QoMEX 12)*, pages 266–271, Yarra Valley, VIC, 2012.

[125] U. Reiter and J. You. Estimating perceived audiovisual and multimedia quality - A survey. In *Proc. of the 14th IEEE International Symposium on Consumer Electronics (ISCE)*, pages 1–6, Braunschweig, 2010.

[126] W. Robitza, M. Garcia, and A. Raake. At home in the lab: Assessing audiovisual quality of HTTP-based adaptive streaming with an immersive test paradigm. In *Proc. of the 7th International Workshop on Quality of Multimedia Experience (QoMEX 15)*, pages 1–6, Pylos-Nestoras, 2015.

[127] L. A. Ross, D. Saint-Amour, V. M. Leavitt, D. C. Javitt, and J. J. Foxe. Do you see what i am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cereb Cortex*, 17:1147–1153, 2007.

[128] V. Santangelo and C. Spence. Multisensory cues capture spatial attention regardless of perceptual load. *Journal of Experimental Psychology: Human Perception and Performance*, 33(6):1311–1321, 2007.

[129] L. Shams and K. Robin. Crossmodal influences on visual perception. *Physics of Life Reviews*, 7(3):269—284, 2010.

[130] N. Staelens, S. Moens, W. Van den Broeck, I. Marien, B. Vermuelen, P. Lambert, R. Van de Walle, and P. Demeester. Assessing quality of experience of IPTV and video on demand services in real-life environments. *IEEE Transactions on Broadcasting*, 56(4):458–466, 2010.

[131] R. L. Storms and M. J. Zyda. Interactions in perceived quality of auditory-visual displays. *Presence: Teleoperators and Virtual Environments*, 9(6):557–580, 2000.

[132] D. Strohmeier. Open Profiling of Quality: A mixed methods research approach for audiovisual quality evaluations. *Doctoral dissertation, Ilmenau University of Technology*, 2011.

[133] G. J. Sullivan, J. R. Ohm, W. J. Han, and T. Wiegand. Overview of the High Efficiency Video Coding (HEVC) Standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 22(12):1649–1668, 2012.

[134] G. J. Sullivan, P. Topiwala, and A. Luthra. The H.264/AVC Advanced Video Coding Standard: Overview and introduction to the fidelity range extensions. In *Proc. of SPIE conference on Applications of Digital Image Processing XXVII*, volume 5558, pages 454–474, 2004.

[135] P. Waingankar and D. Valsan. Audio-video synchronization. In *Proc. of the International Conference & Workshop on Emerging Trends in Technology*, pages 202–205, 2011.

[136] Y. Wang, Z. Liu, and J. C. Huang. Multimedia content analysis-using both audio and visual clues. *IEEE Signal Processing Magazine*, 17(6):12–36, 2000.

[137] I. Wechsung, M. Schulz, K. P. Engelbrecht, J. Niemann, and S. Möller. Rate control for H.264 with two-step quantization parameter determination but single-pass encoding. In *Proc. of the Paralinguistic Information and its Integration in Spoken Dialogue Systems Workshop*, pages 175—186, 2011.

[138] B. Weiss, D. Guse, S. Möller, A. Raake, A. Borowiak, and U. Reiter. *Temporal Development of Quality of Experience. In book: Quality of Experience: Advanced Concepts, Applications and Methods.* S. Möller and A. Raake (eds.). Springer, 2014.

[139] T. Wiegand, G. J. Sullivan, G. Bjøntegaard, and A. Luthra. Overview of the H.264/AVC video coding standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 13(7):560–576, 2003.

[140] S. Winkler and F. Dufaux. Video quality evaluation for mobile applications. In *Visual Communications and Image Processing. Proc. of the SPIE*, volume 5150, pages 593—603, 2003.

[141] S. Winkler and C. Faller. Audiovisual quality evaluation of low-bitrate video. In *Proc. of the SPIE/IS&T Human Vision and Electronic Imaging*, volume 5666, pages 139—148, San Jose, 2005.

[142] S. Winkler and C. Faller. Perceived audiovisual quality of low-bitrate multimedia content. *IEEE, Transactions on Multimedia*, 5150:973—980, 2006.

[143] S. Wolf and M. H. Pinson. NTIA Tech. Rep. 02-392: Video quality measurement techniques. ITS, NTIA, U.S. Department of Commerce, Boulder, 2002.

[144] K. M. Wolters, K. P. Engelbrecht, F. Gödde, S. Möller, A. Naumann, and R. Schleicher. Making it easier for older people to talk to smart homes: the effect of early help prompts. *Universal Access in the Information Society*, 9(4):311–325, 2010.

[145] J. You, U. Reiter, M. M. Hannuksela, M. Gabbouj, and A. Perkis. Perceptual-based quality assessment for audio-visual services: A survey. *Journal of Signal Processing: Image Communication*, 25(7):482–501, 2010.

[146] S. K. Zielinski, F. Rumsey, and S. Bech. Effects of bandwidth limitation on audio quality in consumer multichannel audiovisual delivery systems. *Journal of The Audio Engineering Society*, 51(6):475–501, 2003.

[147] T. Zinner, O. Hohlfeld, O. Abboud, and T. Hossfeld. Impact of frame rate and resolution on objective QoE metrics. In *Proc. of the 2nd International Workshop on Quality of Multimedia Experience (QoMEX 10)*, pages 29–34, Trondheim, 2010.

Part II

INCLUDED PAPERS

A. Quality Evaluation of Long Duration Audiovisual Content

Paper A

Adam Borowiak[1], Ulrich Reiter[1], U. Peter Svensson[1]

[1] *Centre for Quantifiable Quality of Service in Communication Systems, Norwegian University of Science and Technology (NTNU)*

Proceedings of The 9th Annual IEEE Consumer Communications and Networking Conference (CCNC). Special Session on Quality of Experience (QoE) for Multime-dia Communications, pp. 353–357, Las Vegas, 2012.

Abstract

In this paper a new methodology for the evaluation of perceived quality for long duration audiovisual content is presented. This method allows the investigation of unexplored dependencies between perceived quality and impairment variations over extended periods of time. Instead of providing quality scores on predefined rating scales, assessors are asked to adjust the quality level of the displayed content themselves, whenever a degradation in perceived quality occurs. This method approaches the problem of quality evaluation from a different perspective than the existing methods, which might be valuable for future research. Preliminary results of an experiment conducted using this methodology are presented.

A.1. Introduction

The perceived quality of audiovisual and multi-modal systems plays a very important role in today's technologically advanced world. Multimedia services are growing in popularity and technological progress presents new possibilities to end-users regarding types of content (e.g. 2D, 3D), transmission schemes (e.g. traditional broadcast, internet, mobile networks) and devices used to display the content. Having access to the latest technology, users become more demanding and their expectations regarding offered quality are successively increasing. Most of the service/content providers strive to meet user's expectations and provide increasingly higher quality to improve their customers' overall experience. Providing the highest available quality to everyone would be the ideal solution but unfortunately in many cases there are a number of technical limitations (e.g. bandwidth, device constraints) which can degrade the end-user's experience. In such cases, knowing the acceptable level of quality becomes a crucial requirement for service suppliers. For this purpose objective metrics are desirable, and they need to be based on evaluations with human assessors.

Existing measuring techniques for audiovisual services are not optimal yet. These techniques are rather close to the traditional quality of service concept, which aims to capture the system related characteristics and frequently ignoring the perceptual and cognitive sides of Quality of Experience (QoE) [2]. Therefore, there is a need to continue to develop new evaluation methodologies and objective metrics which could, as accurately as possible, assess the complex impression of human audiovisual quality perception. Quality perception is a very subjective process which can be influenced by many factors, especially in multi-modal systems, where audio and video modalities might influence each other. Taking these aspects into account, as well as the fact that current objective metrics often correlate weakly to the user opinion we can easily understand why experiments with human assessors are considered to be the most accurate methods reflecting the human perception [5]. In spite of limitations and weak points, such experiments are invaluable tools, helping us to better understand user responses and providing results which are the foundations for any objective quality metric.

Most of available quality evaluation tests are usually designed to be performed in controlled laboratory environments, where the duration of stimuli is usually very short and where full attention from the assessors is expected. This makes them not reliable for real applications scenarios, in which the perceptual and cognitive factors play a very important role, and which comprise much longer durations of content [2].

Long–term audiovisual quality assessment is still an unexplored field of science. This is due to a very limited number of internationally accepted recommendations in the field. Only one such recommendation exists [5]. Moreover, the recommendation is not suitable for test sequences longer than 30 min and is not intended for the evaluation of a full movie or complete TV programmes. Limiting the duration of test sequences to a few seconds might affect an assessor's judgment and thus the

overall measurement of QoE [7]. Long–term aspects of quality assessment should be investigated more accurately to give us a better understanding of the cognitive side of human perception and also to expand our knowledge in relation to important aspects of QoE such as user interest and expectations for audiovisual material. To do so we need new methods which allow us to collect subjective data in as little invasive way as possible. This way participants could focus on the presented stimuli instead of the usually demanding task of assessing quality.

This paper is organized as follows: in Section A.2, we motivate the need for new quality assessment methods, describing existing methodologies and their major drawbacks. Section A.3 contains the detailed description of a novel method and of a test which was conducted to evaluate it. In Section A.4 preliminary results are presented and discussed. Finally, conclusions are given and potential further work is outlined in Section A.5.

A.2. Motivation

Existing methodologies for perceived quality assessment are typically designed for short duration sequences (from 10 to 15 s). These methods do not consider the time dimension as an important factor contributing to overall QoE. A number of degradations in function of time exert a significant effect on the overall experience. Let's consider a short audiovisual clip with one impairment and a full duration movie, also with one impairment. We can easily convince ourselves that the same distortion will have a much bigger effect on QoE during the short clip than while rating the full movie. This example shows how important the time dimension is with respect to an overall quality evaluation and why it should be further investigated.

The International Telecommunication Union (ITU) provides a number of recommendations for perceived quality evaluation of audio, video and audiovisual material. Only one of those is aimed at long duration audiovisual sequences (up to 30 min) – it is a Single Stimulus Continuous Quality Evaluation (SSCQE) method. This method allows for the continuous evaluation of presented stimuli by using a *"slider that subjects have to move while looking at and /or listening to programmes or scenarios"* [5, 3]. There is no explicit reference given for the comparison and subjects are required to provide a rating for each variation of quality. Using a slider to evaluate the presented material is a precision task which usually requires full concentration from users. Apart from performing the quality assessment, subjects also have to think about positioning the slider which is not easy without looking at the device. These issues may have a direct impact on the assessment of QoE, thus making the method unreliable for real life scenarios.

Existing methods are often very attention demanding. To avoid user fatigue, their duration is limited to 30 min. These methods help us to better understand some of the characteristics of quality perception, but at the same time leave a lot of aspects of QoE, in relation to time, still undiscovered. As subjects can become bored and/or disaffected with a rating process over longer periods of time [1], innovative test procedures are desirable.

A new subjective methodology for long-term quality assessment was recently presented [7]. The proposed method uses full length DVD movies with impairments introduced beforehand. The specially prepared disc is given to subjects together with a questionnaire in a sealed envelope, and their task is to watch the movie in the same environment and under the same conditions as they usually watch television. User feedback is collected through the questionnaire which is filled in immediately after the movie is finished. The subjects provide their overall quality ratings using an ACR scale as well as answer to questions related to perceived degradations. A day after the rating process, a face to face interview is performed.

Although some new insight has been gained using this method, it does not take into account memory effects (for time-varying quality estimation) which seem to be limited to about 15 s [6]. Answering detailed questions about distortion occurring in the beginning of a movie and their annoyance is a difficult task. Moreover, it requires a lot of effort to make the answers ready for further analysis, which is impractical for most real life applications. Finally, the evaluation process is lengthy in nature and hence costly. Summarizing, not too many corresponding research activities can be found on the topic of long-term quality evaluation methods. This significantly slows down progress in the field. New subjective methodologies are desirable to overcome the mentioned problems and to help us better understand users' needs and expectations.

A.3. Proposed Method

In this paper we present a novel method for multi-modal, long-term quality assessment of audiovisual content. The method represents a new approach for understanding the underlying attributes of perceived quality and user behavior. We were particularly interested in a subjective methodology, which could be applicable in a laboratory setup as well as in a real life environment. With respect to limitations of current recommendations and available guidance in this field, answering some of the research questions turns out not to be feasible. To assess the QoE of multi-modal systems we need to think about a methodology which can be used for continuous sessions of long duration, which can work without an explicit quality reference, which limits assessors' fatigue to a minimum, and which allows the subjects to focus on the content instead of directing their attention to the assessment task itself [2]. Taking into account the mentioned requirements we designed a unique experimental setup described in the following. As an outcome of our work we propose a novel subjective assessment method which can be used for quality assessments of audio, video, and audiovisual content.

The method is based on an adjustment of the quality during playback, which is a completely different approach than direct assessment. In most of the existing methodologies, subjects evaluate presented material by giving scores (on existing scales) or by setting a slider to a specific position representing a score value. The method described here allows assessors to adjust the quality to a desired level in case degradations occur, instead of giving a score. This is a purely perceptually

based judgment, not involving any extra processing needed to translate perceived quality into a single number. Giving the power of adjustment to the assessors we can learn more about their expectations and reactions to quality changes over longer periods of time, while possibly requesting less of their attention.

A.3.1. Experimental Design

A number of stimuli with different quality levels are prepared beforehand, for example by using different encoding parameters on a specific encoder. An adjustment instrument (e.g. knob, scroll wheel, etc.) is used for selecting the quality setting of the stimulus. It is important that there are no physical limits in the adjustment mechanism. The device itself should not provide any extra information to the user (no clicks or anything that could influence the perception of quality change, only smooth moves). The only feedback the user can notice should be related to changes seen on the screen or heard from the speakers, i.e., changes in the quality itself. The main concept of how it works is shown in Fig. A.1 (example with a knob). Min and Max on this drawing are only for explanatory purposes and are not marked on the device. The dotted mark indicates the transition point from increasing to decreasing quality, and the solid one shows the lower quality limit which cannot be under-passed from either direction.

Turning the device clockwise increases the quality level of the presented stimuli until the maximum is attained. Rotating it further clockwise starts decreasing the displayed quality. This behavior could be considered a sort of penalty for going 'too far' on the scale. This is a completely reversible process, which means that the assessor can come back to the maximum displayed quality by turning the device counterclockwise. A similar paradigm holds in case of a scroll wheel with up and down directions. The device sensitivity of the device should be set to an appropriate level to avoid too rapid or too slow changes while being turned.

During the assessment, automatic changes in the quality are introduced periodically or at random and stepwise (e.g. the degradation process starts every 3 min and thereafter quality levels decrease every 10 s). The test participant should have enough time to notice the change and to react by rotating the device. Assessor's input will stop any further automatic degradation of the quality introduced by the system. Subsequently, the user adjusts the displayed quality to a satisfying level (maximum desirable).

The next degradation procedure always starts from the quality level which was displayed last.

The method represents an alternative for continuous quality assessment. It has been designed to minimize user fatigue related to the assessment task, and at the same time to provide new information about QoE during long-term procedure. The evaluation process is fully automated and the results are collected in a format which is relatively easy to use for further analysis. Besides the numerical data (time and quality level), also information about types of change (by user or by the system) is written into a log file.

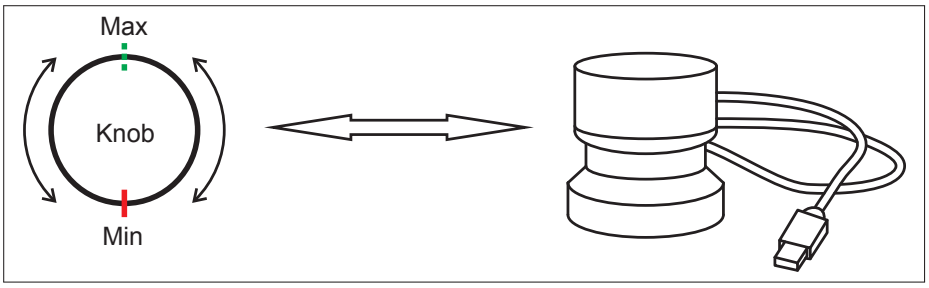


Figure A.1.: Principle of operation of an adjustment device (the knob example).

A.3.2. Test

To evaluate the proposed method an audiovisual experiment was conducted. A software was developed for the automatization of the testing procedure. The general conceptual structure of the experimental setup is presented in Fig. A.2.

The test software used pre-processed raw video files and an uncompressed audio file for playback. Video clips were generated by applying different values of quantization parameter (QP) implemented in H.264/AVC encoder and then by decoding them to raw YUV format. Assessors were using the knob (mentioned in Section A.3.1) to adjust the quality level by seamlessly switching between video files on the fly. The device had no physical limitations with respect to rotation mechanism (lack of start and end point). To switch between adjacent levels, a 90 degree turn was required. Sensitivity of the adjustment instrument was set according to feedback received from subjects during a pilot test, to avoid too rapid or too slow changes. Video and audio were synchronized and only video artifacts were introduced during the playback.

A.3.2.1. Subjective testing procedure

A total of 22 subjects (eight female, 14 male) participated in the subjective experiment, aged between 24 and 61 years. Participants were screened for visual acuity, resulting in two subjects having slightly worse vision (20/30) than average. Subsequently, results provided by these two subjects were excluded from further analysis. Seven out of all participants had worked professionally with HD content before and/or had watched movies in HD quality on a daily basis. They can be regarded as experienced HD quality viewers. The test procedure was divided into three main parts.

The first part was designated to visual acuity test, instructions (written and oral) and training. Assessors were instructed that in case of visible degradation of presented material, their task is to return to reference quality using a knob. The training lasted for 10-15 min. Two short clips were presented to the participants (5 min each), so they had time to familiarize themselves with the setup, find out how

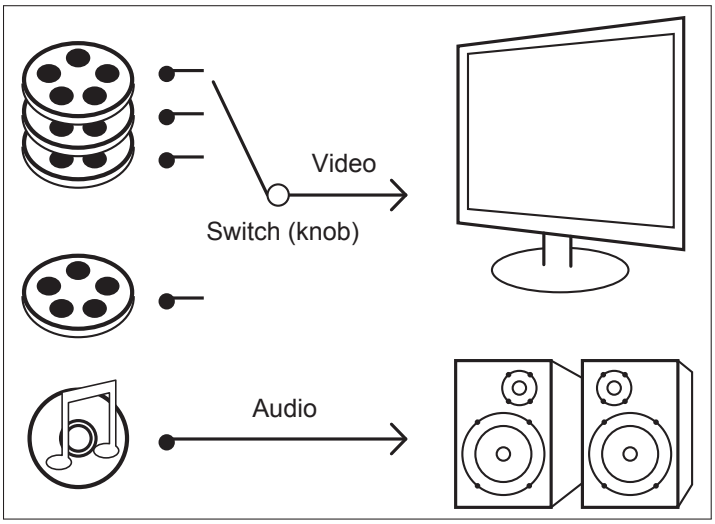


Figure A.2.: Conceptual structure of the experimental setup.

sensitive the device was, see the types of distortions used and how different levels of the quality looked like. Subjects could ask questions during the first part.

In the second part over 30-min long clip was evaluated. For the first 3 min the highest quality was displayed. Before the start, participants were informed that the first quality degradation would occur at 1 to 5 min into the clip. This allowed them to familiarize themselves with the highest quality before the degradation process started. The degradation process was introduced every 3 min with the quality levels further decreasing every 10 s. In the third part, subjects were asked to answer questions regarding the experiment. Here, we tried to get feedback on difficulties the participants had (if any), and positive aspects of the evaluation method. We also were interested in assessors' attention to the task/content as well as in their interest in the presented material.

Video content was displayed on a 50-inch plasma screen (Pioneer PDP-5000EX), and two loudspeakers (Dynaudio BM6A) were used for sound reproduction. The experiment was conducted in a room particularly designed for video and audio quality tests, with appropriate lighting and acoustics. Viewing, listening and lighting conditions were set according to ITU recommendations: BT.500-11 [5] and P.911 [4]. Participants were sitting in comfortable, cinema style chairs with a slide-out tray, which was used for placing the knob on.

A.3.2.2. Content and its preparation

With respect to the duration and nature of the experiment, a meaningful and interesting audiovisual content had to be selected. Due to a lack of databases containing

high definition (HD) content of long duration, freely available, commercial material was employed.

The selected audiovisual clip was taken from the first episode of a nature documentary series titled 'Life' (made by BBC television). High resolution (1920 x 1080 pixels) and high quality (Bluray version) copy was used in the test. The original material was ripped from a Bluray disc onto the hard drive (raw .m2st file). Subsequently, H.264/AVC encoder (x264) was used for further processing. The test material was prepared such that the steps between quality levels were similar with respect to the just noticeable difference (JND) threshold. For this purpose, the QP for each level was set according to [9, 8] and also based on our own investigations in this field. Finally, all video clips were decoded to YUV format (4:2:0). The original audio track was used (stereo, 48 kHz) for audio playback. The experimental conditions are summarized in Table A.1.

Table A.1.: Test conditions

Duration	30 min 55 s
Video resolution	HD 1080p (1920 × 1080 pixels)
Video frame-rate	25 fps
Video color scheme	16-bit YUV 4:2:0
Audio format	WAV, stereo, 48 kHz, 512 kbps
Video format	H.264/AVC
QP values	0,16,22,24,28,30,32,36,38,40,44,48

A.4. Preliminary Results

We expected to find that the sensitivity to quality changes over long periods of time, as well as dependencies related to users' expectations. Fig. A.3 presents average responses of experienced subjects vs. naïve ones. As was mentioned previously, the experiment began with the highest quality level as a reference for a period of 3 min. We can clearly see from presented figures that participants were accepting a lower quality almost from the very beginning of the assessment task, in spite of the fact of having continuous possibilities to adjust it. Returning to the top level turned out to be not an obvious task, even right after the reference was presented.

Inspection of Fig. A.3 shows significant discrepancies in the responses by experienced and by naïve participants. The first group tends to be more demanding with respect to acceptable quality, keeping it on average at a higher level. Naïve users seem to be less sensitive to changes and require more time to return to higher quality levels. One should notice, that in spite of displaying the 'anchor' during the first 3 min, an almost immediate drop in the quality can be observed. It is related to particular scenes (where foreground was much more focused than the background) which for inexperienced subjects seemed to contain artifacts. Another reason for this phenomenon might be a high attention to the task in the early stage of the

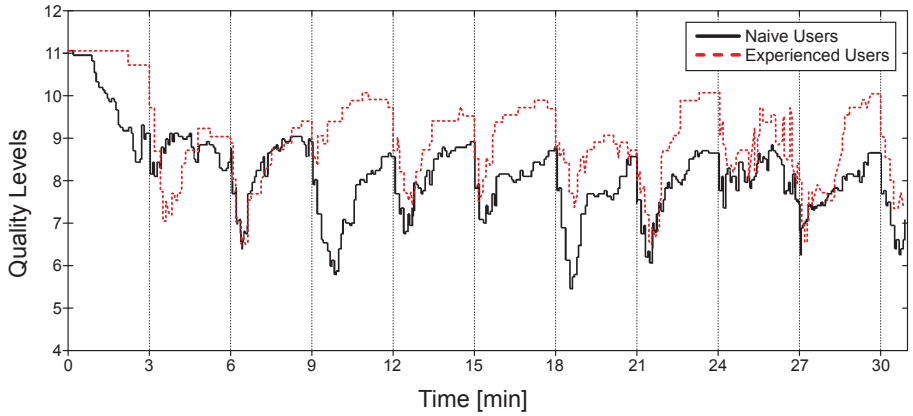


Figure A.3.: Average responses of experienced assessors vs. naïve ones.

experiment. Naïve participants were highly focused on the detection of impairments during the first several minutes, but after this period their attention to the task significantly decreased. This finding is based on answers given by participants during the last part of the test procedure.

Fig. A.3 suggests that the average participant was very consistent in his/her choices after a warm-up phase of around 5 min. Visible variations in the responses are most probably related to the content itself and user individual interest in the content. This need to be further investigated and particular scenes have to be analyzed in detail.

A.5. Conclusions

In this paper a novel subjective methodology for the quality assessment of long duration content has been proposed. The method is based on quality adjustment during playback, which represents a different approach towards evaluation of QoE. The method allows continuous assessment of real life applications (e.g. movies, TV programmes, video-conferencing, etc.). The results of such a setup can provide a direct answer to the question: which quality level is acceptable for a potential user? This fact might be of interest for content/service providers who often strive to meet end-user expectations. The described methodology requires low-attention from assessors (according to feedback received from participants right after the experiment), making the evaluation process easy to understand, pleasurable for the participants, and most important – accurate. Assessors do not have to translate perceived quality into a numerical score or an equivalent semantic designator. Having the possibility to adjust the quality instead of judging it on existing scales allows assessors to focus on the content itself and not on the artefacts and degradations, just like in a real-life application scenario.

The experiment conducted for verification of the new method shows its suitability for long-term quality evaluations and provides us with some interesting findings. However, it is not utterly clear how big influence on obtained results and on the evaluation process has the chosen content. Therefore other types of content have to be employed to the next experiments. Analysis of collected data is still in progress and more detailed results will be presented in a follow-up article. The method is still under development and its usability in other modal contexts as well as constraints and possible improvements will be studied further. Moreover, comparison of the new method with the standardized subjective quality assessment methodology will be investigated in the nearest future. The long-term goal for this research is to demonstrate usefulness of presented methodology in different contexts and for different applications. We are going to support the idea of quality adjustment by better understanding its benefits and limitations in a process of audiovisual quality assessment.

Acknowledgment

This work was performed within the PERCEVAL project, funded by The Research Council of Norway under project number 193034/S10.

References

[1] R. P. Aldridge, D. S. Hands, D. E. Pearson, and N. K. Lodge. Continuous quality assessment of digitally-coded television pictures. In *Proc. of IEEE Vision Image Signal Processing*, volume 145, pages 116–123, 1998.

[2] A. Eichhorn, U. Reiter, O. Tomic, and K. I. Frostervold. PERCEVAL - Perceptual and cognitive quality evaluation techniques for audiovisual systems. 2009. Project proposal to the Research Council of Norway.

[3] R. Hamberg and H. de Ridder. Continuous assessment of perceptual image quality. *Journal of the Optical Society of America*, 12(12):2573–2577, 1995.

[4] ITU-T Rec. BT.500-12. Methodology for the subjective assessment of the quality of television pictures. Int. Telecomm. Union, Geneva, 2009.

[5] ITU-T Rec. P.911. Subjective audiovisual quality assessment methods for multimedia applications. Int. Telecomm. Union, Geneva, 1998.

[6] M. H. Pinson and S. Wolf. Comparing subjective video quality testing methodologies. In *Visual Communications and Image Processing. Proceedings of the SPIE*, volume 5150, pages 573–582, 2003.

[7] N. Staelens, S. Moens, W. Van den Broeck, I. Marien, B. Vermuelen, P. Lambert, R. Van de Walle, and P. Demeester. Assessing quality of experience of IPTV and video on demand services in real-life environments. *IEEE Transactions on Broadcasting*, 56(4):458–466, 2010.

[8] H. Wang, X. Qian, and G. Liu. Inter mode decision based on just noticeable difference profile. In *Proc. of 2010 IEEE 17th International Conference on Image Processing*, pages 297–300, Hong Kong, 2010.

[9] X. Yang, Y. Tan, and N. Ling. Rate control for H.264 with two-step quantization parameter determination but single-pass encoding. *EURASIP Journal on Applied Signal Processing*, 2006:1–13, 2006.

B. Quality Evaluation of Long Duration AV Content – An Extended Analysis using a Novel Assessment Methodology

Paper B

Adam Borowiak[1], Ulrich Reiter[1]

[1] *Department of Electronics and Telecommunications, Norwegian University of Science and Technology (NTNU)*

Multimedia Tools and Applications Journal, volume 74(2), pp. 367-380, 2015

Abstract

This paper is an extension of our previous work describing a novel methodology for quality assessment of long duration audiovisual content. In this article we focus on data analysis of results obtained from two experiments conducted using the new methodology. In the first study, we found that the time dimension does not influence participants' expectations with respect to perceived video quality and that a possible increase or decrease in acceptable quality level is rather directly related to the presented material itself. Moreover, we found that participants are less sensitive to quality changes when the process is controlled externally than when they are in charge of the quality adjustment. A second experiment (Study 2) was performed to evaluate the effect of simultaneous quality changes in the two modalities (audio and video) which confirmed the previous results.

B.1. Introduction

Quality assessment of multi-modal systems is a subjective task involving highly complex interactions in the human brain. There have been many attempts of modeling such processes for the purpose of objective metrics' development - so far without any major success. Therefore, evaluations with human assessors are still considered to be the most accurate methods providing fundamental data, which is then further used for objective metrics design and their validation. For human beings, the quality evaluation of an audiovisual content, at first glance, seems to be a relatively straightforward process. We can quickly decide whether the level of perceived quality is acceptable or not [15]. From our own experience, we know that when it comes to the translation of our impression into a numerical score or semantic designator, we tend to have much more difficulties. It seems like interpretation of the magnitude of an audiovisual stimulus on descriptive or numerical rating scales is still one of the fundamental perceptual problems.

Unfortunately, most of the existing standardized methods for the quality evaluation require such interpretations from assessors. In the majority of the quality assessment techniques (which are usually based on a mean opinion score (MOS) approach), participants are asked to quantify a sensory experience using a single number (or word) present on an ordinal or interval scale. This is a difficult task and always burdened by errors related to such a simplified description of subjects' experience [17, 20]. In order to eliminate the aforementioned interpretations, methods based on direct comparisons of stimuli (e.g. one clip played after another or both clips played simultaneously side by side) [2, 12] can be used. However, these methods are suitable for short test-material only according to the memory effect (for time-varying quality estimation) which seems to be limited to about 15 s [13]. Therefore, such methods are not appropriate for the quality evaluation of longer duration content. Time-varying and scene-dependent effects of impairments cannot be assessed this way. In general, quality evaluation of long duration content has not received much attention from the researchers in the field. Moreover, only one standardized method suitable for this purpose exists. This is the Single Stimulus Continuous Quality Evaluation (SSCQE) method which is based on the use of a slider to continuously rate the perceived quality [12, 9]. Although the SSCQE allows assessing the quality of longer duration content (up to 30 min) it is not free from disadvantages and ambiguities. The method demands constant concentration and attention from the assessors [18] which makes the evaluation process too involving for real life applications. Moreover, a fatigue related to continuous operation of the slider mechanism and possible loss of absolute slider position on the scale might lead to problems with reliability of the derived results [13]. To overcome limitations of the SSCQE, the development of new methodologies adequate for quality assessment of long duration content is required.

Recently we proposed a novel methodology designed for quality evaluation of audio, video and audiovisual material of long duration. The method represents a novel approach towards evaluation of Quality of Experience (QoE) in mono-

and multi-modal systems. Initial results obtained with this method, including the detailed description of the experimental design and test procedures, have been described previously in [5]. In this paper we want to focus on the data analysis and interpretation of the results as well as discuss what additional insights into the quality perception this method brings. To strengthen the obtained results and explore suitability of the novel method in a mixed-media context with both modalities being impaired, an additional experiment was conducted and analyzed.

The remainder of this paper is organized as follows: in Section B.2, operational principles of the new assessment method are presented. In Section B.3, the experimental setup and procedures are discussed. Section B.4 consists of a detailed description of the results of two experiments and methods for the data interpretation. Finally, conclusions are drawn and future work is discussed.

B.2. Method Description

Instead of giving numerical scores on predefined scales, the method proposed by us in [5] allows participants to adjust the quality to a desired level in case degradation occurs. As this is a purely perceptually based judgment, there is no need for assessors to translate their impressions into a numerical value or semantic designator, thus avoiding most of the problems typical for MOS related approaches, e.g. scale usage heterogeneity [15], nonlinearity of the scales [17], etc. (for more examples see [6, 17]). To select the most appreciated quality level a rotary controller device is used (e.g. knob, scroll wheel, etc.). It is crucial that the device mechanism has no physical limitations (such as restricted amount of rotary movement) and does not provide any tactile feedback to the user (such as clicks during the quality change). Quality adjustments should be based on perceptual judgment of the presented stimulus only. The system automatically degrades the quality in equal time intervals (or at random) and stepwise in time (or immediately in one step) during the evaluation task. In this study for example, the degradation procedure occurs at the beginning of each 3 min time interval, with quality levels further decreasing every 10 s. Whenever the participant acts on the device, the quality will no longer be degraded further. Rotating the knob clockwise increases the quality level of the presented stimulus up to the point where the highest quality is achieved. Turning the device further clockwise causes a gradual decrease in quality. This is a sort of penalty introduced when the maximum quality level is crossed. The process is reversible and by rotating the knob in the opposite (counterclockwise) direction the assessor can return to the highest quality level. The quality levels are prepared beforehand (or on the fly, e.g. using scalable video coding - SVC), and each of them has assigned a numerical value (e.g. lowest quality: 0, highest: 11). Acquisition of the user's responses is performed automatically, at a time resolution of one millisecond and with values which lie within the range of the corresponding quality levels. Besides the numerical data (time and values corresponding to quality levels), also information about the source of the quality change (system or user) is gathered. A more detailed description of the method can be found in [5]. The proposed methodology

can be regarded as a sort of extension of the *Method of Adjustment* (also called the method of minimal changes) originally developed by Fechner [7]. The main purpose of Fechner's method is to determine a threshold which represents the level of the stimulus property at which it becomes detectable. It is achieved by gradually increasing (decreasing) the intensity of the studied property in discrete steps until it is detectable (not detectable). In our case, quality represents the property which is gradually decreased (by the system, in order to find the level at which the assessor notices the change in the quality) and increased (by participants - in order to find the maximum level of quality). The aim is to find the acceptability threshold over time with respect to the quality of the displayed material. This is based on direct quality levels discrimination and not as in MOS-related methods, on yes-no or ordinal judgments.

B.3. Details of Subjective Study

An audiovisual clip extracted from the first episode of a nature documentary series titled *Life* (produced by BBC television) was used in the first study. With respect to the experimental design the original material was cut to 30 min and 55 s while still preserving a semantic structure. The video contained a variety of camera angles and shots, including ultra-slow motion captures as well as action scenes with a lot of movement, details and close-ups. In order to characterize the stimulus in the visual domain, the amount of spatial and temporal information (SI and TI accordingly) was calculated. The SI generally denotes the amount of spatial details in the picture. It is computed as a standard deviation over the pixels in each Sobel-filtered frame (luminance plane) [11]. More spatially complex scenes generate higher SI values. The TI provides information about the amount of temporal changes in the video, and it is computed as the standard deviation of the difference between pixel values in successive frames (luminance plane). Higher TI values denote more motion in adjacent frames [11]. The SI and TI values averaged over 4500 frames (this equals to 3 min and was chosen to match the interval of the degradation periods) are presented in Fig. B.1.

High quality material (1080p version) extracted from a Bluray disc served as a starting point for further processing using an H.264/AVC encoder (x264). In total, 12 different quality levels were created and variations in-between the levels were achieved by using different settings of quantization parameter (QP). More specifically, the Constant Quantizer mode (Single Pass) in x264 encoder was used for this purpose. The steps between such prepared quality levels were similar with respect to the threshold of just noticeable differences (JND). The values of QP were set according to [16, 19] and also based on our own investigations. The highest quality was achieved by setting the QP in H.264/AVC encoder to 0. To produce the lowest quality used in this experiment, QP was set to 48. Subsequently, all video clips were decoded to YUV format (4:2:0). The original soundtrack containing speech, background music as well as nature sounds were used for audio playback. For the purpose of the second study, a 30 min and 20 s long extract from the ninth

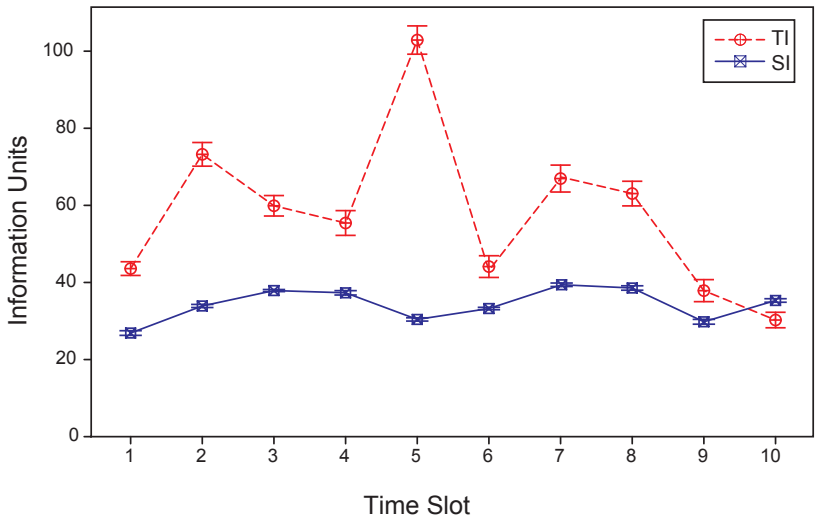


Figure B.1.: Spatial and temporal information of the test clip used in study 1. Error bars show 95% confidence interval.

episode of the same series was selected. Spatial and temporal characteristics of the video can be seen in Fig. B.2.

Visual material was prepared in a similar way as in the first study, but with the number of quality levels reduced to 11 (lowest one was removed). The associated audio stream (2ch, 1411 kbps, 16 bit/44.1 kHz) was encoded as MP3 with LAME encoder [1] at 11 different bit-rates (the lowest at 32 kbps and the highest at 320 kbps) with VBR (variable bit-rate) off. To preserve loudness consistency among all audio quality levels the volume level was normalized (unweighted level). The summary of test conditions for both studies is shown in Table B.1.

Table B.1.: Test conditions of study 1 and study 2.

	Study 1	Study 2
Video color scheme	16-bit YUV 4:2:0	16-bit YUV 4:2:0
Video properties	HD 1080p (1920 × 1080), 25fps	HD 1080p (1920 × 1080), 25fps
Audio properties	2ch, 16-bit/44.1 kHz	2ch, 16-bit/44.1 kHz
Audio compression rates (kbps)	512	32, 48, 56, 64, 80, 96, 112, 128, 160, 192 and 320
QP values	0, 16, 22, 24, 28, 30, 32, 36, 38, 40, 44, 48	0, 16, 22, 24, 28, 30, 32, 36, 38, 40,44

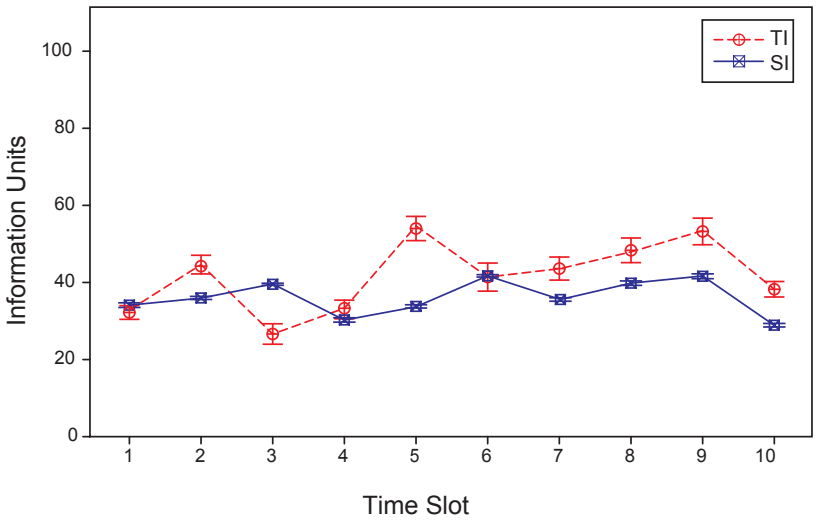


Figure B.2.: Spatial and temporal information of the test clip used in study 2. Error bars show 95% confidence interval.

B.3.1. Participants

Twenty-two subjects (eight female and 14 male, mixture of students and employees from the Norwegian University of Science and Technology in Trondheim / Norway) took part in the first experiment. The mean age was 33 years (age range 24 – 61 years). Participants were screened for visual acuity using a Snellen chart, resulting in 2 subjects having slightly lower visual acuity (20/30) than the others. Consequently, results provided by these two subjects were removed from data analysis. Seven participants had worked with HD content on a professional level and/or had watched movies in HD quality on a regular basis. They can be regarded as experienced viewers with respect to HD audiovisual material. In the second study 20 subjects, seven females and 13 males were employed. The mean age was 31 years (age range 22 – 43 years). Thirteen participants also took part in the first study. All subjects had normal or corrected to normal visual acuity (according to the Snellen's test) and hearing (according to their self-report).

B.3.2. Test Procedure

The test procedure was identical for both studies and consisted of pre-test, test and post-test session. In the pre-test session, each subject received instructions (written and oral) and watched two 5-min long training clips. The purpose of the training was to familiarize participants with the experimental setup, types of distortions, and with the sensitivity of the adjustment instrument. Assessors were allowed to ask questions at any time of the pre-test session. The training clips were selected

to span the same quality range as the test clip. In the second part, participants evaluated the quality of an over 30-min long audiovisual material. To familiarize subjects with the highest quality of the presented stimuli, the maximum quality was presented for the first 3 min. Participants were told that the first deterioration of the quality would take place between the 1st and the 5th minute of the clip. In the first study, participants were asked to react to video only quality changes, whereas in the second experiment subjects had to respond to both video and audio quality degradations (impairments were introduced simultaneously in both domains). In the post-test session, qualitative data was gathered. Participants were asked questions regarding difficulties (if any), and positive/negative aspects of the evaluation method. We were also interested in assessors' attention to the task/content as well as in their interest in the presented material. Additionally, for the second experiment we wanted to get to know how participants discovered the quality drop (by hearing or by seeing it) and how (if at all) quality improvement of one modality was helping to improve the quality of another one. We were also interested to see which modality (audio or video) was provoking the decisions during the process of final quality adjustments. The clips were presented on a 50-inch full HD, Pioneer PDP-5000EX plasma screen. Two professional grade active monitor loudspeakers, Dynaudio BM6A, were used for sound reproduction. Both experiments took place in a room suitable for audio and video quality assessment tests, with appropriate lighting and room acoustic treatment. Viewing, listening and lighting conditions were set conforming to ITU recommendations P.911 [12] and BT.500-12 [10]. The test-environment simulated a casual home cinema setup (with respect to type of furniture and placement) rather than rigorous lab conditions, under which subjects are mainly concentrated to perform the assessment task.

B.4. Results

B.4.1. First Study

We were particularly interested in investigating relations between the three-min time slots (from one to another automatic degradation procedure) and users' responses to quality changes. The objective of the first study was to find answers to the following questions:

- Do the quality expectations decrease over time and with increased involvement in the content?
- How fast can participants notice quality changes and at which quality level does this happen?
- Is the quality level at which the change is noticed similar to the quality level set by the participant?

To proceed with data analysis which would allow us to investigate above inquiries we created three subsets of data. These subsets consist of:

1. quality levels averaged over last minute of each 3 min time section (AQL)
2. response time to quality changes introduced by the system (RT) after the start of each 3 min time slot
3. quality levels at the time when a user reacted to a quality change (QLRT)

Each of these data subsets had the same size 20 x 9, where 20 corresponds to the number of participants and nine to the number of 3–min time periods. The first time slot was excluded from data analysis as for the first 3 min the reference quality was displayed and no user reaction was expected during this time period. For the subset 1) we decided to focus on the last 60 s of each time section as it should be the period where users have established their preferences with respect to the perceived quality. To reduce unwanted noise the results provided by assessors during these periods were averaged. Analysis of Variance (ANOVA) was used as a tool for extraction of information from this data. ANOVA is a well established statistical method that compares the deviation between means of several groups to the random deviation within groups [3]. With respect to the assumptions required to use ANOVA the data was checked for normality and homogeneity of variance across assessors as well as across time slots. All data sets showed normal distribution (Kolmogorov-Smirnov test; $p>0.05$) and homogeneous variance (Levene's Test; $p>0.05$). A mixed-effects model of ANOVA was applied to each of the mentioned subsets of data with participants representing random-effect type factor and time slots representing fixed-effect type factor. Using such a model we can better understand which factor is responsible for most of the variation in the data and also compare the main effect of each of them. The ANOVA test used for the following analysis was conducted at the 0.05 significance level.

Results for the subset data 1) are presented in Table B.2.

Table B.2.: Results of ANOVA for the first data subset of study 1.

Source	Df	Sum of Squares	Mean Square	F	p
User	19	125,66	6,614	4,19	0,000
Time Slot	8	11,479	1,435	0,91	0,511
Eror	152	240,154	1,580		
Total	179	377,299			

We can see that the mean quality levels for time slots are not significantly different from each other ($F = 0.91$; $p>0.05$), whereas the mean quality levels across users greatly differ in the statistical sense ($F = 4.19$; $p<0.05$). The results indicate that indeed there were no significant changes in the quality expectations over the entire length of the clip. Assessors were quite consistent in their choices during the whole period of time and variation across time slots (see SS of time slots in ANOVA table) may be due to presented content/the particular scene. On the other hand we can conclude that there were big variations in means of quality levels across participants, which is presumably due to differences between the ways assessors

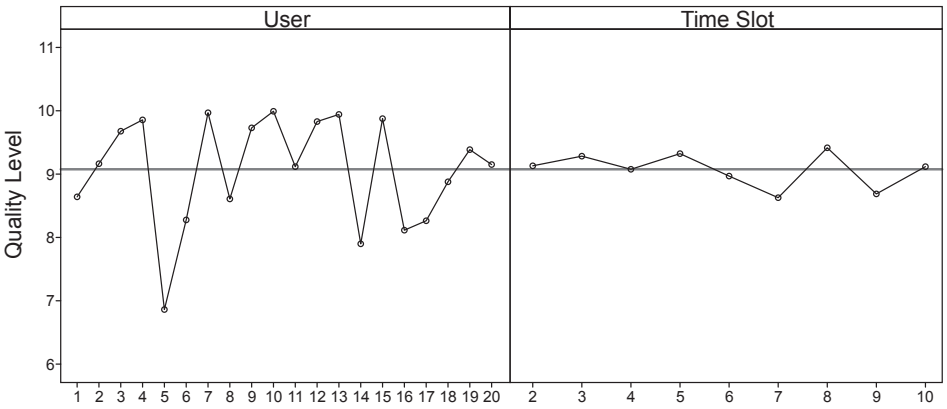


Figure B.3.: Main effects plot for quality levels in study 1, averaged over last minutes of each 3 min time section (AQL).

were using the adjustment device to improve the quality. From Fig. B.3 we can observe that variation due to user differences is much larger than variation due to time slot differences. The mean values for time slots are relatively alike to each other. Similar results were obtained for the data subset 2).

Table B.3 shows that there were no significant differences in reaction time to quality changes across all time sections (F = 0.69; p>0.05). The average difference between the start of each degradation process and the time at which users noticed a change in the quality was around 26 s. This corresponds to a 3 levels drop in the quality before assessors reacted. From Table B.3 it can also be seen that mean reaction times for users are significantly different (F = 2.02; p<0.05).

Table B.3.: Results of ANOVA for the second data subset of study 1.

Source	Df	Sum of Squares	Mean Square	F	p
User	19	13476,1	709,3	2,02	0,010
Time Slot	8	1928,5	241,1	0,69	0,703
Eror	152	53378,1	351,2		
Total	179	68782,7			

Looking at Fig. B.4 we can see more clearly how means for reaction times across users as well as across time slots were distributed.

Table B.4 shows again that there is no significant difference between means for time sections in subset data 3) (F = 1.60; p>0.05), whereas we can see statistical significance with respect to means comparison across the users (F = 5.00; p<0.05). One could notice that this time the p value in the first case is relatively low compared to subset 1). This is due to high values in one particular time slot (see Fig. B.5).

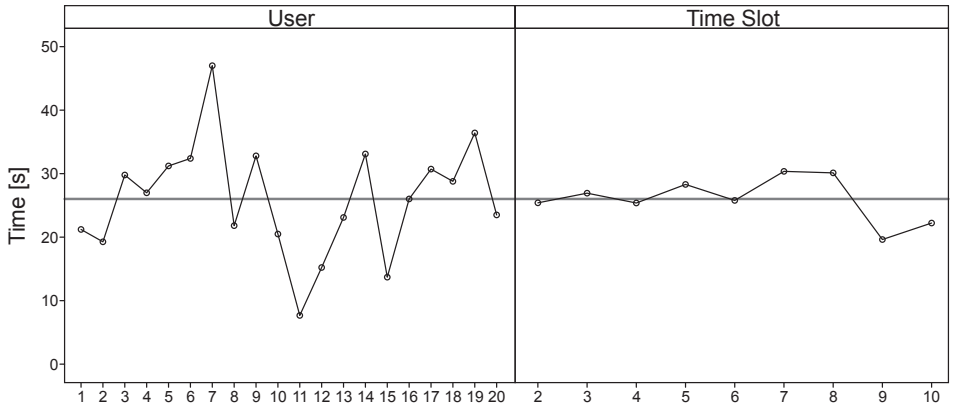


Figure B.4.: Main effects plot for reaction time (RT) in study 1.

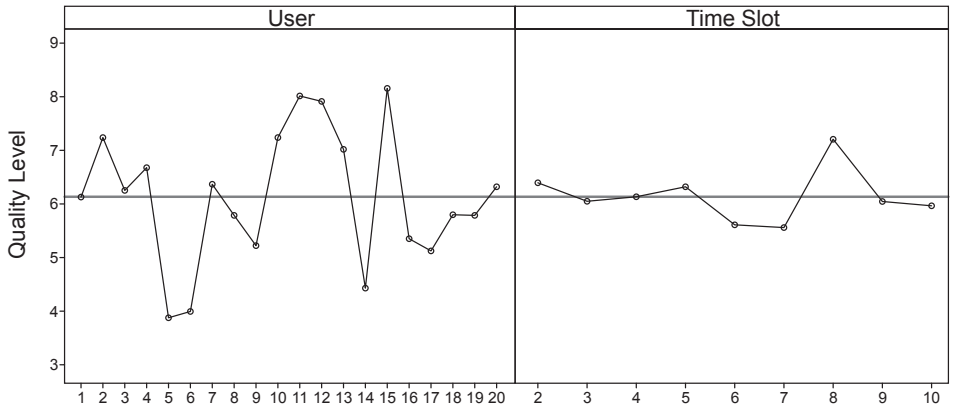


Figure B.5.: Main effects plot for quality levels in study 1, corresponding to reaction time (QLRT).

Table B.4.: Results of ANOVA for the third data subset of study 1.

Source	Df	Sum of Squares	Mean Square	F	p
User	19	263,756	13,882	5,00	0,000
Time Slot	8	35,611	4,451	1,60	0,128
Eror	152	421,944	2,776		
Total	179	721,311			

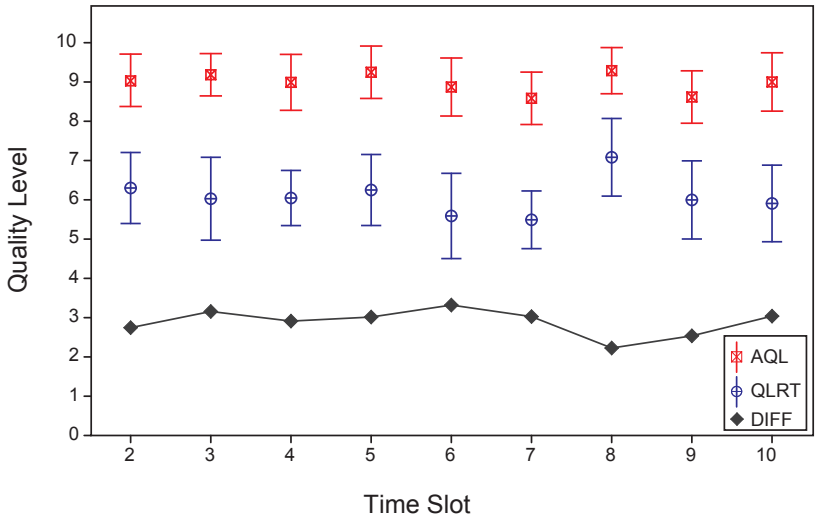


Figure B.6.: Comparison of subjects' sensitivity to audio quality changes under different conditions (AQL vs. QLRT) in study 1. Error bars show 95% confidence interval.

Fig. B.6 shows the comparison between AQL and QLRT. The bottom plot (DIFF) represents the actual differences between them. It can be seen that these differences are alike for each time period and in average correspond to three quality levels. This clearly suggests that it is easier for assessors to distinguish between neighboring quality levels when they perform control over the displayed quality themselves (e.g. during the adjustment procedure) than when the process is independent from them and happens at random. We could also conclude that quality level set by assessors towards the end of each time slot is not necessarily the one which represents the acceptable quality level for most of the participants. The averaged acceptable quality level is rather related to the one corresponding to reaction times.

In addition to the above considerations, the experience and gender influence on the obtained results was investigated. The difference between results provided by experienced users and the naïve ones was found to be statistically significant. This holds for the case when AQL's for both groups are compared (ANOVA; F = 0.90;

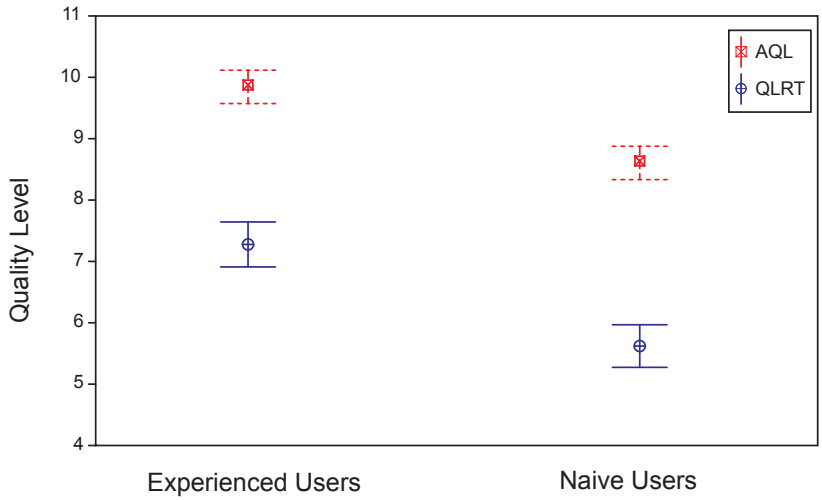


Figure B.7.: Main effects plot for comparison between naïve and experienced users with respect to AQL and QLRT in study 1. Error bars show 95% confidence interval.

$p<0.05$) as well as when QLRT's are considered (ANOVA; F = 1.15; $p<0.05$). The results of such comparisons can be seen in Fig. B.7.

A similar procedure was employed to investigate the possible gender effect. It turned out that gender does not play a significant role in such a setup, no matter whether we consider AQL (ANOVA; F = 1.15; $p>0.05$) or QLRT (ANOVA; F = 0.99; $p>0.05$).

B.4.2. Second Study

The purpose of the second study was to check the suitability of our method in case when both modalities are impaired at the same time, as well as to strengthen the results from the first study. This time, the assumption of the data normality was not met (Kolmogorov-Smirnov test; $p<0.05$) which resulted in a non-parametric method being used for statistical considerations. The Kruskal-Wallis (K-W) test (equivalent of one way ANOVA) for independent group comparisons was employed for further data analysis. The data was pre-processed in exactly the same way as it was done for the previous experiment.

The results for subset data 1) (see Fig. B.8) again show that for a similar content type the time dimension does not play a major role in the process of quality reconstruction (K-W; H(8) = 13.86, $p>0.05$).

Participants' expectations do not change over time, but might vary between particular time periods as a result of different stimulus properties (e.g. amount of motion and details in the video). Similar results were obtained for subset data 2

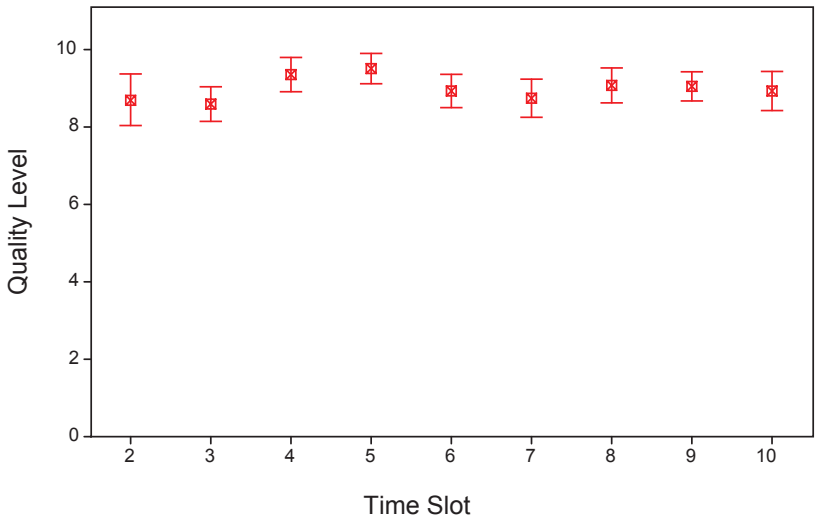


Figure B.8.: Main effects plot for quality levels in study 2, averaged over the last minute of each 3 min time section. Error bars show 95% confidence interval.

(K-W; H(8) = 14.33, p>0.05) and 3 (K-W; H(8) = 15.45, p>0.05). Once more, it turned out that the quality level at which the impairments are noticed (QLRT) by the test-subjects remains rather constant across all the time intervals (see Fig. B.9). An attentive reader may also notice that the mean difference between AQL (see Fig. B.8) and QLRT (see Fig. B.9) is similar as described in study 1 (see Fig. B.6). This finding confirms that the awareness of the process of change in the quality may have a significant influence on human expectations.

Those participants who took part in both study 1 and 2 reported that degradations introduced in both modalities simultaneously make the process of quality adjustment easier for them. This seems to be confirmed by comparison of the results for subset data 1) for both studies. Such a comparison makes sense in this case as both clips are of similar content type and fall into the same category with respect to SI and TI (according to ITU-T P.910 guidance on this topic [11]), even with the clip from the first study having a higher amount of temporal information (compare Fig. B.1 and Fig. B.2). The above phenomenon has also been confirmed in one of our later studies, in which for the same clip different impairment conditions were compared [4].

According to the feedback received from participants, the visual modality, in general, is the one driving the process of quality improvement. The auditory modality might be helpful in some cases (3 subjects reported that it is the other way around). This might be due to the types of impairment being used for both modalities as well as due to the content type – previous studies have suggested that the dominance of one modality over the other is directly related to content type and features [14]. However, this may also be related to participants' cognitive styles [8].

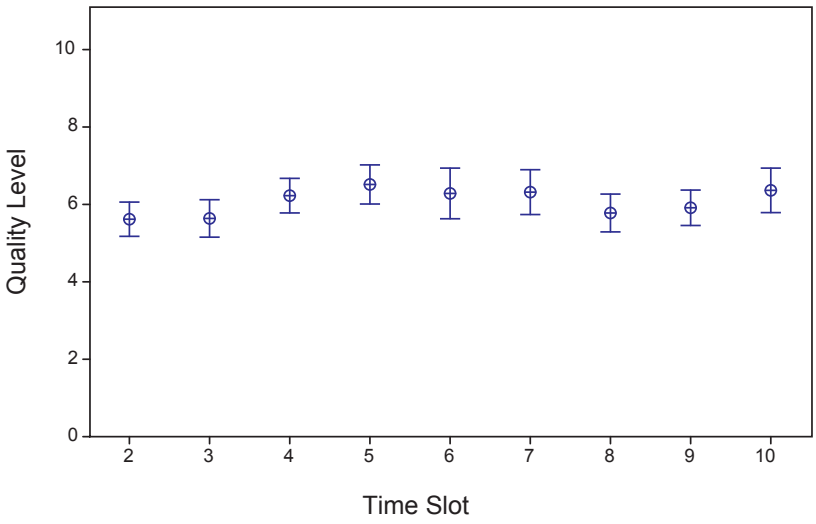


Figure B.9.: Main effects plot for quality levels in study 2, corresponding to reaction time (QLRT). Error bars show 95% confidence interval.

In order to help the reader better understand subjects responses to the stimuli over time, an average time history of the data set 2 (all participants included) against a single user response is presented in Fig. B.10. Analogous plot showing an average time history for data set 1 can be found in [5].

B.5. Conclusions and Future Work

We carried out two experiments in which the usability of a novel subjective assessment method for quality evaluation of long duration content was proven in two different scenarios: a) video quality evaluation (with accompanying audio), b) audio and video quality evaluation. The paper defines ways of processing users' responses collected with the help of our methodology and provides a statistical guidance of how to analyze the data.

The results obtained from the first study confirm some of our previous conclusions and thoughts, and also deliver new findings. We discovered that quality expectations over extended periods of time are rather constant and that the same holds for the reaction time to quality changes. These results suggest that the time dimension is not necessarily a factor influencing participants' quality expectations. Therefore, the reason behind fluctuations in the quality perception might be directly related to the test material itself and/or personal involvement in the content. In general, subjects reported great interest in the presented stimuli which might have an impact on the obtained results, but this needs to be verified. More interestingly, the data analysis showed that participants are less sensitive to quality changes when the

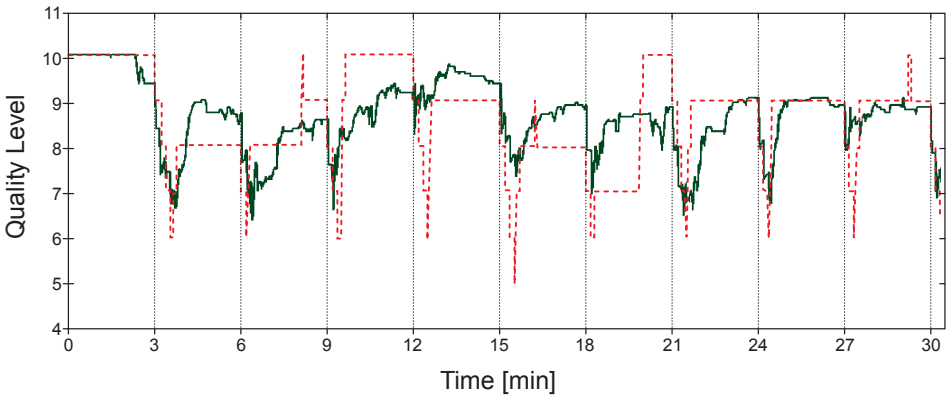


Figure B.10.: Results showing averaged responses of all participants (*continuous line*) against an individual user response (*dotted line*) in study 2.

process is controlled externally than when they have the possibility to adjust the quality themselves. We also showed that participants categorized as experienced HD quality viewers are able to set the quality to higher levels, as well as notice the quality change earlier than the less experienced users. It seems that frequent interactions with the HD material increase our sensitivity to distortion and make us more demanding with respect to the quality of an audiovisual content.

Results of the second study strengthen the above conclusions and also reveal interesting findings for scenarios in which both modalities are degraded simultaneously. We have found that in such cases the process of quality adjustment is easier for participants (less variation in the data) and that the quality levels set by them are high on average. Based on the above we can clearly see that our novel methodology can produce results which contribute to a better understanding of assessor's behavior regarding quality selection, expectations and reactions to quality changes over extended periods of time. Furthermore, it can improve our knowledge about QoE, providing results which cannot be obtained using popular MOS based approaches. The method gives us a direct answer, of which quality level is acceptable to a subject at a given point in time. The above is attained without a need for subjects to translate the perceived quality into a numerical score or an equivalent semantic designator. It is clearly an advantage when compared with the conventional evaluation techniques, wherein such translation is essential and where an acceptability threshold is hard to be determined (e.g. 'fair' or '3' on the rating scale does not tell us whether the quality is acceptable or not).

The aim of our future work is to investigate usability of the method in a context where audio only content is used. The influence of content representing different types of audio/video properties (and of even longer duration) on users' performance and results will be investigated in detail. Other types of artifacts, e.g. quality degradations caused by packet loss, might be considered at a later time.

Acknowledgment

This work was performed within the PERCEVAL project, funded by The Research Council of Norway under project number 193034/S10.

References

[1] LAME (Lame Ain't an MP3 Encoder). The Hydrogenaudio recommended MP3 encoder, http://lame.sourceforge.net.

[2] T. Alpert and L. Contin. DSCQE (Double Stimulus using a Continuous Quality Evaluation) experiment for the evaluation of the MPEG-4 VM on error robustness functionality. In *ISO/IEC – JTC1/SC29/WG11, MPEG 97/M1604*, 1997.

[3] S. Bech and N. Zacharov. *Perceptual audio evaluation – theory, method and application.* John Wiley & Sons, 2006.

[4] A. Borowiak and U. Reiter. Long duration audiovisual content: Impact of content type and impairment appearance on user quality expectations over time. In *Proc. of the 5th International Workshop on Quality of Multimedia Experience (QoMEX 13)*, pages 200–205, Klagenfurt, 2013.

[5] A. Borowiak, U. Reiter, and U. P. Svensson. Quality evaluation of long duration audiovisual content. In *Proc. of the 9th Annual IEEE Consumer Communications and Networking Conference. Special Session on Quality of Experience (QoE) for Multimedia Communications*, pages 353–357, Las Vegas, 2012.

[6] K. T. Chen, C. C. Wu, Y. C. Chang, and C. L. Lei. A crowdsourceable QoE evaluation framework for multimedia content. In *Proc. of the 17th ACM international conference on Multimedia*, pages 491–500, Beijing, 2009.

[7] G. T. Fechner. Elements of psychophysics. volume 1, New York, 1966. Original work published in 1860. Translated by H. E. Adler Holt, Rinehart and Winston.

[8] G. Ghinea and S. Y. Chen. The impact of cognitive styles on perceptual distributed multimedia quality. *British Journal of Educational Technology*, 34(4):393–406, 2003.

[9] R. Hamberg and H. de Ridder. Time-varying image quality: Modeling the relation between instantaneous and overall quality. *SMPTE Motion Image Journal*, 108(11):802—811, 1999.

[10] ITU-T Rec. BT.500-12. Methodology for the subjective assessment of the quality of television pictures. Int. Telecomm. Union, Geneva, 2009.

[11] ITU-T Rec. P.910. Subjective video quality assessment methods for multimedia applications. Int. Telecomm. Union, Geneva, 2008.

[12] ITU-T Rec. P.911. Subjective audiovisual quality assessment methods for multimedia applications. Int. Telecomm. Union, Geneva, 1998.

[13] M. H. Pinson and S. Wolf. Comparing subjective video quality testing methodologies. In *Visual Communications and Image Processing. Proceedings of the SPIE*, volume 5150, pages 573–582, 2003.

[14] U. Reiter. Towards a classification of audiovisual media content. In *Proc. of the 129th Convention of the Audio Engineering Society*, 2010.

[15] P. Rossi, Z. Gilula, and G. Allenby. Overcoming scale usage heterogeneity: A bayesian hierarchical approach. *Journal of the American Statistical Associa-tion*, 96(453):20–31, 2001.

[16] H. Wang, X. Qian, and G. Liu. Inter mode decision based on just noticeable difference profile. In *Proc. of 2010 IEEE 17th International Conference on Image Processing*, pages 297–300, Hong Kong, 2010.

[17] A. Watson and M. A. Sasse. Measuring perceived quality of speech and video in multimedia conferencing applications. In *Proc. of ACM Multimedia*, pages 55–60, 1998.

[18] S. Winkler and F. Dufaux. Video quality evaluation for mobile applications. In *Visual Communications and Image Processing. Proc. of the SPIE*, volume 5150, pages 593—603, 2003.

[19] X. Yang, Y. Tan, and N. Ling. Rate control for H.264 with two-step quantization parameter determination but single-pass encoding. *EURASIP Journal on Applied Signal Processing*, 2006:1–13, 2006.

[20] R. B. Zajonc. Feeling and thinking: Preferences need no inferences. *American Psychologist*, 35(2):151–176, 1980.

C. Audio Quality Requirements and Comparison of Multimodal vs. Unimodal Perception of Impairments for Long Duration Content

Adam Borowiak[1], Ulrich Reiter[1], U. Peter Svensson[1]

[1] *Department of Electronics and Telecommunications, Norwegian University of Science and Technology (NTNU)*

Paper C

Journal of Signal Processing Systems, volume 74(1), pp. 79-89, 2013

Abstract

Using our novel methodology for quality evaluation of long duration multimedia content, the effect of the time dimension on quality ratings and user responses is investigated. Particularly, the influence of audio artifacts related to different compression rates on participants' reactions to quality changes over extended periods of time is examined. Results of the first study suggest that participants' quality expectations are rather constant throughout the entire duration of the 30-min long clip, which also holds for subjects' reaction time to quality degradations. Furthermore, it turns out that the test persons are more sensitive to quality changes when they are able to influence the quality themselves. In addition to the first study, two experiments were conducted in which cross-modal effects between the visual and auditory modality were investigated. The findings indicate that it is significantly easier for participants to discover quality changes when impairments are introduced in both the auditory and visual modality at the same time than when distortions occur in the audio or video domain solely.

C.1. Introduction

Traditional techniques for quality assessment are mainly designed for short clips and do not take into account temporal variations of the quality. The clips are usually viewed in randomized order and with constant quality throughout their entire (and short) duration. After the stimulus has been presented, the quality rating is requested from the assessors, most often on a 5-point MOS scale. In such a situation, assessors are not really involved in the audio-video presentation, focusing rather on the evaluation task itself. This is fairly uncommon in a natural viewing environment, where a stimulus usually is of longer duration (e.g. full movie) and where visibility of the distortion, and hence perception of the quality, varies as a function of time and scene content [15, 2]. For such situations, non-intrusive, continuous measurement methods which allow for evaluation of long duration test material seem to be more suited, as they promise results more closely related to real-world viewing scenarios.

To cope with some of the above requirements, the Single Stimulus Continuous Quality Evaluation (SSCQE) method has been developed and incorporated into ITU recommendation BT.500-7 [12] in 1996. The SSCQE allows for continuous evaluation of presented material (up to 30-min long) by using a slider to indicate the perceived quality. The slider represents a simple scale (typically from 0 to 100) and can be adjusted any time the user chooses to. It has been reported that the SSCQE is too demanding for assessors performing a real evaluation task and that continuous operation of the slider can be distracting [6]. Moreover, the method is designed only for quality assessment of video material where accompanying audio might be introduced. In spite of the above shortcomings, the SSCQE remains the only internationally accepted recommendation allowing for instantaneous quality assessment of long duration content.

Recently, however, increased interest towards replacement/improvement of the SSCQE has been observed. Alternative methods able to catch momentary changes over longer period of time by means of different types of rating instruments (e.g. joystick [16], glove [7], steering wheel [14], etc.) have been proposed. The aim of these methods is to relax constraints related to the traditional slider mechanism and also to extend their suitability for different usage scenarios (e.g. mobile context). Nevertheless, any improvement over the SSCQE with respect to accuracy of the obtained results has not been proven. Moreover, user fatigue related to continuous operation of such devices for a prolonged period of time has not been investigated. In general, apart from the different devices being used to gather quality ratings, no major methodological changes in comparison to the SSCQE have been introduced. All the mentioned methods (including the SSCQE) elicit information reflecting the perceived quality using alike type of rating scales (typically form 0 to 100, partitioned into five equal units analogous to the ordinal five-point quality scale) affiliated with each of the devices. Additionally, affective and cognitive attributes (e.g. annoyance, information loss) cannot be determined by this type of procedures.

The lack of alternative methodologies which would allow to overcome the mentioned limitations and which would expand our knowledge in relation to cognitive aspects of user's experience significantly hinders progress in this field. Therefore, in [4] we have proposed a different approach towards continuous examination of quality variations in audio, video or audiovisual stimuli over extended periods of time (the general description of our method is presented in Section C.2).

So far, the suitability of the technique for quality assessment of video (with accompanying, undistorted sound track) has been demonstrated in [4] and [5]. The objective of the work described in this paper is to apply and further investigate our methodology. More specifically, two studies will be presented. In the first study (experiment 1) the effect of the time dimension upon the perception of impaired audio quality will be investigated. The second study (experiment 1, 2 and 3) focuses on cross-modal perception of quality changes when only auditory or only visual, or both modalities, are impaired at the same time.

The paper is organized as follows. Section C.2 briefly summarizes the methodology used for all experiments. The experimental setup in terms of test material, participants as well as test conditions and procedures is described in Section C.3. In Section C.4, results of both studies are presented and discussed. Finally, the conclusions are given in Section C.5.

C.2. Method Description

In [4] we have proposed a novel methodology which represents a different approach towards continuous quality evaluation of long duration material.

Instead of measuring the quality by using traditional Mean Opinion Score (MOS) based approaches, the method allows participants to select the most appreciated quality themselves. In case quality degradation occurs subjects have the possibility to adjust the quality to a desired level by using a rotary controller (e.g. scroll wheel, knob, etc.). The optimal (possibly the highest) quality level can be achieved solely by appropriate adjustment based on perceptual appreciation of what is seen or heard. There is no need for assessors to translate their sensations into a value on a numerical scale, thus avoiding most of the typical problems associated with MOS related concepts [8]. During the assessment task, automatic quality alterations are introduced at random or periodically and stepwise (e.g. the degradation procedure begins every third minute and subsequently, the quality level decreases every 10 s). The participant's response (movement of the controller) to a quality change stops the automatic degradation procedure and provides him/her with full control over the quality adjustment. Turning the knob clockwise increases the quality level of the displayed material up to the point where the highest quality is attained. Rotating the device further clockwise introduces a gradual decrease in quality. This is in a way a penalty introduced when the maximum quality level is being surpassed. The process is reversible and by rotating the knob in the opposite direction the subject can return to the highest quality level (or decrease the quality level, if the maximum quality level has not been surpassed). A number of stimuli with different

quality levels are produced beforehand, and for each of them a numerical value is allocated internally (e.g. lowest quality- 0, best- 10). Rotation of the knob selects between them. Users' responses are collected automatically by the system, at a time resolution of one millisecond and with values which lie within the range of the corresponding quality levels. The method allows for gathering information about user's behavior and expectations in a non-intrusive way which is an advantage when cognitive aspects of quality, like annoyance, information loss are considered. By letting users adjust the perceived quality according to their own expectations we can learn more about the complex process of multi-modal quality evaluation, which in turn can help in the development of more accurate objective models. A more detailed explanation of the operation principles of this method can be found in [4].

C.3. Subjective Evaluation

C.3.1. Test Subjects

Twenty participants, 14 males and six females, took part in the first experiment and received a cinema ticket for their participation. The participants' mean age was 31 years (age range 22–61 years). Ten out of twenty subjects had participated in earlier video quality assessments using the same testing methodology. Nineteen participants reported to have normal hearing and vision whereas one subject reported a doubt with respect to hearing acuity before the start of the main task. With regards to the purpose of the experiment, results provided by this person were excluded from further data analysis.

For the additional experiments two different groups of ten subjects were employed. They were recruited conforming to the same sensory requirements as in the first experiment and were remunerated in the same way. The mean age of participants of the second experiment was 32.1 years (age range 26–48) and participants of the third one - 30.4 years (age range 25–42).

C.3.2. Test Material and Its Preparation

A 32 min and 2 s extract from the third episode of the *BBC* nature documentary series titled *Life* was used in all three tests. The duration of the clip was selected with respect to the experimental design while still maintaining a logical structure with a beginning and end. The auditory part of the material contained speech, background music, nature sounds, and also dynamic, action-type music. The visual part was full of different shots and camera angles, including slow motion as well as action scenes with fine details, closeups and movement. A high quality version was extracted from a Bluray disc edition and served as a starting point for further processing.

For the first experiment the original audio track (DTS, 5.1 ch, 16 bit/48 kHz, 1536 kbps) was downsampled to 44.1 kHz and downmixed to two channels PCM

format (CD quality). Subsequently, the prepared audio stream was encoded as MP3 with LAME encoder [1] at 11 different bit-rates (32, 48, 56, 64, 80, 96, 112, 128, 160, 192 and 320 kbps) with VBR off. The compression rates were selected according to results from a pilot test. The volume level of all audio quality levels was normalized (unweighted sound level) to maintain loudness consistency among them. The original video clip (1080p version) was decoded to the YUV format (preserving the quality and resolution) and used for video playback. Upon playback, both modalities were in synch at all times. The test conditions are summarized in Table C.1.

Table C.1.: Test conditions of experiment 1.

Clip duration	32 min 2 s
Audio properties	2 ch, 16-bit/44.1 kHz
Audio compression rates (kbps)	32, 48, 56, 64, 80, 96, 112, 128, 160, 192 and 320
Video properties	HD 1080p (1920 × 1080), 25 fps
Video color scheme	16-bit YUV 4:2:0

For the purpose of the second experiment 11 different video quality levels were created using various values of a quantization parameter (QP) in *H.264*/AVC encoder (x264). To maintain the minimal difference that can be detected between two neighboring quality levels (just noticeable difference) the QP value for each level was chosen according to results in [20] and [19]. Afterwards, all video clips were decoded to YUV format (4:2:0). The original sound track (uncompressed version) was used for audio playback. A summary of the technical parameters can be found in Table C.2.

Table C.2.: Test conditions of experiment 2.

Clip duration	32 min 2 s
Audio properties	2 ch, 16-bit/44.1 kHz, 1411 kbps
Video properties	HD 1080p (1920 × 1080), 25 fps
Video color scheme	16-bit YUV 4:2:0
QP values	0, 16, 22, 24, 28, 30, 32, 36, 38, 40, 44

The test material for the third experiment was made from a mixture of 11 audio quality levels created in the first experiment and 11 video quality levels created in the second experiment. Detailed information can be found in Table C.1 for audio and Table C.2 for video.

C.3.3. Test Procedures

Participants received written and oral instructions prior to each of the experiments. The main part of all the experiments was preceded by a training session which con-

sisted of a 10-min long clip selected to span the same quality range as the test clip. During the training, subjects had time to familiarize themselves with the test methodology, the sensitivity of the rotary knob device and also with the different levels of impairments. Questions were allowed throughout the whole training session.

In the main section of the experiment an over 30-min long audiovisual clip was evaluated (as described above).

In the first experiment assessors were asked to react to audio quality changes only in case it was really audible for them. The first 3 min were used to familiarize participants with the reference sound quality (320 kbps). Thereafter, an automatic degradation process was introduced every 3 min with the quality levels further decreasing every 10 s. The subjects were advised that the first degradation procedure would start at 1 to 5 min into the clip.

For the second and third experiment the test procedure was practically the same as for the first one, with the difference that participants had to respond to video only or audio and video quality changes instead. In the third experiment, changes in audio and video quality levels were appearing simultaneously and in the same order (from good to bad).

Right after the main task participants were asked several questions regarding the easiness of the task, positive/negative aspects, difficulties (if any), and their overall experience regarding the methodology used for this test. Subjects were also asked about their interest in the presented material.

The experiment took place in a room designed to provide high quality listening and viewing conditions according to ITU recommendations BT.500-7 [12], P.911 [13] and BS.1116 [11]. Two DynAudio BM6A active loudspeakers were used for sound reproduction and a Pioneer PDP-5000EX plasma screen served as a display for video content. A USB control knob (PowerMate made by Griffin Technology) was used to instantaneously adjust the quality level in case of perceived degradation. To avoid too slow or too sudden changes, the sensitivity of the knob was set according to users' feedback from the previous experiment. Consequently, a 90° rotation was required to switch between adjacent levels but this was unknown to the participants. The total duration of the experiment ranged between 50 and 55 min and only one participant was performing the test at a time. Subjects were sitting in a cinema-style seat with a pullout tray, which was used as a stand for the controller.

C.4. Results and Discussion

C.4.1. Study 1: Evaluation of Audio Quality Requirements

The first experiment had two main objectives. Firstly, the goal was to further examine our novel method for continuous quality evaluation and secondly, we wanted to investigate how audio impairments related to various compression rates affect participants' responses to quality fluctuations over extended periods of time.

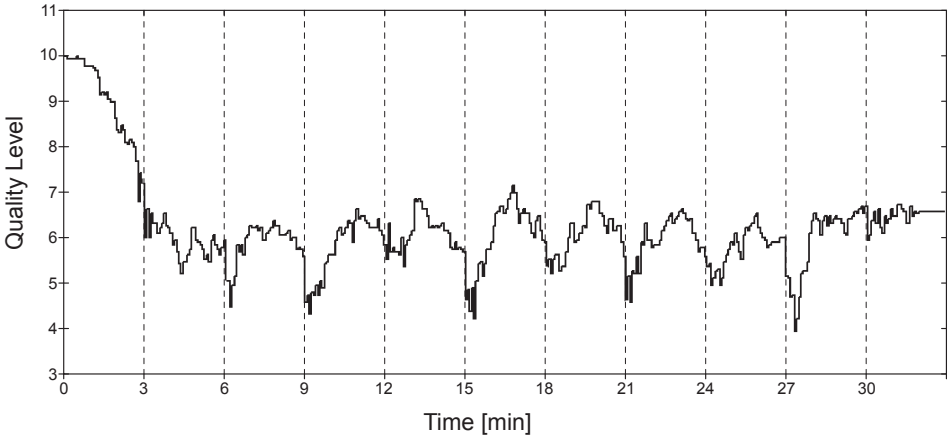


Figure C.1.: Results showing averaged responses of all participants with respect to variations in the sound quality of the audiovisual clip.

As mentioned previously, the test started with the maximum audio quality which lasted for the first 3 min, until the first degradation process began. One would expect a lack of subjects' responses (no movement of the knob) during this period, which turned out not to be the case (see Fig. C.1, where the values decreasing from 10 during the first 3 min indicate that assessors were decreasing the quality by themselves). One of the reasons for such a situation is the users' reaction to the noises created by the nature scenes (e.g. an arctic wind). Those noises caused the impression of audio quality degradation despite the fact that compression artifacts were not present. Trying to improve the quality of the sound by rotating the knob clockwise, participants started to decrease it. The second reason for this phenomenon might be related to some subjects being overly eager to use the controller during the first minutes of the test.

After this period of time participants were quite consistent in their choices and occurring variations are most probably related to particular scenes which represent different audio attributes. Looking at Fig. C.1 we can notice little variation in the means of users' responses during the last minutes of the main test. This is related to a specific type of audio material appearing at this particular time—mainly speech without background music. Subjects reported that it was easier for them to detect distortions at higher quality levels while speech was present. This is in accordance with results presented in [10]. Moreover, some of the test-persons declared that accompanying, high quality video diminished the effect of audio quality degradation, making them more tolerant to lower bit-rates (cross-modal masking effect).

Figure 1 suggests that the average participant would be satisfied with a quality level between 4 and 7, which corresponds to compression rates between 80 and 128 kbit/s. The overall mean value (6.16) implies that quality level 6 (112 kbit/s)

would be satisfactory for most of the participants throughout the entire duration of the clip.

To better understand the above considerations and also extend our knowledge about users' responses to quality changes a statistical analysis was performed.

Three subsets of data were created to proceed with further data examination. These subsets included:

a) average quality level of the last minute of each 3 min time slot (AQL) (this represents established/stationary quality preferences of the subjects)

b) response time to the automatic quality degradation right after the start of each 3 min time slot (RT)

c) quality level at the time when a user reacted to a quality change (QLRT)

In the following we will use the above abbreviations to refer to the description of those data subsets.

Due to the fact that the first 3 min were designed only to make participants familiar with the top quality the results for this time section were excluded from further analysis. The final size of the data matrix for each of the subsets was 19×9, where 9 corresponds to the number of three-minute time sections (without the first one) and 19 to the number of participants that were included.

The Analysis of Variance (ANOVA) was used as a tool to reveal dependencies between periods of time from one to another automatic degradation procedure (3 min time slots) and participants' reactions to quality variations. With ANOVA, the variability of scores between conditions and within conditions is analyzed and compared. This helps to find out if the independent variable has a significant effect on the dependent variable.

For validity of the results the data was checked for normality and homogeneity of variance across time slots as well as across users. All data sets showed close to normal distribution and close to homogeneous variance. A mixed-effects model of ANOVA with participants representing a random-effect type factor and time slots representing a fixed-effect type factor was used. Such a model can help to better understand which factor is responsible for most of the variation in the data and also compare the main effect of each of them. To justify a claim of a statistically significant effect the 0.05 level of significance was used.

Detailed results of the ANOVA test conducted on the subset data a) are presented in Table C.3 and in Fig. C.2.

We can see that dissimilarities between mean quality expectations between users are statistically highly significant ($F = 6.374$; $p<0.005$) and that variations in mean quality among participants are quite big (see Sum of Squares of User in the ANOVA table). This might be due to differences between the ways assessors were using the knob for the adjustment of quality or due to individual dissimilarities in hearing acuity or different interpretation of the instructions. On the other hand it can be noticed that such a phenomenon is not present if time sections are considered. In the statistical sense the differences between mean quality expectations between the

Table C.3.: Results of ANOVA for data subset a) in study 1.

Source	df	Sum of squares	Mean square	F	p
User	18	356,049	19,781	6,374	0,000
Time Slot	8	3,914	0,489	0,158	0,996
Error	143	443,768	3,103		
Total	169	803,731			

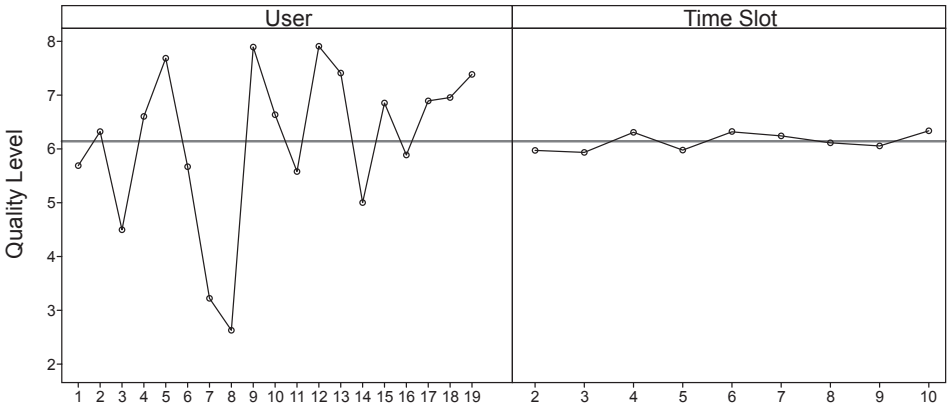


Figure C.2.: Main effects plot for quality levels in study 1, averaged over the last minute of each 3 min time interval (AQL).

time sections are not significant (F = 0.158; p>0.005) which denotes that subjects were quite consistent in their choices throughout most of the clip's duration. This follows the pattern from a previous study on video quality assessment using the same method [5].

We were also curious to know if differences between the users depended on the level of the time section factor and vice versa. No such interaction between these two factors has been found.

For data subset b), Table C.4 shows that differences between mean reaction times between users are statistically significant (F=3.042; p<0.005). Participants were reacting to the automatic quality changes differently, starting usually from different quality levels.

This fact explains such big variations in mean reaction times between participants. Contrary to the between-user factor, no significant dissimilarities in reaction time to quality changes across time sections have been found (F = 0.826; p>0.005). The mean time until people reacted to gradual quality decreases was roughly 24 s which corresponds to a 3 level drop in quality before the test-subject reacted.

From Fig. C.3 we can see the distribution of means of reaction time with respect to users and time slots. It can be noticed that the variation due to time intervals is much smaller than the variation due to users' differences.

Table C.4.: Results of ANOVA for data subset b) in study 1.

Source	df	Sum of squares	Mean square	F	p
User	18	1,640E10	9,113E8	3,042	0,000
Time Slot	8	1,980E9	2,476E8	0, 826	0, 581
Error	137	4,104E10	2,996E8		
Total	163	5,942E10			

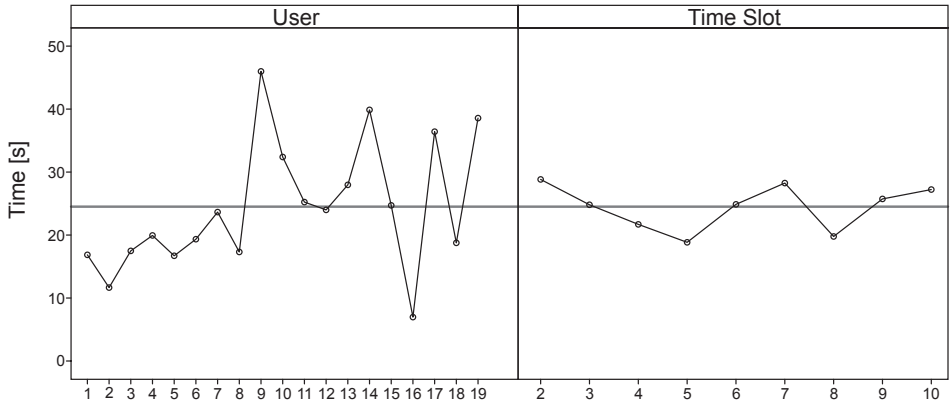


Figure C.3.: Main effects plot for reaction time (RT) in study 1.

Similar results were obtained for the data subset c) (see Table C.5 and Fig. C.4).

Table C.5.: Results of ANOVA for data subset c) in study 1.

Source	df	Sum of squares	Mean square	F	p
User	18	382,436	21,246	6,177	0,000
Time Slot	8	30,367	3,796	1,104	0, 365
Error	137	471,230	3,440		
Total	163	884,033			

Figure C.5 shows the relation between the average quality levels set by participants during the last minute of each 3 min time interval (AQL), and the average quality levels corresponding to the time at which participants detected a change (QLRT). One could notice that except for the very beginning the difference between these two plots with respect to quality levels is relatively constant and on average equal to three levels. The smaller difference occurring in the beginning might be related to the bigger attention participants paid during the first minutes. However, as observed in Table C.5, there was no significant effect of this tendency. The above suggests that it is easier for a person to distinguish between neighboring quality levels while concentrated on the task (e.g. quality adjustment) than when

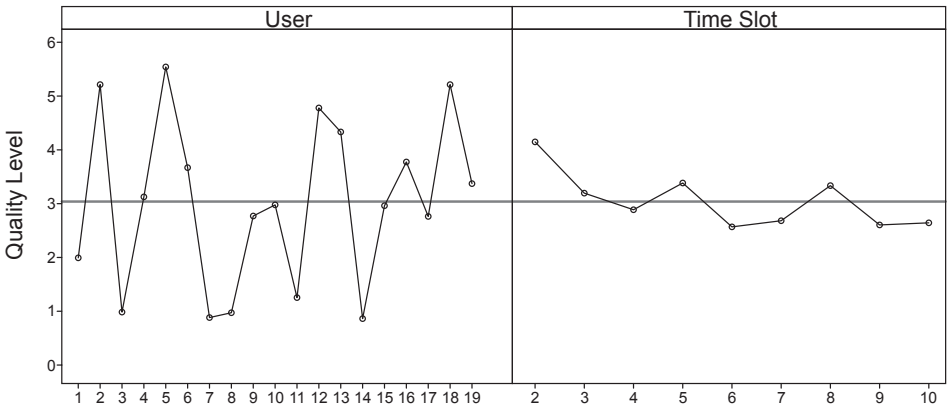


Figure C.4.: Main effects plot for quality levels, in study 1 at the time when a user reacted to quality change (QLRT).

the change happens at random and is independent of his/her actions (e.g. automatic degradation procedure). In addition, the performance of those test-subjects who had participated in the previous experiment was checked against those for whom the methodology was new. It turned out that this learning process did not affect the users' performance; no significant difference was found between these two groups. This might imply that the method is quite intuitive and easy to follow from the very first time it is used.

C.4.2. Study 2: On the Cross-Modal Perception of Impairments

The aim of this study was to examine relationships between results obtained from all three experiments. Specifically, we were interested to find out if there is a difference in subjects' reactions to quality changes of audiovisual presentation in case when only one of the modalities is distorted (audio or video) or when both modalities are impaired at the same time.

In order to proceed with the data analysis, two subsets of data (AQL and QLRT) for each of the additional experiments were created in the same manner as described in Section C.4.1. Those data subsets of size 10×9 (10 - number of participants, 9 - number of time slots) served as a basis for further investigations.

A 2×2 fixed-effects, between subject factorial design ANOVA was computed to compare means of the three different impairment conditions. These impairment conditions refer to an audiovisual clip containing:

1. audio impairments solely (AI)
2. video impairments solely (VI)
3. both: audio and video impairments (AVI)

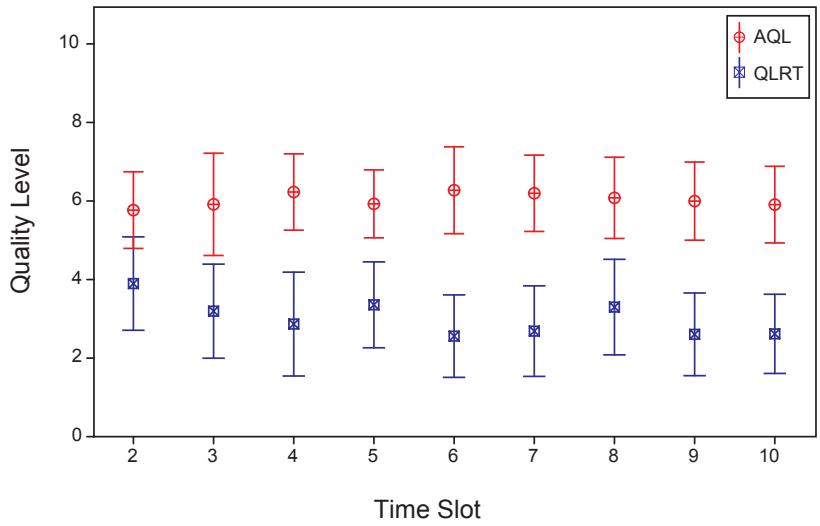


Figure C.5.: Comparison of subjects' sensitivity to audio quality changes under different conditions (AQL vs. QLRT) with 95% confidence intervals in study 1. According to the ANOVA test, the differences between AQL and QLRT are statistically significant across all the time slots (F=151.156; p<0.0001).

Comparisons were drawn between combinations of all of the above (AI vs. AVI, VI vs. AVI, and AI vs. VI) and for 2 data subsets (AQL and QLRT).

A graphical representation of results for the first pair comparison (AI vs. AVI) is shown in Fig. C.6. We can clearly see that differences in means between AI and AVI in both plots are relatively big and that associated confidence intervals are not overlapping (except for one time slot). It implies that dissimilarities between AI and AVI in both cases are highly significant. This is supported by results of the ANOVA test for subset data a) (F(1.243) = 73.714; p<0.0001) and subset data c) (F(1.243) = 93.315; p<0.0001), respectively. Apparently, participants are able to set the quality on a higher level and discover the quality changes earlier when impairments are introduced in both the auditory and visual modality at the same time. These results could be caused by the specific choice of degradation levels for the audio and video impairments, respectively, so the comparisons below address that aspect.

The comparison of VI and AVI can be seen in Fig. C.7. This time the confidence intervals for means are overlapping for most of the time slots in both plots. However, the ANOVA exhibits significant effect of means comparison for subset data a) (F(1.243) = 32.239; p<0.0001) as well as for subset data c) (F(1.243) = 55.980; p<0.0001). The results seem to uphold our previous finding that impairments introduced in both modalities make the process of quality selection and reaction to quality changes more effective and easier.

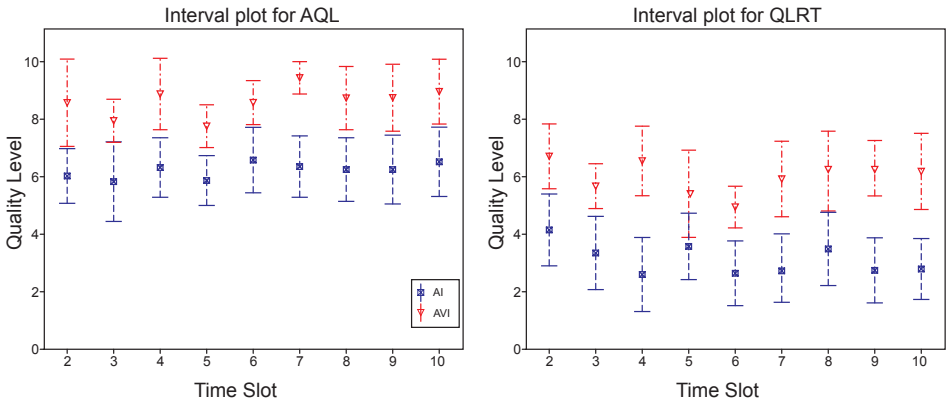


Figure C.6.: Comparison of mean quality levels for AQL (*left plot*) and QLRT (*right plot*) with 95% confidence intervals at a specified time slot between a clip with audio impairments solely (AI) and same clip with audio and video impairments (AVI).

The last comparison is between VI and AI. We can notice that mean quality levels for particular confidence intervals are much closer to each other (see Fig. C.8) than it was observed in previous comparisons. Furthermore, the time intervals greatly overlap for all of the time slots. However, calculated results for subset data a)($F(1.172) = 11.346$; p=0.001) and subset data c) ($F(1.172) = 8.555$; $p = 0.004$) show again significant effect at the 0.01 level of significance.

Figure C.9 shows two interval plots of means with respect to different impairment conditions. The mean differences (MDs) between the above conditions are summarized in Table C.6. Interestingly, the difference between mean quality levels is larger between the multi-modal and the visual case (AVI vs. VI; MD=1.39) than that between the two unimodal cases (VI vs. AI; MD=1.03). This is presumably due to superadditivity effects between the modalities as described in [9]. A more careful balancing of the corresponding impairment/quality levels in the two modalities could be achieved by using the experiment described in [17].

Table C.6.: Mean differences (MDs) between impairment conditions for AQL and QLRT (all statistically significant at level 0.01).

Comparison	MDs for AQL	MDs for QLRT
AVI vs. AI	2.42	2.90
AVI vs. VI	1.39	1.95
VI vs. AI	1.03	0.95

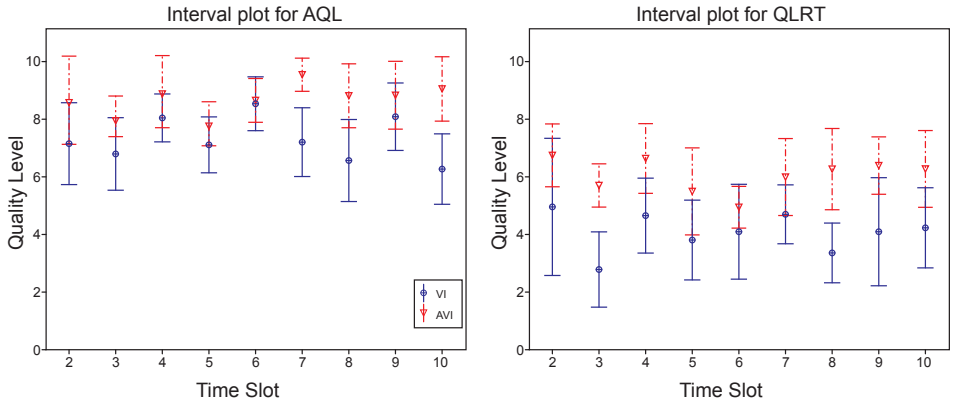


Figure C.7.: Comparison of mean quality levels for AQL (*left plot*) and QLRT (*right plot*) with 95% confidence intervals at a specified time slot between a clip with video impairments solely (VI) and same clip with audio and video impairments (AVI).

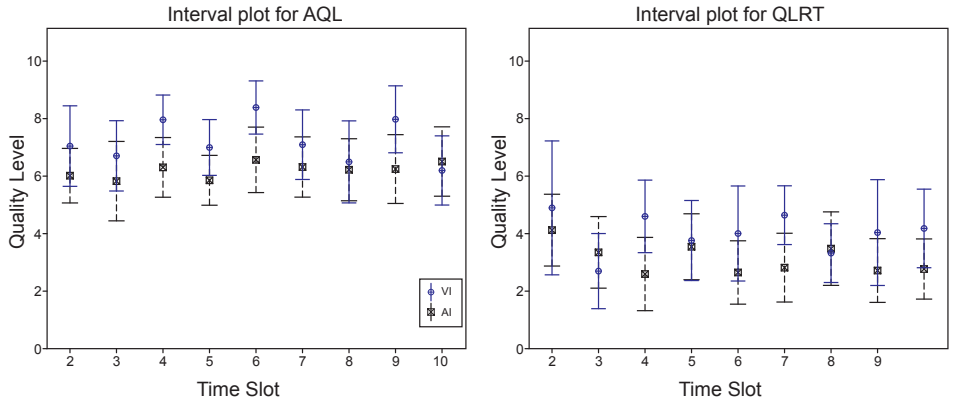


Figure C.8.: Comparison of mean quality levels for AQL (*left plot*) and QLRT (*right plot*) with 95% confidence intervals at a specified time slot between a clip with audio impairments solely (AI) and same clip with video impairments solely (VI).

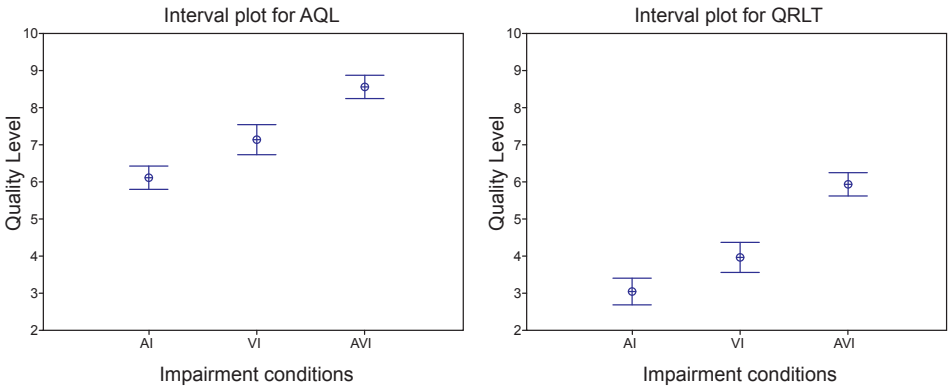


Figure C.9.: Comparison of mean quality levels for AQL (*left plot*) and QLRT (*right plot*) with 95% confidence intervals at different impairment conditions (AI, VI, and AVI).

Summarizing, it seems that deterioration of the audio quality solely in a high definition audiovisual presentation influences subjects' perception of quality changes to a smaller extent than when only video impairments are present or when both modalities are distorted simultaneously. The results suggest that high visual quality can mask the effect of audio degradation and vice versa (but to a lesser degree), which is in line with previous results presented e.g. in [3, 18] and elsewhere. The studies presented here confirm that this also holds true for long duration content.

C.5. Conclusions

The main part of this work was dedicated to present the experimental setup and results of a long duration audiovisual content audio quality experiment. It has been found that participants' preferences regarding audio quality are relatively constant with respect to time when the content is of a relatively similar type. A mixed-effects ANOVA model was used to gain insight into the relationship between means of specific factors. It turned out that the time factor does not influence quality ratings, whereas the between-user factor does. Similar conclusions can be drawn with respect to the response time to automatic quality changes. The between-user factor is responsible for most of the variations in the data, and the dissimilarities among participants (whether related to selected quality level or reaction time) are statistically significant. Furthermore, it has been shown that subjects are substantially more sensitive to quality changes when they themselves are in control of the quality adjustment process, than when the quality degradation process is controlled externally. We can see that the time dimension does not influence the audio quality expectation, which is contrary to what could be expected. The outcome might be

different for much longer duration stimuli (e.g. full length movie), but this needs to be studied.

In addition, the suitability of our method for quality assessment of long duration audio streams (with accompanying, high quality video) has been demonstrated. The results uncover time/quality related dependencies and expand our knowledge of users' responses to quality changes over extended periods of time.

Besides the main study, two additional experiments were conducted with the same test material and under the same test conditions as in the first study. The aim of the additional work was to investigate possible dissimilarities in reactions of participants to quality changes of audiovisual presentation in case of audio only, video only, and simultaneous audio and video distortions. The results show significant differences between each of the above cases, i.e., subjects reacted faster to quality changes and preferred higher quality levels when impairments were introduced in both modalities simultaneously. However, so far this has only been verified for one type of content. Our next step will be to check if these findings hold when different types of content are considered.

Acknowledgment

This work was performed within the PERCEVAL project, funded by The Research Council of Norway under project number 193034/S10.

References

[1] LAME (Lame Ain't an MP3 Encoder). The Hydrogenaudio recommended MP3 encoder, http://lame.sourceforge.net.

[2] R. P. Aldridge, J. Davidoff, M. Ghanbari, D. S. Hands, and D. E. Pearson. Measurement of scene-dependent quality variations in digitally coded television pictures. In *Proc. of IEEE Vision Image Signal Processing*, volume 142, pages 149–154, 1995.

[3] J. G. Beerends and F. E. De Caluwe. The influence of video quality on perceived audio quality and vice versa. *Journal of the Audio Engineering Society*, 47(5):355–362, 1999.

[4] A. Borowiak, U. Reiter, and U. P. Svensson. Quality evaluation of long duration audiovisual content. In *Proc. of the 9th Annual IEEE Consumer Communications and Networking Conference. Special Session on Quality of Experience (QoE) for Multimedia Communications*, pages 353–357, Las Vegas, 2012.

[5] A. Borowiak, U. Reiter, and O. Tomic. Measuring the quality of long duration AV content. Analysis of test subject/time interval dependencies. In *EuroITV - Adjunct Proceedings*, pages 266–269, Berlin, 2012.

[6] A. Bouch and M. A. Sasse. The case for predictable media quality in networked multimedia applications. In *Proc. of ACM/SPIE Multimedia Computing and Networking (MMCN)*, pages 188–195, San Jose, 2000.

[7] S. Buchinger, W. Robitza, M. Nezveda, M. Sack, P. Hummelbrunner, and H. Hlavacs. Slider or glove? Proposing an alternative quality rating methodology. In *Proc. of the 5th International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, Scottsdale, Arizona, 2010.

[8] K. T. Chen, C. C. Wu, Y. C. Chang, and C. L. Lei. A crowdsourceable QoE evaluation framework for multimedia content. In *Proc. of the 17th ACM international conference on Multimedia*, pages 491–500, Beijing, 2009.

[9] N. P. Holmes and C. Spence. Multisensory integration: Space, time and superadditivity. *Current Biology*, 15(18):762–764, 2005.

[10] R. Huber and B. Kollmeier. PEMO-Q—A new method for objective audio quality assessment using a model of auditory perception. In *Proc. of IEEE Transactions on Audio Speech and Language Processing*, volume 14, pages 1902–1911, Piscataway, 2006.

[11] ITU-R Rec. BS.1116. Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems. Int. Telecomm. Union, Geneva, 1997.

[12] ITU-T Rec. BT.500-7. Methodology for the subjective assessment of the quality of television pictures. Int. Telecomm. Union, Geneva, 1996.

[13] ITU-T Rec. P.911. Subjective audiovisual quality assessment methods for multimedia applications. Int. Telecomm. Union, Geneva, 1998.

[14] T. Liu, G. Cash, N. Narvekar, and J. Bloom. Continuous mobile video subjective quality assessment using gaming steering wheel. In *Proc. of the 6th International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, page 6, Scottsdale, Arizona, 2012.

[15] N. K. Lodge and D. Wood. Subjectively optimizing low bit-rate television. In *Proc. of the IEEE International Broadcasting Convention IBC 94*, pages 333–339, Amsterdam, 1994.

[16] O. Nemethova, M. Ries, A. Dantcheva, and S. Fikar. Test equipment of time-variant subjective perceptual video quality in mobile terminals. In *Proc. of International Conference on Human Computer Interaction*, Phoenix, 2005.

[17] U. Reiter and J. Korhonen. Comparing apples and oranges: Subjective quality assessment of streamed video with different types of distortion. In *Proc. of the International Workshop on Quality of Multimedia Experience (QoMEX)*, pages 127–132, San Diego, 2009.

[18] R. L. Storms and M. J. Zyda. Interactions in perceived quality of auditory-visual displays. *Presence: Teleoperators and Virtual Environments*, 9(6):557–580, 2000.

[19] H. Wang, X. Qian, and G. Liu. Inter mode decision based on just noticeable difference profile. In *Proc. of 2010 IEEE 17th International Conference on Image Processing*, pages 297–300, Hong Kong, 2010.

[20] X. Yang, Y. Tan, and N. Ling. Rate control for H.264 with two-step quantization parameter determination but single-pass encoding. *EURASIP Journal on Applied Signal Processing*, 2006:1–13, 2006.

D. Long Duration Audiovisual Content: Impact of Content Type and Impairment Appearance on User Quality Expectations Over Time

Adam Borowiak[1], Ulrich Reiter[1]

[1] *Department of Electronics and Telecommunications, Norwegian University of Science and Technology (NTNU)*

Proceedings of The 5th International Workshop on Quality of Multimedia Experience (QoMEX), pp. 200–205, Klagenfurt, 2013.

Paper D

Errata

pp. 110, paragraph 2, line 5: ”QP = 0,16,22,26,28,30,32,36,40,44” should be ”QP = 0,16,22,24,28,30,32,36,40,44”

Abstract

In this paper, two questions related to long duration audiovisual content are addressed: firstly, we investigate whether users' quality expectations/requirements develop differently over extended periods of time for different types of content. We have found indicators suggesting that high spatial and temporal activity indices decrease quality requirements over time. Secondly, we show that for long duration content, viewers' quality requirements are independent of magnitude and appearance (gradual increase vs. spontaneous leap) of quality impairments introduced.

D.1. Introduction

Subjective evaluation of video quality is usually performed using short video clips. Only more recently has there been an increased interest in long duration content and its quality evaluation as witnessed in the studies by Staelens et al. [13] and Borowiak et al. [3].

The majority of standardized subjective methods specified by the International Telecommunication Union (ITU) is not adequate for this type of study/content. The standardized methods are typically intended for short clips (less than 10 s duration), with a constant quality throughout the entire clip length. Moreover, in such procedures time-varying and scene-dependent effects of impairments cannot be assessed, making them not appropriate for real-life viewing scenarios. The Single-Stimulus Continuous Quality Evaluation (SSCQE) method set forth in ITU recommendation BT.500-7 [8] was developed to overcome the mentioned limitations. This is a continuous quality assessment method allowing participants to rate the quality of a long video sequence with the help of a slider mechanism along an associated quality scale. Although it allows capturing time varying impairments it's not free from disadvantages as described in [6] and [5]. Moreover, SSCQE as well as other mean opinion score (MOS) based approaches makes participants focus more on the detection of impairments than on the perception and cognition of the stimulus they are exposed to [12]. In our previous work we have therefore developed a subjective assessment methodology [3] that overcomes most of the drawbacks of other long duration content methods, and which was also employed here. Contrary to other methods, it allows assessing viewers' quality requirements directly by giving control over presentation quality to the test subject.

The goal of this study was twofold: firstly, we wanted to investigate the influence of different types of long duration content on quality perception under specific degradation conditions. This work is a continuation of our previous study in which the time dimension was found not to play any major role with respect to users' quality expectations [4, 2]. More specifically, it was shown that the quality expectations are rather constant throughout the entire stimulus duration. These earlier conclusions were drawn based on one type of content only, but as reported by Kortum et al. [11], the type of content may have a significant effect on subjective quality ratings. They found that quality of desirable audiovisual content is rated significantly higher compared to content which is neutral or undesirable. Their study involved 2-min long movie clips from 20 different movies released by major film studios. The results have been confirmed in their later work, where even more pieces of content were employed [12]. Here, we wanted to study whether a similar content dependency can be observed for long duration audiovisual content. In our study, however, the focus is rather on investigation of content influence on users' expectations developing over extended periods of time than directly on the quality ratings. The second goal of our study is to find out whether different impairment appearances, i.e. slow gradual decrease in quality vs. large instantaneous

(catastrophic) quality loss, influence viewers' quality expectations in long duration content scenarios.

The remainder of the paper is organized as follows: Section D.2 describes the experimental study performed to answer the above two research questions. Section D.3 presents an analysis of the collected data and discusses the results. Finally, Section D.4 summarizes the main findings and draws some conclusions.

D.2. Study Design

D.2.1. Test Methodology

For the experimental purpose our methodology described in [3] was used. The method allows participants to adjust the quality to a desired level in case of quality degradation. By using a rotating adjustment device (e.g. knob) users can select the most appreciated quality level themselves at any time of the test. Rotating the device clockwise increases the displayed quality until the maximum level is attained. Further clockwise rotation begins a process of gradual quality decrease. In case the maximum quality is over-passed the counterclockwise rotation of the device allows returning to the highest level. Every so often (periodically or at random) an automatic degradation procedure takes place (e.g. the quality decreases gradually or in one step). As soon as the user responds to the quality change by using the rotating device, the degradation procedure stops and full control over the quality adjustment is given to her/him. More detailed operational principles of the method can be found in [3].

D.2.2. Participants

Twenty one participants (14 male, seven female) aged between 16-50 years (M = 30.6, SD = 8.2), mostly naïve or untrained, took part in the study. All subjects reported to have normal hearing and normal or corrected to normal vision.

D.2.3. Content Selection and Processing

Four high definition (HD) audiovisual sequences (Bluray edition, 1080p, 25 fps), representing various types of content, were selected for the experiment. The genres of these sequences can be described as follows: Computer Animation, Action Movie, Opera and Nature Documentary. The video and audio characteristics, such as the amount of temporal information (TI), spatial information (SI) and the type of audio track can be found in Table D.1. According to ITU-T Recommendation P.910 [9], the SI indicates the amount of spatial detail of a picture and was calculated as a standard deviation of Sobel filtered luminance plain. The TI indicates the amount of temporal changes of a video sequence and was calculated as the standard deviation of the difference between pixel values in successive frames.

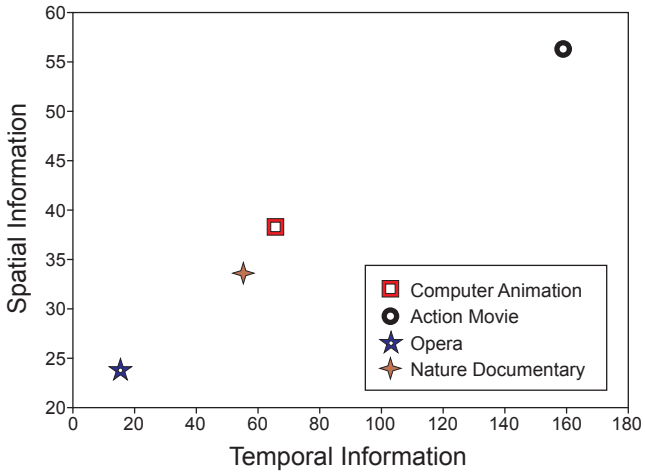


Figure D.1.: Spatial and temporal information of the test clips.

The numbers reported in Table D.1 are means of SI and TI values over all frames. The averaging over smaller parts of the clips (2250 frames/90 s) resulted in similar differences (with respect to TI and SI) between the mentioned content types. A graphical representation of the calculated spatial and temporal indices is shown in Fig. D.1. Audio was categorized base on dominant audio signal characteristics (e.g. speech, vocal, music, other sounds like natural noises, nature sounds or special effects).

Table D.1.: Selected sequences and their properties.

Genre	TI	SI	Audio
Computer Animation	66.34	38.07	Speech/Sound
Action Movie	159.85	56.67	Sound
Opera	15.68	23.70	Vocal/Music
Nature Documentaryl	56.96	33.40	Sound/Speech

From each sequence, a 9-min long clip (semantic structure preserved), was extracted and served as a starting point for further processing. The x264 (H.264/MPEG4-AVC) encoder [1] was used to prepare different quality levels for each of the test clips by employing various quantization parameter (QP) values. In total, 10 quality levels (QP = 0,16,22,26,28,30,32,36,40,44) were created for each of the four test clips. The selection of the QP's was based on results of our previous studies [3, 4]. Subsequently, the prepared video sequences were decoded to YUV format (4:2:0) and presented to participants together with related sound tracks of constant quality (2ch, 44.1 kHz, 16-bit, 1411 kbps). Audio and video were in synch at all times.

D.2.4. Test Procedures

The experiment consisted of one, around 56-min long session divided into 3 stages: instructions & training (15 min), main test (36 min) and questions (5 min).

At the beginning of each session, subjects were given initial instructions for the test procedure and operational principles of the method. In order to make participants familiar with the technique and the adjustment device, a training sequence of 3-min duration was viewed. The training comprised an audiovisual clip representing the same quality range as used in the main test. The subjects were allowed to ask questions during the training.

The actual test consisted of four clips played one after another with a 5 s break in-between, during which a gray image was displayed on the screen. The playback order was randomized individually for each participant to avoid the sequencing effect. The lowest quality level was a starting point for all the clips and subjects were supposed to adjust the quality to their preference. Subsequently, an automatic degradation procedure was introduced every 90 s into the clip. The procedure was instantly decreasing video quality level from currently displayed level to the lowest one. Immediately after the main task, subjects were asked questions about their interest/involvement in the presented material and their general experience regarding the methodology used.

D.2.5. Presentation of the test material

The experiment was conducted in a room fulfilling ITU recommendations BT.500-12 [7] and P.911 [10] with respect to lighting, viewing and acoustical conditions. The clips were displayed on a 50-inch plasma screen (Pioneer PDP-5000EX) and the sound was reproduced through two active monitor loudspeakers (Dynaudio BM6A). A USB controller (Griffin PowerMate) served as the quality adjustment device. The conceptual structure of the experimental setup can be seen in Fig. D.2.

D.3. Results and Discussion

Prior to statistical analysis, a subset of data consisting of quality levels averaged over the last 30 s of every 90-s long time interval was created for each of the 4 stimuli. The 90 s refer to the time between the automatic degradation procedures of which the last 30 s represent the established/stationary quality preferences of the subjects. Each of the data subsets was of size 21 x 6, where 21 refers to the number of participants and 6 to the number of time slots. The data turned out not to be normally distributed (Kolmogorov-Smirnov; $p<0.05$) which resulted in a non-parametric test being used for further data analysis. As a non-parametric method allowing comparison of more than two independent samples the Kruskal-Wallis (K-W) test was used. The significance level of 0.05 was adopted for all statistical tests.

Closer inspection of Fig. D.3 suggests that the earlier reported phenomenon of constant quality expectation throughout the whole duration of a long duration clip

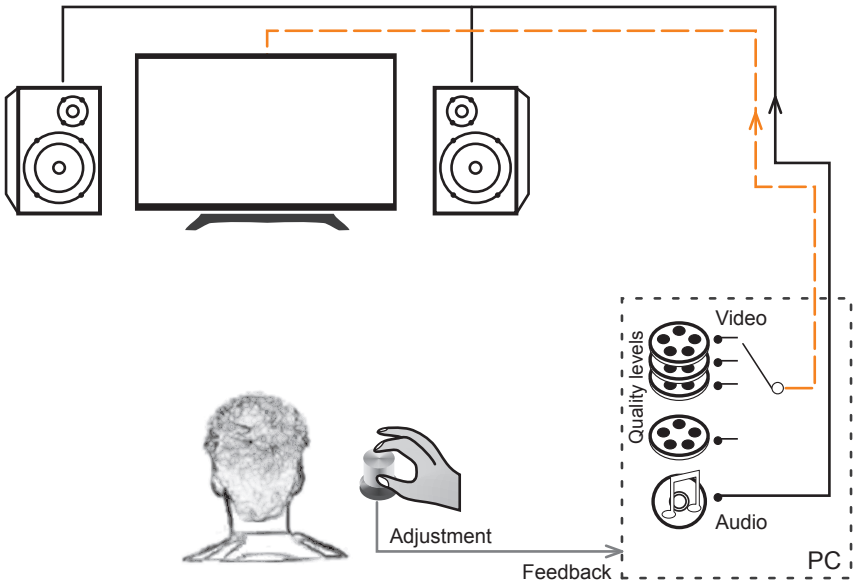


Figure D.2.: Experimental setup of the test.

does not hold for one of the employed content types – Action Movie, but that it's true for the remaining three.

The K-W results of the time slot means comparison for the Action Movie proved significant differences among them ($H(5) = 21.424$, $p<0.001$; Fig. D.3). The post-hoc analysis showed that mean quality levels of the first two time intervals were significantly different to the means of the remaining four time slots ($p<0.001$). It seems that the very fast action and significant amount of details present in the clip could cause the distortions being less visible (or less annoying) for the participants after a while. This might be a consequence of the phenomenon of distortions masking effect which occurs when high spatial and temporal activity, such as very high-motion scenes with a lot of details, occur [14]. Moreover, the majority of the test-subjects reported a great interest in such type of content during the post-test session. The combined effect of the above facts might be a reason behind the obtained results; however, a general influence of such type of content on quality expectations over extended periods of time needs to be verified.

The same K-W test was performed for the remainder of the stimuli. Results of the time slot means comparison for the Animation ($H(5) = 7.751$, $p>0.05$; Fig. D.3), Opera ($H(5) = 6.885$, $p>0.05$; Fig. D.3) and Nature Documentary ($H(5) = 3.974$, $p>0.05$; Fig. D.3) prove not to be significantly different, which is in line with results of our previous study (Study A) [4] in which nature documentary type of content was used.

Fig. D.4 presents a direct comparison of mean quality levels of all four content types. One can easily notice that the animation clearly stands out with higher

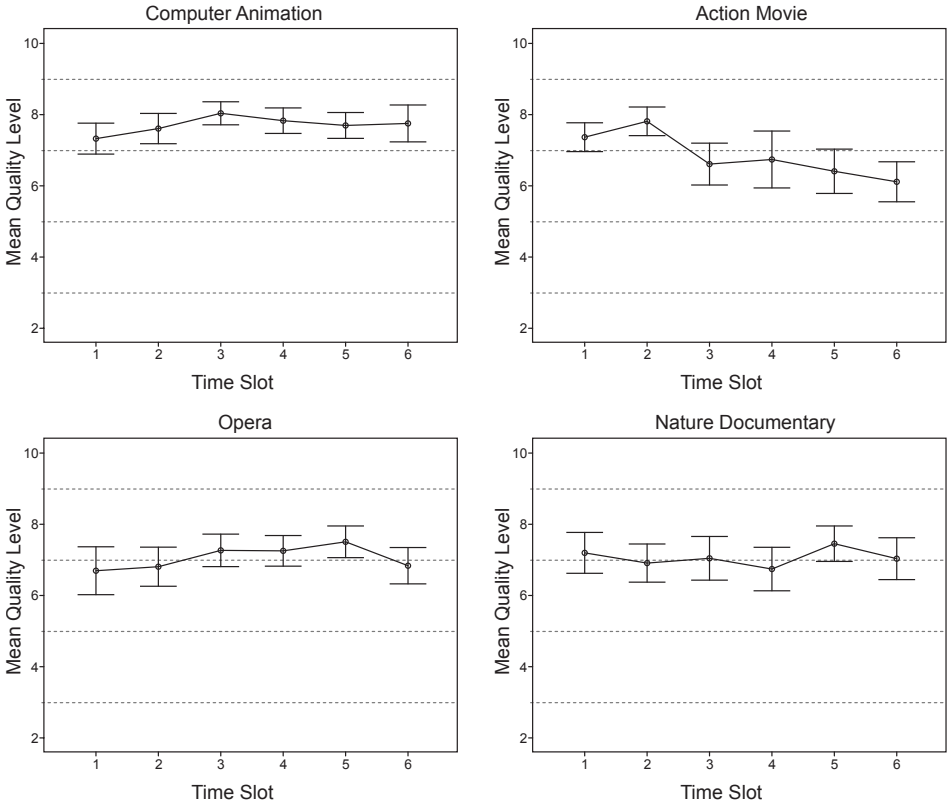


Figure D.3.: Mean quality levels set by test subjects for various content types. Error bars show 95% CI of mean.

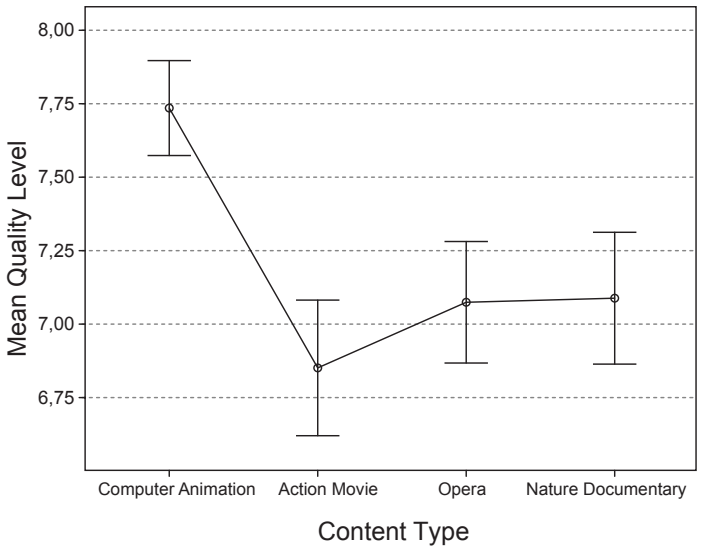


Figure D.4.: Comparison of mean quality levels for different content types. Error bars show 95% CI of mean.

mean quality level than the remaining content types. The observation has been confirmed by results of the K-W tests (H(3) = 33.068, $p<0.001$). This has nothing to do with the amount of spatial and temporal information (see Fig. D.1) but is rather directly related to the unique characteristics of the animated content type (such as color uniformity, stationary background, etc.). These characteristics make the quality degradations easier to being noticed by participants as they are better visible while parts of the picture remain unchanged for some period of time.

Furthermore, most of the participants reported less interest in the animation, which might make them more focused on the detection of quality changes than on the content itself.

A comparison of the mean quality levels of Action Movie, Opera and Nature Documentary does not reveal significant differences (K-W; H(2)=0.865, $p>0.05$).

We were also interested in comparing the impact of a test sequence being launched with the lowest quality level (lack of introduction of the reference quality) vs. starting with the reference quality level as in our previous studies. Furthermore, we looked into the impact of instantaneous catastrophic video quality degradation (as opposed to slowly developing quality impairment) on users' acceptance level. This was done in conjunction with results obtained in Study A [4] in which the same methodology and similar experimental material (30-min extract from another episode of the same nature documentary series) were used. In Study A, the quality levels were degraded stepwise while in this study the quality was degraded to the lowest level in one step. There is also a difference in number of quality levels used in both studies (12 in the first one vs. 10 in the current one). However, this is not

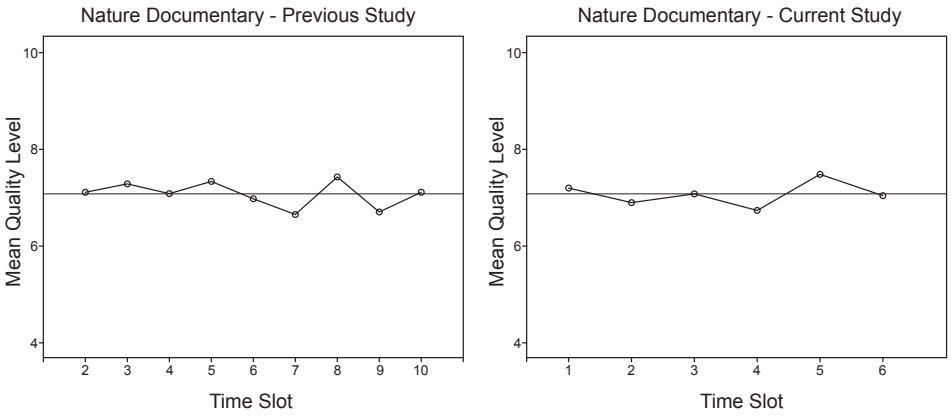


Figure D.5.: Comparison of two studies with different initial conditions. Left plot – stimulus starting with the reference quality level, degradations appear stepwise in time; right plot – stimulus starting with the lowest quality level, degradations appear immediately and in one step.

a major change as we have only removed the lowest quality levels with QP = 48 and QP = 38 from the set of 12 quality levels. This enables a direct comparison.

The results (see Fig. D.5) show that no matter whether the presentation starts with highest or lowest quality and no matter how the degradation process is handled – gradually or in one step, participants' final expectations are very similar on average, i.e. they set the quality to a similar level. The latter can be explained by the fact that subjects become more focused on the task when in charge of quality adjustment, which makes them more sensitive to smaller impairments [4, 2].

D.4. Conclusions

In this study we have performed a subjective assessment investigating the effects of time dimension and quality impairment appearance on users' quality expectations for different content types. It has been shown that the nature of content might have a significant effect on quality acceptance level over extended periods of time. However, this was proven only for the action-movie type content with a very fast motion and vast amount of details. The finding gives us only an indication that such phenomenon might occur and should not be treated as an absolute truth. The limited number of content types probably does not allow for conclusive demonstration of the results, however, it suggests which direction should be chosen. Therefore, further research with a broader range of different content types with similar spatio-temporal characteristics is needed.

In addition, it was found that the participants' quality expectations are similar regardless of whether the reference quality is introduced at the beginning of the

test clip or not. Based on the obtained results, we can also conclude that the way the degradations appear in the presented material (immediate drop from good to bad or stepwise in time) does not have an effect on the quality level that viewers are satisfied with.

Moreover, this study demonstrated the repeatability of results obtained using our method, thus making the method more desirable for purposes of quality assessment of long duration content than other traditional methods.

Acknowledgment

This work was performed within the PERCEVAL project, funded by The Research Council of Norway under project number 193034/S10.

References

[1] x264 - H.264/MPEG-4 AVC encoder released under the terms of the GNU GPL. http://www.iab.net/media/file/long-form-video-final.pdf.

[2] A. Borowiak, U. Reiter, and U. P. Svensson. Evaluation of audio quality requirements over extended periods of time using long duration audiovisual content. In *Advances in Multimedia Information Processing. PCM*, volume 7674 of *Lecture Notes in Computer Science*, pages 10–20, 2012.

[3] A. Borowiak, U. Reiter, and U. P. Svensson. Quality evaluation of long duration audiovisual content. In *Proc. of the 9th Annual IEEE Consumer Communications and Networking Conference. Special Session on Quality of Experience (QoE) for Multimedia Communications*, pages 353–357, Las Vegas, 2012.

[4] A. Borowiak, U. Reiter, and O. Tomic. Measuring the quality of long duration AV content. Analysis of test subject/time interval dependencies. In *EuroITV - Adjunct Proceedings*, pages 266–269, Berlin, 2012.

[5] A. Bouch and M. A. Sasse. The case for predictable media quality in networked multimedia applications. In *Proc. of ACM/SPIE Multimedia Computing and Networking (MMCN)*, pages 188–195, San Jose, 2000.

[6] K. T. Chen, C. C. Wu, Y. C. Chang, and C. L. Lei. A crowdsourceable QoE evaluation framework for multimedia content. In *Proc. of the 17th ACM international conference on Multimedia*, pages 491–500, Beijing, 2009.

[7] ITU-T Rec. BT.500-12. Methodology for the subjective assessment of the quality of television pictures. Int. Telecomm. Union, Geneva, 2009.

[8] ITU-T Rec. BT.500-7. Methodology for the subjective assessment of the quality of television pictures. Int. Telecomm. Union, Geneva, 1996.

[9] ITU-T Rec. P.910. Subjective video quality assessment methods for multimedia applications. Int. Telecomm. Union, Geneva, 2008.

[10] ITU-T Rec. P.911. Subjective audiovisual quality assessment methods for multimedia applications. Int. Telecomm. Union, Geneva, 1998.

[11] P. Kortum and M. A. Sullivan. Content is king: The effect of content on the perception of video quality. In *Proc. of the Human Factors and Ergonomics Society 48th Annual Meeting*, volume 48, pages 1910–1914, Santa Monica, 2004.

[12] P. Kortum and M. A. Sullivan. The effect of content desirability on subjective video quality ratings. *The Journal of the Human Factors and Ergonomics Society*, 52(1):105–118, 2010.

[13] N. Staelens, S. Moens, W. Van den Broeck, I. Marien, B. Vermuelen, P. Lambert, R. Van de Walle, and P. Demeester. Assessing quality of experience of IPTV and video on demand services in real-life environments. *IEEE Transactions on Broadcasting*, 56(4):458–466, 2010.

[14] S. Winkler and P. Mohandas. The evolution of video quality measurement: From PSNR to hybrid metrics. *IEEE Transactions on Broadcasting*, 54(3):660—668, 2008.

E. Momentary Quality of Experience: User's Audio Quality Preferences Measured Under Different Presentation Conditions

Adam Borowiak[1], Ulrich Reiter[1], U. Peter Svensson[1]

[1] *Department of Electronics and Telecommunications, Norwegian University of Science and Technology (NTNU)*

Journal of The Audio Engineering Society, volume 62(4), pp. 235-243, 2014.

Paper E

Is not included due to copyright

Appendix

Key details of the experimental studies carried out in Papers A-E.

	Content (sourced from BD*)/ description[1]/ test material	Purpose of the study/ research questions	Participants/ test environment[6]/ procedures[7]	Method of data analysis/ key findings
PAPER A	*Challenges of Life; Life* by BBC/ (ND[1]); Animal life stories supported by human narration and emotional music pieces/ Duration: 30min55s; Video: HD[2]; Audio: WAV[3], 512 kbps; QP: 0,16,22,24,28,30,32,36,38,40,44,48	- Is the proposed method suitable for momentary quality assessment of long duration AV content? - Does the time dimension affect user quality expectations over time?	22P (8F, 14M; 7 experienced HD quality viewers)/ LRoom[6]/ training (10-15min), test (32min), and post-test session (5min); best start q, degradation every 3min and in steps	Average over raw data/ - the method produced plausible data indicating its suitability for a purpose of the experiment - averaged quality preferences were constant across stimulus duration - experienced participants on average set the quality to higher levels than naïve ones
PAPER B	*Challenges of Life; Life* by BBC/ (ND[1]); Animal life stories supported by human narration and emotional music pieces/ Duration: 30min55s; Video: HD[2]; Audio: WAV[3], 512 kbps; QP: 0,16,22,24,28,30,32,36,38,40,44,48 *Plants; Life* by BBC/ (ND[1]); Plants life stories supported by human narration and emotional music pieces/ Duration: 30min20s; Video: HD[2]; Audio: WAV[3]; AC-R[4]: 32,48,56,64,80,96,112,128,160,192,320 QP: 0,16,22,24,28,30,32,36,38,40,44	- Do the quality expectations decrease over time and with increased involvement in the content? - How fast can participants notice quality changes and at which quality level does this happen? - Is the quality level at which the change is noticed similar to the quality level set by the participant? - Is proposed method suitable to test stimulus with both modalities being distorted simultaneously?	T1[9]: 22P (8F, 14M; 7 experienced HD quality viewers)/ LRoom[6]/ training (10-15min), test (32min), and post-test session (5min); best start q, degradation every 3min and in steps T2[9]: 20P (13M, 7F)/ LRoom[6]/ training (10-15min), test (32min) and post-test session (5min); best start q, degradation every 3min and in steps	ANOVA and Kruskal-Wallis test/ - q expectations weren't changed over time ($F=0.91$; $p>0.05$) the same held for the reaction time to q changes ($F=0.69$; $p>0.05$) - higher sensitivity to q changes when users' self-controlled the q adjustment than when the q degradation process was controlled externally (based on a graphical representation) - the process of q adjustment was easier for participants when both modalities were distorted (less variation in the data) compared to only one and the q levels set by them were high on average - the method produced plausible data indicating its suitability for a purpose of the experiment
PAPER C	*Mammals; Life* by BBC/ (ND[1]); Animal life stories supported by human narration and emotional music pieces/ Duration: 32min2s; Video: HD[2], T1[9]: AC-R[4]: 32,48,56,64,80,96,112,128,160, 192,320; Audio: WAV[3]; T2[9]: QP: 0,16,22,24,28,30,32,36,38,40,44; Audio: WAV[3], 1411kbps T3[9]: AC-R[4]: 32,48,56,64,80,96,112,128,160, 192,320; Audio: WAV[3]; QP: 0,16,22,24,28,30,32,36,38,40,44	- Are user's quality preferences changing over time? - Is the proposed method suitable to test AV stimulus with only audio modality being distorted? - Do participants react differently to quality changes in an AV stimulus in case of audio/video only distortions and in case when both modalities are distorted at the same time?	T1[9]: 20P (14M, 6F)/LRoom[6] T2[9]: 10P (6M, 4F)/LRoom[6]/ T3[9]: 10P (5M, 5F)/LRoom[6]/ The same test procedure for T1[9], T2[9], T3[9] but with different modality being impaired in the test: training (10-15min), test (35min), and post-test session (5min); best start q, degradation every 3min and in steps	ANOVA/ - users' preferences regarding audio q were relatively constant over time when the content was of a relatively similar type - the method produced plausible data - deterioration of only the audio q in HD AV presentation influenced users' perception of q changes less than when only video impairments were present or when both modalities were distorted simultaneously ($F(1.243)= 73.714$; $p<0.0001$), ($F(1.243)= 93.315$; $p<0.0001$), ($F(1.243) = 32.239$; $p<0.0001$) - subjects reacted faster to q changes and preferred higher q levels when impairments were introduced in both modalities simultaneously
PAPER D	*Hunters and Hunted; Life* by BBC/ (ND[1]); Animal life stories supported by human narration and emotional music pieces/ SFA[5]: Duration: 9min; Video: HD[2]; Audio: WAV[3], 1411 kbps; QP: 0,16,22,24,28,30,32,36,40,44; *Lego Star Wars: The Padawan Menace*/ (CA[1]); A computer-animated comedy mainly with speech and special sound effects/ SFA[5]; *Transformers: Dark Side of the Moon*/ (AM[1]); A movie with fast motion scenes supported by dialogs and special sound effects/ SFA[5]; *The Phantom of the Opera at the Royal Albert Hall*/ (OP[1]); Opera singers performing on stage accompanied by orchestra musicians/ SFA[5]	- Does different type of long duration AV content affect perception of quality changes over time? - Does start quality (reference vs. worst) affect user quality preferences over time? - Does instantaneous catastrophic video quality degradation influence users' quality preference?	21P (14M, 7F),(MD=30,6; SD=8,2)/LRoom[6]/ training (10-15min), test 4x9min (36min), and post-test session (5min); worst start q, degradation every 90s (big q leap)	Kruskal-Wallis test/ - fast action scenes with many details may have caused distortions being less visible (or less annoying) after a while ($H(5) = 21.424$, $p<0.001$) - participants' q expectations were similar regardless of whether the reference q was introduced at the beginning of the test clip or not (based on a graphical representation and mean q level) - the way the degradations appear in the presented material (immediate drop from good to bad or stepwise in time) did not have an effect on the q level that viewers were satisfied with (based on a graphical representation and mean q level)
PAPER E	*Santana: Live at Montreux 2011*/ (JF[1]); Singers and musicians performing on stage. Relatively static performance with instrumental music/ Duration: 13min; Video: HD[2]; AC-R[4]: 32,40,48,56,64,80,96,112,128,192,320; Audio: WAV[3] *The Beyonc´e Experience Live*/(PC[1]); A pop singer performing on stage (dynamic music performance)/ Duration: 9min; Video: HD[2]; AC-R[4]: 32,40,48,56,64,80,96,112,128,192,320; Audio: WAV[3] *The Phantom of the Opera at the Royal Albert Hall*/ (OP[1]); Opera singers performing on stage accompanied by orchestra musicians/ Duration: 9min; Video: HD[2]; AC-R[4]: 32,40,48,56,64,80,96,112,128,192,320; Audio: WAV[3]	- Do the audio quality preferences (without an accompanying video) change over time? - Do the audio quality preferences become different when an accompanying visual stimulus is present? - Does the consecutive order of a presented material influence audio quality perception?	32P (13F,19M) (MD=32,7; SD=7,4)/ LRoom[6]/ training(10min), test (32min) and post-test session (5min), JF: T1[9]: AV, best start q, degradation every 110s and in steps; T2[9]: same but AS[8], PC: T1[9]: AS, worst start q, big q leap every 90s; T2[9]: same but AV, OP: T1[9]: AV, worst start q, 1 continuous clip; T2[9]: same but clip in 9, 1min pieces played at random	Kruskal-Wallis test/ - audio q preferences did not change over clip's duration regardless of AS[8] or AV scenario considered (T1[9]: $H(5)=4.27$, $p=0.511$; T2[9]: $H(5)=2.38$, $p=0.794$) - bit-rate levels set by participants in the AS[8] scenario were significantly higher than in the AV scenario ($H(1) = 43.631$, $p<0.001$) - perceptual fluency might have an impact on subject's audio q preferences ($H(1) = 12.385$, $p<0.001$)

[1]ND – Nature Documentary; CA – Computer Animation; AM – Action Movie; OP – Opera/Musical; JF – Jazz Festival; PC – Pop Concert; [2]HD – 1080p, 25fps, 16-bit YUV 4:2:0; [3]WAV – 2ch, 44.1kHz; [4]AC-R - audio compression rates (kbps); [5]SFA – same for all; [6]LRoom - a laboratory room following recommendations in ITU rec. BT.500, P.911 and BS.1116; [7]a method proposed in this thesis was used in all performed experiments; [8]AS – audio solely, [9]Ti – Test i ; q – quality

* The original source material (Bluray edition) is primarily intended for the home cinema scenario usage. This seems to confirm the ecological validity of the chosen content with regard to the purpose of the experimental design described in this thesis. However, it has not been verified if the content would have the same impact on the obtained results in a similar context but outside the laboratory setup.