

Chapter 2

State of the Art

2.1 Quality of Experience (QoE)

Multimedia contents, such as music or video for example, are around us constantly, and consumers are sometimes paying more and sometimes less attention to them. Especially when watching or listening to these contents as a primary task, the experienced quality is very important and may cause deeper engagement into the content when the quality is high. Stronger engagement may lead to more usage of a certain service, which may have conscious or sub-conscious reasons. These internal evaluations of quality are happening even inside naïve users. Sometimes customers come to the conclusion: this is ‘bad quality’. Lower quality is usually more conspicuous compared to high quality. Thus, customers are more obliged to change services when they experience bad quality. Obviously, this is highly dependent on the context in which they experience the multimedia content, and on the customer itself. Due to this, service providers need to ensure that they find an optimum tradeoff between the delivered quality and the used bandwidth. Therefore, they have to have a concrete value describing the average delivered quality. This value has been described as the *Quality of Experience (QoE)*. The *Qualinet* community, which is an international group working on several aspect of quality, defines QoE according to their white paper [3] as follows:

Quality of Experience (QoE) is the degree of delight or annoyance of the user of an application or service. It results from the fulfillment of his or her expectations with respect to the utility and/or enjoyment of the application or service in the light of the users personality and current state.

This definition is different from the definition given by the International Telecommunication Unit (ITU), as it moves the user of the application or service more in the focus. ITU-T Rec. P.10 (Amendment 2, 2008) [4] defines QoE as:

Quality of Experience (QoE): The overall acceptability of an application or service, as perceived subjectively by the end-user.

NOTE 1 Quality of experience includes the complete end-to-end system effects (client, terminal, network, services infrastructure, etc.).

NOTE 2 Overall acceptability may be influenced by user expectations and context.

In the case of analyzing the quality, a feature extraction is (sub)consciously performed by the test participant. A *quality feature* is defined according to [5]:

Quality feature is the perceived characteristics of an entity “that is relevant to the entity’s quality”.

The evaluation of the QoE can be influenced by context and user expectations as mentioned above. These factors can be summarized as influence factors. These are crucial when it comes to the perception of subjectively experienced quality. In [3], these *Influence Factors (IF)* are defined as: “Any characteristic of a user, system, service, application, or context whose actual state or setting may have influence on the Quality of Experience for the user.”

The influence factors can be divided in: Human (HIF), System (SIF), and Context (CIF) influence factors [6]. Here, HIF describe possible factors connected to the observer, such as social-demographic factors. SIF are factors coming from the system under test, and CIF are factors describing the environment the observer is situated in [3]. The most unpredictable component (usually) is the human. HIF factors are defined as: “A Human Influence Factor (HIF) is 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” [6]. Rather easy to obtain data from the HIF are e.g. demographics and the socio-economic situation. These are quite stable over a longer period. The emotional state or mental constitution, however, may play a more crucial role and are varying by form of the day or even situation, and are more difficult to measure.

The assessment of the emotional state can be performed subjectively, using the Self-Assessment-Manikin scales (SAM), as proposed by [7]. Here, the level of arousal, valence, and dominance are being assessed. Arousal describes the level of excitement (on a scale from bored to being excited), valence refers to the liking or level of happiness (ranging from unhappy/sad to happy), and dominance assessing the level of control (on a scale from being controlled to having the control over a situation). On the one hand, the emotion of the test participant may be important, and on the other hand, the presented material can provoke emotions. Although, current research is not arguing that emotions are influencing the experienced quality, it is still not clear in which way they do [8]. Thus, presenting emotionally neutral stimuli is preferred in a quality evaluating scenario.

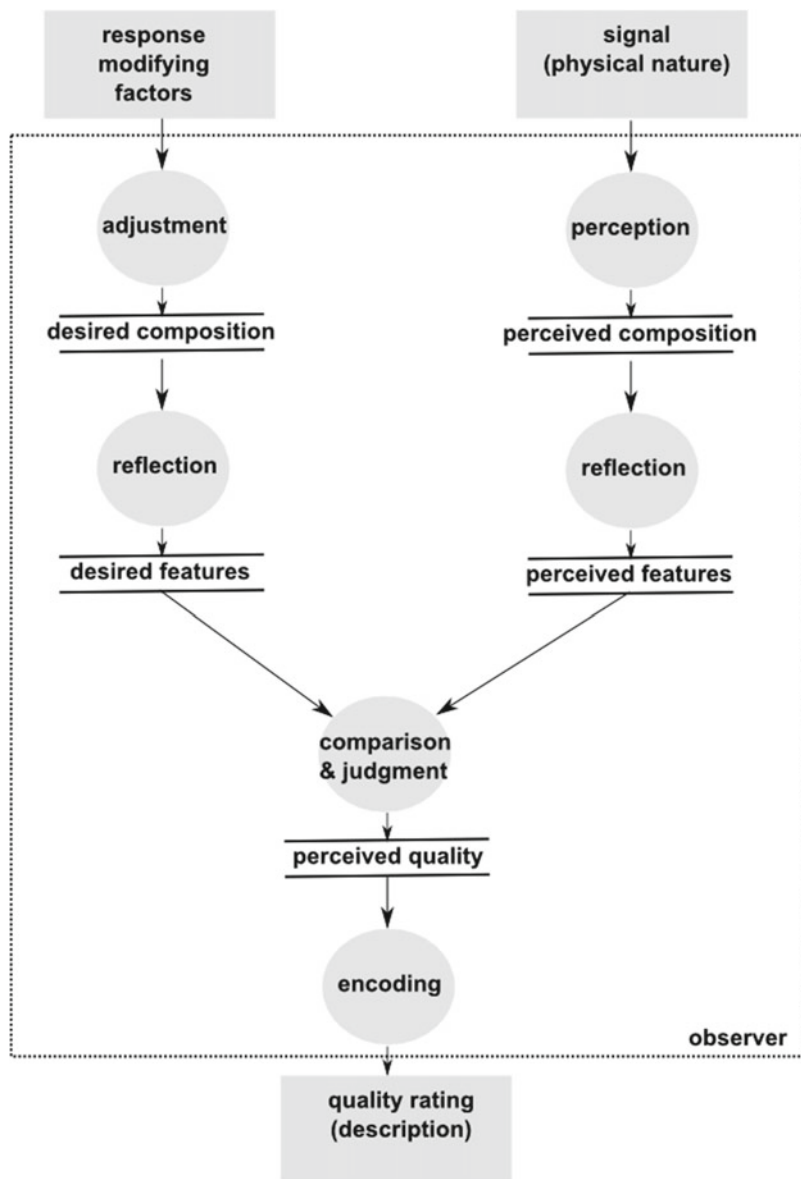


Fig. 2.1 Model for the quality perception and rating process, taken from [9, 10]

The model depicted in Fig. 2.1 describes the quality perception and rating process, derived from models developed by Jekosch [9] and Raake [10]. The model takes the *response modifying factors* which include the HIF, CIF and SIF into account, and it takes the presented *signal*, i.e. the actual stimulus. Based on these two, the

process of quality perception and rating begins. The result is the product of an internal comparison process, between the currently perceived quality features and the desired quality features. The external stimulus goes along the stimulus perception path, along the way the stimulus is *perceived* by the observer, and is being *reflected*, the *perceived quality features* are being extracted. These perceived quality features are *compared* to the *desired quality features*, and eventually lead to a *perceived quality*. The result of this *comparison process* needs to be *encoded* to the used *test scale*. Thus, all processes leading to this final evaluation are happening inside the test participant.

The initial assumption of the model is a comparison of the perceived stimulus with an internal reference. This eventually results in the quality description process. The human observer is basically assumed to be a black box and modeled with some states and processes derived from psychology. It remains unclear how these processes are exactly arranged and happen in the case of forming a quality description. Physiological measures may be helpful to gather more insights into this process, as they directly obtain reactions of the human due to external sensation.

2.2 Quality Evaluation Methods

2.2.1 Stimulus Material

To evaluate the experienced quality, subjective quality tests are being conducted. In these tests, a specific stimulus material is used which is selected following certain criteria. Entertaining movies are often longer than one hour, and episodes of series have a duration of several minutes. Therefore, these cannot usually be used for quality tests. Furthermore, they usually only show a small subset of possible scene arrangements. Additionally, they try to evoke specific emotional responses from the viewer. All these factors make it hard for subjective quality tests to use standard film materials, as testing time is limited and different technical settings have to be evaluated. Therefore, stimulus material specifically designed for quality evaluation tests is produced by researchers and is provided in databases. One commonly used database is the cdvl.org [11] database recommended by the video quality experts group (VQEG).

Uncompressed audiovisual material can only be obtained from professional cameras and high-quality microphones, and the outcome of those contains a vast amount of data. This cannot be handled by physical media or current networks. Also the outcome of semi-professional cams is too massive to be transmitted via networks. In order to reduce the complexity of data to a manageable level, video compression algorithms are necessary. The video stream is encoded on the sender side, and after transmission over the network, decoded on the receiving side. During encoding, decoding, or transmission, different errors on the video material can appear. These can be clustered into two classes: spatial and temporal artifacts [12]. Spatial artifacts in a video can be observed even in a still frame of the video, and temporal

artifacts can only be observed while playing the video sequence. The first class of distortions (spatial artifacts) contains: blockiness, blurring, slicing, color bleeding, staircase effect, and ringing. Temporal artifacts include: jerkiness, mosquito noise, and spatial artifacts lasting for several frames. A more exhaustive list and a more detailed description of distortions can be found in [13]. Obviously, codecs usually contain a combination of the aforementioned distortions. To quantify the effect on the user's quality perception when combining different degradation types, a reference impairment system for video (RISV) is proposed by ITU-T Rec. P.930 [14]. Here, the frame rate, blurriness level, edge busyness, blockiness and noise can be adjusted accordingly and is leading to the above listed artifacts. Furthermore, the distortions can be applied sequentially and therefore, a huge variety of degradations can be accounted for, using the RISV. In case of audiovisual stimuli, the described video degradations can be combined with the range of audio distortions. In case of audio three dimensions for audio quality were identified in [15]: coloration, noisiness, and discontinuity. In the auditory domain, a standardized method to produce distortions exists as well. It makes use of the modulated-noise reference unit (MNRU) [16] which adds signal-correlated noise on the signal in different intensities.

In addition to these described content variations, the video content can be described by its technical parameters. Here, spatial perceptual information measurement (SI) and temporal perceptual information measurement (TI) are used. SI describes the spatial complexity of a presented video, thus, if many or rather few edges and details are present in the video. This can be estimated by using a spatial Sobel filter. For example, a blue sky with almost no edges has a lower SI when compared to a video containing a scene in a nicely decorated room which has many more edges and therefore a larger SI. The temporal perceptual information measurement (TI) describes the temporal complexity of a stimulus: many changes within the scene result in a larger TI, whereas less changes result in a lower TI. Using both measurements, a spatial-temporal-map can be spanned.

Subjective quality tests are being conducted to gain insights into the subjective quality perception of a user. In these tests, different technical parameters are evaluated, such as certain compression algorithms, codecs, or technical setups. These perform differently towards the combination of different SI and TIs, also the viewer perceives quality differently in these scenarios. Thus, in subjective tests, it is desirable to have the SI-TI-map covered completely, or due to time constraints at least at specific points. These may vary depending on the goal of the experiment. In case of audiovisual material, this map has to be extended for a third dimension, namely for audio. In the case of speech, three perceptual dimensions have been identified: coloration, noisiness, and discontinuity [17]. For more detailed information on perception of speech quality, the reader is referred to [15].

In addition to the distortions which occur independently on each modality, effects of asynchronous audiovisual material can be analyzed. This describes an offset between the audio and the video track. This can often only be identified when a person is speaking, and audio as well as video are focused on this detail. The perceptual threshold for detecting asynchronous stimulus material is analyzed in these scenarios. Due to the human perception system the threshold to detect audio leading

video, and video leading audio is different, see [18] for more information on this. The ITU standard describing the corresponding parameters is the ITU-R Rec. BT.1359 [19].

Obviously, any combination of the mentioned distortions can occur and be analyzed in quality perception tests. Note that the combination of distortions in different modalities may also lead to a different evaluation of the distortion in one modality compared to an isolated presentation as it is initially shown in [20].

In order to quantify these impairments, subjective quality tests are being conducted. These tests typically aim at short duration of stimuli, i.e. approx. 10 s long, and try to represent a whole range of typical video sequences. Typically clips transporting a minimum of emotional content are chosen for these tests, as evoked emotions may influence the behavioral pattern of a test participant and thus, the quality judgment [8]. In order to accommodate for variety different scenes e.g. sports, news, or group conversations, are used (see scene content categorization in ITU-T P.910 [21]). In case of conducting audiovisual experiments, the described content has to be expanded towards several sound sources. Single or multiple speakers can be present, either as a background narrator or as people visually seen in the video material. Furthermore, these speech sources can be combined with background noise or music (see ITU-T Rec. P.911 [22] for more details).

2.2.2 *Subjective Quality Tests*

The quality is traditionally assessed using subjective opinion tests. Here, the International Telecommunication Unit (ITU) has proposed several standards on how exactly these tests have to be performed for each type of content (image, audio, video, audiovisual). These standards are summarized in the corresponding recommendations: ITU-T Rec. P.800 is used for audio quality tests, ITU-R Rec. BT.500 for image quality, ITU-T Rec. P.910 for video quality tests, and ITU-T Rec. P.911 for audiovisual quality tests. In the context of this work, only quality test methodologies for video and audiovisual content will be explained in more detail.

A result of these opinion tests is an averaged quality judgment over several participants for each tested condition. This value is the mean opinion score (MOS).

Generally, two different test materials have to be distinguished, namely, short stimuli which have a typical length of 10 s, and long stimuli, which can last several minutes.

For short stimuli, two methodologies can be distinguished: absolute ratings and reference ratings. For the absolute rating, the absolute category rating (ACR) is employed. In this case, one stimulus is presented to the test participant which has to be rated after its presentation (Fig. 2.2a). Hereby, the rating is conducted on a discrete scale with the labels ‘excellent’, ‘good’, ‘fair’, ‘poor’, and ‘bad’ (Fig. 2.2b). If a finer grading is desired by the experimenter, a 9-point or 11-point scale can also be used, which depends on the needs of the experiment.

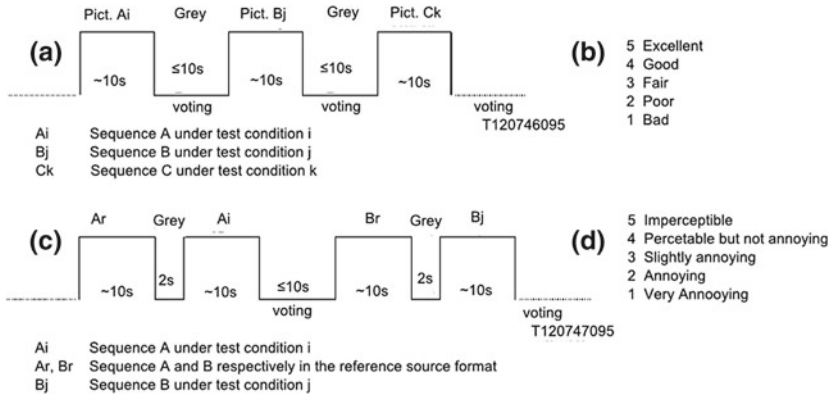


Fig. 2.2 Illustration of (a) ACR test procedure and (b) corresponding rating scale, as well as sequence of (c) DCR test procedure and (d) their corresponding scale. Figures taken from ITU-T Rec. P.910 [21]

In the case of presenting stimuli in pairs, as it is done in reference ratings, different experimental designs exist. Two of the most commonly used rating methodologies are degradation category rating (DCR) and pair comparison method (PC). Using PC, two consecutive videos are shown and the test participant has to decide which of the two they prefer. In an extended test setup it is also asked to quantify this perceptual quality distance of the two presented conditions. In the DCR test scenario, the video presented first is always the reference video (which is also known to the participant) and the second one is the stimulus under test (Fig. 2.2c). Here, the rating is performed on a scale using the labels: ‘Imperceptible’, ‘Perceptible but not annoying’, ‘Slightly annoying’, ‘Annoying’, and ‘Very annoying’ (Fig. 2.2d). These are usually shown on a discrete scale. How many steps this scale has, again depends on the needs of the experiment, as in the ACR case. Furthermore, when presenting purely visual stimuli, a presentation of both conditions simultaneously (e.g. on two screens) can be an option, using either the PC or DCR method. This is more time saving while conducting the experiment, as two conditions can be tested at the same time, and might lead to a finer quality differentiation. Due to the more complex setup, especially when using moving videos, it might be only suitable for non-novices in quality evaluation tasks [23]. For audiovisual material, the quality judgment is assessed individually for both modalities, and as an overall audiovisual quality judgment. The ratings can be either obtained in an absolute quality evaluation scenario, with the corresponding ACR scale, or in the PC or DCR scenario. It is left unspecified, in ITU-T P.911, in which order the scales are evaluated.

When using longer stimuli spanning several minutes, ratings can be obtained continuously. The test participant is given a slider with which they can communicate their experienced quality instantaneous to the current quality of experience. The methods to be used are described for speech in [24], called Continuous Evaluation of Time Varying Speech Quality (CETVSQ), and for video in [25], called Single

Stimulus Continuous Quality Evaluation (SSCQE). The main idea in both setups is that the test participant has a slider with a scale which they can adjust throughout the stimulus is played. These adjustments should be performed as an instantaneous reaction towards a perceived quality change in the presented material. The scale has labels that are the same as used in the ACR method. This method has the disadvantage that the test participant has to avert their gaze from time to time in order to know on what specific position the slider currently is or where to push it to. To overcome this, different input devices to rate the quality e.g. a steering wheel, joystick [26], or a glove [27] have been proposed. A different approach was suggested by Borowiak et al. [28]. Here, the test participant is adjusting the quality instead of evaluating it. Consequentially, whenever a quality change is introduced by the testing system, participants have to readjust the quality with the help of a knob, until they perceive again optimal quality.

All the described methodologies were developed for passive audiovisual scenarios, hence, no direct interaction of the participant was desired. Recent work shows that also interactive scenarios need such quality evaluations, such as e.g. video calls. For this area, different subjective tests and materials have to be used. The current work is only using passive scenarios, therefore, the interactive part will be excluded. For deeper insight into the latter work, the interested reader is referred to [29].

2.2.3 Instrumental Estimation

Conducting subjective quality tests is not always possible, since these tests are time and money consuming, as outlined above. Different quality prediction algorithms have been developed, to estimate the perceived quality. This is especially helpful for existing networks or services, as here the delivered quality cannot (easily) be evaluated using subjective test methods, as it usually would annoy the customer to give a rating of the used service. The quality prediction algorithms can be divided into three sub-categories, based on the input information they require:

- *parametric*: these rely solely on descriptive parameters of the material and/or network; neither the original signal, nor the processed signal is available for quality estimation (e.g. TV-model [30])
- *non-intrusive signal-based (no-reference)*: only the output signal is available and all estimations are made based on this signal
- *intrusive signal-based (full-reference)*: the input and output signal are compared to each other and the difference is reported (e.g. PEVQ [31])

These three basic mechanisms of models exist in the speech domain as well as in the video domain. To predict the audiovisual quality, the *quality-based model* can be used, as it is introduced in [32]. Here, the individual quality scores for both modalities are included and the overall audiovisual quality is predicted based on the individual quality scores. General descriptions and more details for instrumental quality estimation for video services can be obtained from [33].

2.2.4 Summary

This section described the current state of the art and challenges in conducting subjective quality tests using audiovisual stimuli. One of the major drawbacks with all these subjective assessment methods is, they are not gathering insights into underlying physiological responses of the test participant. The final rating of a test participant is eventually the end of a process which is described in [9] and the model in Fig. 2.1. To better understand these processes, measurements of physiological responses can give more detailed insights into those processes. Additionally, measures of neural activity derive the response directly from the source, i.e. the brain, and eventually lead to a better understanding of quality perception and their ratings.

Furthermore, quality prediction models rely on conducted subjective quality tests. Even better estimation of the perceived quality may be obtained when employing models that are based on physiological responses. The predicted quality scores may be more precise because they incorporate the current characteristics, e.g. mood, of a user into the predicted quality values. Furthermore, they may help to differentiate between reliable and unreliable test participants.

The obtained judgments are also very much dependent on the used content in the experiment, as mentioned in the beginning. They rely on a quality judgment obtained subjectively from self-assessed judgments, as outlined. The reactions towards a stimulus can be captured directly from the participant using physiological measures. Therefore, the encoding step in the model from Fig. 2.1 can be omitted and a less biased judgment may be obtained. Furthermore, a better understanding on how the quality judgment is being formed internally can be derived.

2.3 Physiological Measures

Physiology is the scientific study of functions in living systems. It is concerned with the physical and chemical processes of organs, cells, and bio-molecules. Physiological measurements can be monitored from the human using different measurement devices. These can be categorized into different classes, an overview of different physiological measures can be seen in Fig. 2.3. This chart is not supposed to be complete but should give an overview of measurements which will be used within this work. These physiological measures are obtained from the human body and are roughly divided into three subcategories: neurophysiological, peripheral physiological, and movement capturing methods. All reactions of the body, whether these are voluntary or involuntary, are based on neurophysiological reactions.

Two basic methodologies can be considered for neurophysiological recordings: imaging and electrical techniques. Medical imaging is picturing the anatomy of internal structures. For example fNIRS (functional near-infrared spectroscopy) reveals the blood flow of oxygenated and deoxygenated blood and therefore, lets researchers draw conclusions about which parts in the brain are active during certain processes.

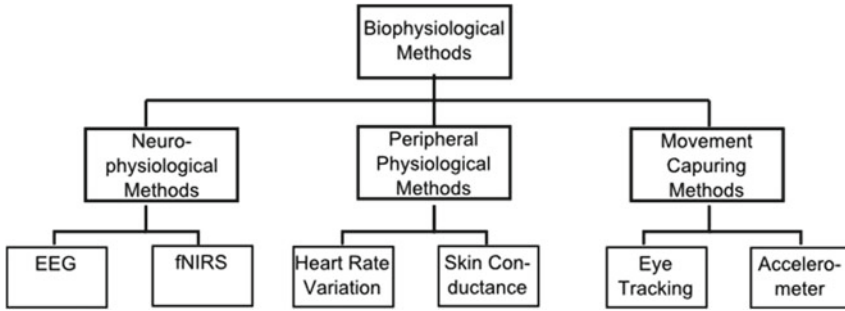


Fig. 2.3 Overview of physiological measures

It is assumed that areas which are active, should induce higher concentrations of oxygenated hemoglobin, and lower in deoxygenated [34]. The technique of fNIRS is based on absorbed near-infrared light which is emitted and collected by diodes, placed on the scalp's surface. It is depicting the blood flow and therefore, it has a very high spatial resolution. However, it has a rather poor temporal resolution due to the fact that the blood flow is the reaction of firing neurons and therefore, a physiological reaction lags up to 8 s.

EEG (electroencephalography) and MEG (magnetoencephalography) belong to the recording techniques which are based on direct neural reactions and therefore, can be observed using tools for capturing electrical activity. In contrast to NIRS, EEG has a high temporal resolution, as brain responses are measured directly from the scalp's surface, but it has a rather poor spatial resolution. The main focus of this work is on EEG, therefore, the details will be explained in Sect. 2.4.

The second class mentioned in Fig. 2.3 are peripheral physiological recordings which are obtained from the peripheral nervous system (PNS), the part of the nervous system which does not belong to the central nervous system, including brain or spinal cord. This includes among others:

- *Heart-rate (variability) (HRV)*: using electrocardiogram, measures electrical activity which is due to the heart beat, applied usually between common carotid artery and arm artery. Fundamental information on cardiovascular physiology and its basis can be found in [35].
- *Blood Volume Pulse (BVP)*: using photoplethysmograph, measures the change in blood volume, is usually applied at the index finger [36].
- *Skin conductance (EDA)*: using electrodes which are put on either arm, hand, finger, or foot which measure the change in conductance between the two points. Theoretical foundations of EDA can be found in [37].
- *Electromyography (EMG)*: measures muscle tension between two applied electrodes, can be measured basically anywhere on the body [38].

All these measures can give information about the current physiological state of the human. Emotions, among other things, can be classified based on these parameters, an interesting overview on emotions using peripheral measures can be found in [39].

Usually it is easier and cheaper to obtain these measures from the human, compared to neurophysiological measures.

The last class mentioned in Fig. 2.3 are reactions which result in movement, either of the whole body or only in parts of the body. These can be captured when e.g. an accelerometer is attached to the desired extremities. How still a person is sitting can give information about the current level of their immersion [40]. Additionally, eye movements can be recorded throughout an experiment. Eye tracking can be useful in visual scenarios where the analysis of areas of interest is important. The length of saccades or pupil dilation can give information about the level of attention or the cognitive state, among other things [41]. Another possible method to obtain rough information about eye movement is capturing an electrooculogram (EOG). Here, two pairs of electrodes are applied to the test participant, one being attached above and below one eye to record horizontal eye movements, i.e. saccades. The second pair is attached to the outer canthus of each eye [42]. EOG gives information about eye blinks and large horizontal eye movements. In some experimental setups such as auditory studies, it might be sufficient to obtain only these rough eye movement data as it is not of interest where exactly the test participant is looking at.

In order to understand underlying processes of behavior, the study of neural correlates is necessary. Here, especially the area of electrophysiology is interesting, as it studies directly the electro-chemical transmission within the nervous system. Hemodynamic measures such as NIRS are one opportunity to gain insights, but it takes a few seconds until a change in blood flow (due to oxy- and deoxygenated blood) can be observed. Using EEG, a response towards an external stimulus can be observed immediately. Another advantage is that the used apparatus is rather easy to apply and gives a more natural setting than e.g. using MEG which is very spacious and loud when operating which makes it almost impossible to perform audiovisual tests. Therefore, electroencephalography (EEG) is the method of choice in this work; the fundamentals of EEG are lined out in the next Sect. 2.4.

2.4 Electroencephalography (EEG)

Electroencephalography (EEG) measures electrical potentials as they occur on the scalp's surface. Generally, EEG can be obtained noninvasively from the scalp's surface, or invasively using implanted electrodes (intracranial EEG, iEEG). Only the former will be used and described throughout this work.

Neural activity in the brain is based on physico-chemical processes. Their recordings in the case of using EEG are based on electrophysical processes. The obtained electrical potentials originate from neural activity inside the brain. There are billions of neurons and synapses which have an electrical field around them, if enough neurons are activated this electrical field can be measured on the scalp's surface. To draw conclusions about the exact area the activity is originating from, requires solving an inverse problem (i.e. to draw conclusions based on an observed measurement towards information on their processes), this does not have a unique solution. Thus, based on

knowledge of the internal structures of the human brain, conclusions may be drawn using methods like LORETA (low resolution electromagnetic tomography) in order to calculate approximately where activation is coming from [43].

The recorded oscillatory neural activity is a response from the central nervous system (CNS). The possibility to perform these kinds of recordings, was first discovered by Berger in 1929 [44]. The analysis of the recorded signal is mainly divided into two parts: on the one hand, there is the spontaneous EEG, which is measured continuously and represents the current level of neural activity. With the help of this, assumptions about the mental and cognitive state can be drawn. On the other hand, event-related potentials (ERP) can be measured. These are reactions towards an explicit external event, and are linked in time with the start of this event (i.e. time-locked). In the following, both concepts will be described in more detail, and their corresponding features which are used in this work are explained extensively.

The recorded EEG signal can also be analyzed for its spatial distribution. Although, it may not be completely clear where the activity is exactly originating from the measured voltage differences, as they occur on the scalp's surface still give enough information on rough spatial distribution. Therefore, spatio-temporal maps are employed that show the measured activity at each electrode accumulated over a certain time. Coherence is one possible spatial analyzing technique and is used to draw conclusions about synchronization activity between two electrodes [45].

Standard electrode positions exist to ensure that same or at least similar sites are recorded between different test participants. Here, the 10–20 system has been introduced by Jasper [46]. It describes the distance between two adjacent electrodes such that these are 10 and 20 % respectively apart of the front–back or left–right distance on the skull.

For EEG recordings, often a reference electrode is being placed. This is to measure all electrodes versus one reference point. Therefore, often the tip of the nose is used or the mastoid, which is the bone behind the ear. At both locations, no brain activity is to be expected and they will not have any large artifacts as they cannot (or only hardly) be moved.

2.4.1 Event-Related Potential

First recordings of event-related potentials (ERP) were performed by Davis in 1939 [48]. ERPs are the response of the human brain towards external events, or are preceding a movement before it is performed by the human (so-called motor events) [49]. The ERP is a time-amplitude signal, and consists of several components, see Fig. 2.4 for an example of an ERP and a selection of components. These components are representing different underlying processes. The amplitudes and latencies of those components depend on several factors. The recorded activity is very small in amplitude which is in the range of a few microvolts. Due to this, the signal is very noisy. Hence, in order to obtain a reliable ERP curve, a certain number of repetitions is necessary. The number of repetitions is depending on the complexity

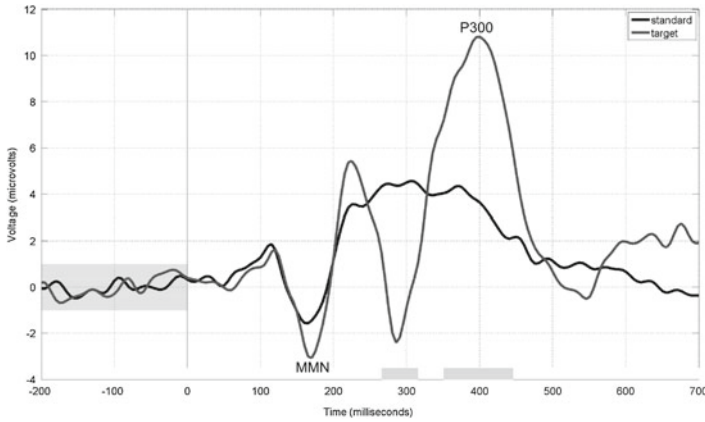


Fig. 2.4 Example of an event-related potential (ERP). Using an auditory oddball paradigm, the figure is taken and adopted from [47]

of the experiment and what explicit features are of interest. For simpler setups, there should be at least 40 repetitions per condition [49]. The more complex an experiment gets, the more repetitions are desirable. For analysis, the recorded data for each condition and participant is averaged.

The summed averaged signal is also called grand average. The signal may additionally be contaminated with unwanted components, therefore, artifact rejection is typically performed before analysis. This can be done e.g. on a single trial. Therefore, a certain threshold is being determined, and if the signal exceeds this threshold, the trial will be excluded from further analysis as this trial is suspected to contain movement artifacts. Another artifact rejection technique is based on independent component analysis (ICA). Here, individual components over the entire recording time are being identified. These components can either be related to actual cognitive processes or be due to body and eye movement. The technique of ICA is often used to remove eye movement artifacts. Horizontal movements and eye blinks will be identified individually as they are composed of two different ICA components [50].

A specifically developed test paradigm in ERP research is the so-called oddball paradigm [51]. Here, a sequence of standard stimuli is presented to the participant, the cue is interrupted from time to time by deviant stimuli (also called target stimuli), these cues can either be audio, visual, or audiovisual ones. One of the earliest and best studied component of an ERP is the miss-match-negativity (MMN). The MMN, a negative deflection after approximately 200 ms, is elicited in an auditory oddball paradigm and triggered when a deviant stimulus is presented [52]. The MMN is also elicited when a distraction task is introduced, and the detection of the target stimulus is not the primary task of the participant.

Another very well researched feature is the P300 component (also called P3), which is a positive deflection after about 300 ms. In general, the P300 is concerned about feature extraction of a stimulus, and is the result of a test participant's reaction

towards a stimulus [53]. It is elicited when a deviant tone is presented in a series of standard tones, as in the oddball paradigm. The P300 amplitude depends among other things on the probability of the presented deviant (the less occurrences the bigger the amplitude). Furthermore, the complexity of the test setup (e.g. level of difficulty, primary/secondary task) is a factor that can vary the P300 latency and its amplitude [54]. In addition, the latency depends also on the amplitude; the larger the amplitude is the earlier the P300 peaks. The P3 can be split into a P3a and P3b. Here, P3a is rather related to features of the stimulus and its activation is more frontal. P3b is rather task-related and activated areas are more temporal-parietal [55].

One of the later components of an ERP is the N400, a negative component approximately 400 ms after stimulus onset. This component represents already first indications of semantically processed information. The N400 is only triggered when semantically incorrect information are given [56] which is the case for auditory stimulation and for visual representations [57]. Even later components are ascribed for higher cognitive processing and will be disregarded for this work.

2.4.2 *Spontaneous EEG*

In contrast to the ERP, which is time-locked, the spontaneous EEG, also called continuous EEG, represents the current neural level of activity which is measured over a certain amount of time. During frequency analysis, its power is calculated by applying Welch's method which is an approach to spectral density estimation and uses the concept of periodogram spectrum estimates and computes the discrete Fourier transform. The obtained frequencies can be divided into several sub-bands: delta (0.5–4 Hz), theta (4–8 Hz), alpha (8–12 Hz), beta (12–20 Hz), and gamma (30–70 Hz). To each of the bands a certain functionality is being ascribed [58, 59]:

- *delta* : occurs during deep sleep
- *theta* : drowsiness and impaired information processing
- *alpha* : drowsiness and relaxed wakefulness
- *beta* : emotional and cognitive response
- *gamma* : arousal and short-term memory matching

As a result of a frequency analysis of such a complex signal, a distribution of spectral energy is obtained. To analyze those in more detail, the relative change of spectral energy in those frequency power bands is calculated. This is typically done in comparison to some known baseline (condition). Sometimes it is necessary to analyze one of those sub-bands more in detail, therefore, those sub-bands are again sub-divided. This is often done for example for the alpha band, resulting into lower (8–9.5 Hz) and upper (9.5–12.5 Hz) alpha bands. Lower alpha is more relevant for task-unspecific and stimulus-unspecific activity, and power in the upper alpha band represents rather task specific processes [60].

Generally, the alpha band is most interesting when it comes to drawing conclusions concerning cognitive states, such as fatigue, as it is shown in [61] in the context of driver's fatigue, and in [62] for effects of mental fatigue. Here, the alpha band power increases when test participants tend to get more fatigued. Although, there is a fixed band defined in general EEG literature, it has been shown that identifying individual alpha bands (IAB) per subject may be the more accurate way to analyze frequency band powers. In [63], a 2 Hz band around the maximum alpha peak was chosen for each participant. Due to individual difference of the human brain the calculated IABs may be more precise than the broader alpha bandwidth.

It is not only important to analyze if an increase within a certain frequency power band occurred, but also the location is of this effect is of interest, as different locations are due to different processes. For example, an increase of alpha activity in the occipital and parietal areas is a strong indicator of increased level of fatigue. Whereas, an increase in only one of the the frontal hemispheres goes along with a change of the emotional response of a test participant [64]. Similar assumptions apply for the theta power band, as an increase in the occipital area is corresponding to a drowsiness and impaired information processing [58].

2.4.3 Summary

Neurophysiological activity builds the foundation for all conscious and non-conscious behavior of humans which can be measured fairly well using EEG. EEG tries to measure a specific response out of a big noise-chunk, thus, the actual source of activity can only be approximated using time consuming algorithms such as LORETA. For the purpose of this work, source approximation is not of interest, but to understand generally the neural reactions which are due to the change of quality of the presented stimulus material.

In this section, it was shown that the analysis of EEG can be done either in the time-amplitude domain or in the time-frequency domain. Here, either the response towards a specific event is measured (ERP) or the level of neural activity is analyzed using frequency band analysis. The appropriate analysis depends on the test paradigm, and the goal of the study. Both techniques will be used within this work, within their respective scopes.

2.5 EEG in QoE

While at the time that this project was initiated only very limited research had been conducted in the area of QoE and EEG and was mainly performed by Antons and Porbadnigk, now more and more research institutes are conducting studies in this area of research. This is mainly due to the fact that some low-cost EEG devices have become available. In contrast to these customer devices, traditional EEG systems

that are coming from the clinical area are quite expensive. Although, these have a much better signal-to-noise ratio and therefore, the recorded components can be identified much better. Nevertheless, research in the area of EEG and QoE is still rarely distributed. Research conducted in the sub-areas of QoE using EEG or other neurophysiological measures is described in this section.

2.5.1 *Audio*

Very first experiments using signal-related degraded audio files and measuring brain responses were made by Miettinen et al. [65]. They used magnetoencephalography (MEG) and presented the test participants with a low resolution audio file while they were recording brain activity. MEG measures noninvasive the magnetic activity in the brain. They could show that the auditory evoked magnetic field significantly increased in the trials with distorted audio.

First concepts have been introduced by Antons and Porbadnigk to the area of speech QoE using EEG. Therefore, a paradigm which is a classical approach in EEG research was used, namely an oddball paradigm. First studies were using short utterances such as the German phoneme /a/. The presented speech files were partially contaminated with noise. Noise was introduced in a post-processing step using the modulated-noise-reference-unit (MNRU) recommended by the ITU [16]. The presented experimental design was such that the clear and undistorted audio file was the standard file played to the test participant, and files with varying SNR levels were presented as deviants in an oddball paradigm. Four SNR levels were chosen, each with 6 % probability. The participant's task was to answer after each trial whether the last file played was degraded or not. Individual noise levels around the perceptual threshold were determined for each test participant during a calibration phase, preceding the main experiment. It was found that the lower the signal-to-noise ratio was, hence the lower the quality was supposed to be, the larger the P300 amplitude was and the earlier the component reached its maximum (latency) [66]. In an extended study, words instead of phonemes were presented. The same experimental setup was used and similar results were obtained, thus, the worse the quality of the presented audio file was, the earlier and larger the P300 component appeared [67]. In a later classification processing step, it was found that trials where participants did not report to notice any degradations in the signal, the recorded EEG was classified as it was noticed by the participant. Hence, EEG showed to be more sensitive than behavioral answers in this case [68, 69]. These results were also brought to the ITU in order to make more parties aware of this research field, which may bring more objective results to the area of QoE [70, 71].

Later longer stimuli were used, and it was analyzed how the exposure of longer degraded audio files is influencing the mental state of the participant. Here, an audio book was presented to the test participants. From the EEG recordings, the synchronization activity was analyzed, performing a frequency band power analysis. It was shown that the calculated alpha band power was larger in cases of the low quality

audio sequences compared to the better one. The same was the case for the theta band power [72]. Both frequency power bands are indicators of fatigue and drowsiness of the test participant. Thus, Antons et al. conclude from their studies that due to the exposure to low-quality stimuli participants get more fatigued compared to the exposure to high-quality stimuli. In a second study, they analyzed the level of fatigue when applying two different quality profiles to an audio book. They could show that a constant low-quality version of the audio resulted in higher fatigue compared to the case when a variation of quality was introduced. Thus, a higher-quality version within a low-quality sequence is more desirable, when both profiles have in average the same bit rate [73].

Thirdly, auditory stimuli with length of sentences (i.e. approx. 10s long), and therefore, more conform to ITU standards, have been used in several setups. In [74] Antons used a recording of a sentence and applied different levels of reverberation time. In [75], the author of this work used synthetic speech samples with the length of sentences and could show that the P300 elicited is significantly larger for badly synthesized speech samples than for better ones. When analyzing the frontal alpha band power in the left hemisphere, a decrease could be observed which is most likely due to disappointment with the presented quality level [76]. In another study, a variety of different natural sentences was used, and signal-correlated noise was introduced as a distortion. It was again shown that when participants evaluated the quality of the presented stimulus as bad, a larger P300 amplitude was found [77].

These results have been brought together and summarized in [78], where Antons furthermore describes the quality formation process when using EEG for auditory stimuli, adapting the model introduced in Fig. 2.1. In his work, he is modifying the model as he is measuring the ERP between the process of *comparison and judgment* and the *quality event*. The ERP consists of several features, some belonging to non-conscious and some to conscious processing. Furthermore, he is introducing the *participant's state* which in his model is not connected explicitly to any of the other components.

Different work conducted by Creusere describes a study using audio sequences with varying quality [79]. Here, on the one hand, they had trials in which participants rated the quality on one set of stimuli, and on the other hand, a different set of trials was used during which participants did not give any rating. Latter ones were used for classifying the trials into either distortion level, based on frequency band features. However, the study described in the paper consists only of very few test participants and the used methodology is not very well described.

Gupta et al. used the neuro-imaging method of functional near-infrared spectroscopy (fNIRS) to assess the quality perception of participants listening to audio samples. In contrary to EEG, fNIRS has a much higher spatial resolution but lacks precision in the time domain. In this study, it could be shown that synthetic speech, which is evaluated worse subjectively, results in a higher deoxygenated blood flow in the prefrontal cortex [80].

2.5.2 2D-Visual

Shortly after the start of this work, Lindemann et al. used correlates of EEG to estimate the quality of still images. A slightly modified oddball paradigm was used, similarly to the auditory studies. Images with undegraded quality and six versions using JPEG compression, were presented. The results show that images which were stronger compressed led to a larger P300 peak amplitude compared to less compressed images [81].

Scholler et al. used synthetic recordings of water rings and a chess grid background and modified the quality using a codec similar to HEVC (high efficiency video coding). They found a more pronounced P300 with stronger degraded videos. After the experiment, they performed an offline classification. The obtained results show a good classification when hits were classified versus reference conditions. However, when classifying misses the results were rather poor [82].

Mustafa et al. used low-complexity videos, and manipulated a video scene in such a way that a person walking in the video appeared blurry, popping out, or ghostly. They show that analysis of band power is different for the different kinds of distortion, and that they were able to classify the trials on a single trial basis concerning the applied distortion type [83].

Moldovan et al. conducted a study in which they manipulated different video scenes with a change in either bit rate, frame rate, or resolution, and let the videos play long enough to conduct a study similar to the SSCQE paradigm. During their experiment, they recorded EEG activity using the Emotive EPOC system. Here, they used the frustration values provided by the system as a validation tool for the obtained MOS ratings [84].

The Emotive EPOC system provides parameters of engagement, boredom, excitement, frustration, and a mediation level. These values are calculated internally based on the measured activity. The main issue with these parameters provided by the system is that it is not clear how these values are being calculated and therefore, replication with different devices is difficult.

2.5.3 3D-Visual

3D TV has been discussed controversially in the public as people report of subjective drowsiness or sickness when watching 3D TV. Some research has been performed in the area of 3D versus 2D TV perception which concludes that 3D leads to more fatigue than 2D [85].

There has been done only very limited research on using EEG in the 3D context. The most common questions researchers are after, is what visual discomfort of 3D stereoscopic displays means, and how this is reflected in physiological responses. This is commonly done in a comparison between 2D and 3D displays. Work conducted in [86] showed that a much higher inter-subject correlation of neural networks

was found in the case that participants watched 3D contents compared to 2D. Additionally, subjective reports suggested more immersion in case of 3D compared to 2D. Furthermore, EEG data could be classified to the corresponding class of stimuli and level of immersion.

Kroupi et al. present a study where 2D and 3D material was presented. For both technologies, stimulus material in a high-quality and a low-quality version was available. They could show that the low-quality version is affecting the cognitive state in both versions (2D and 3D), in such a way that the frontal alpha band power is indicating a rather positive response for the high-quality contents [87].

2.5.4 Summary

This section gave an overview on what has already been done in the area of QoE and physiology, and more particular using EEG. It can be seen that most of the work started only when the work at hand was already in progress. The monetary expenses to purchase an EEG system, especially with clinical standards, are quite high. Also the complexities, in designing studies and analyzing recorded data while using physiological techniques are very high. Consumer-grade products have entered the market, such as Emotive EPOC or Neurosky. Some of the practical problems with those are that no individual electrode layouts can be used. In addition, they drive the approach of one size fits all, which is particularly difficult when having participants with small heads. Furthermore, the data quality lacks on the one hand, due to not perfect sitting caps, and on the other hand the electrodes and amplifier are not of very good quality especially compared to clinical setups.

Most of the labs which conducted studies in the area of QoE started using simple stimuli. Besides Antons et al. all other labs used only short duration stimuli and analysis of ERPs. The common finding was when using correlates of ERPs, that the stronger a stimulus was distorted, the larger the P300 peak amplitude was. How this short-term effect is affecting the users' state in rather longer sequences was examined in two studies by Antons. It was shown that participants tend to become more fatigued for low-quality versions. So far, work has been performed in most areas which are covered by the classical QoE, including this work: speech, imagery, and audiovisual.

Some studies also used values which were pre-defined from the EPOC system, that represent emotional states. These values rely on frequency analysis. However, as it is not clear how they are being calculated, therefore, not too much reproducibility with other systems and setups should be expected.

2.6 Summary and Open Questions

This chapter gave first an overview on what Quality of Experience (QoE) is, why it is important to conduct studies on quality perception, and how to obtain subjective quality ratings. In addition, an overview of typical degradations for audiovisual material

was provided. Afterwards, an introduction into basic physiological data acquisition was given, with emphasis on EEG as this is the method of choice in this work. Here, both main analysis techniques were presented, namely ERP and frequency band power analysis. Finally, an overview was given on what has already been done in the area of QoE using neurophysiological methods, and especially EEG.

From the overview given in Sect. 2.5, it is shown that for the auditory domain a coherent series of studies and designs was derived and explored to gain insights in the quality perception of auditory stimuli. In contrast, in the visual domain only a few studies conducted by different labs have been performed. Thus, one of the objectives of this work is to fill this gap, and present an elaborated series of studies and designs in order to support findings by Antons [78], and transfer the method to audiovisual stimuli.

In Fig. 2.1, one of the current models for quality perception and evaluation was presented. This model will be used and adapted to the specific needs throughout this work. Based on a series of experiments presented in Chap. 3, it will be introduced and discussed how in particular the P300 component can be assessed. The questions which will be answered here is if the measured EEG response, i.e. P300, is dependent on the quality of the distorted stimulus. Furthermore, what influences does (un)distorted accompanying audio have with (un)distorted video. Thus, it will be clarified whether the assumptions drawn by Antons [88] are also applicable for multimodal stimuli.

In Chap. 4, the need of introducing the mental state into the model will be laid out. It will be derived how the perceived quality is influencing the participants mental state. In particular, spectral power analysis of the recorded EEG will be conducted to analyze this. This part will be based on two conducted experiments. Given the result from Chap. 3 that short-term influences on quality can be measured within the EEG signal, the question which arises is what are the influences on long-term degraded audiovisual stimulation.

Finally, in Chap. 5 two studies which are more conform with standard subjective quality evaluations are presented. In that chapter, especially the drawbacks and challenges of using EEG in the domain of quality assessment will be discussed. This includes discussions on the question if and how EEG can be used as a complement to standard subjective quality ratings.

In the last chapter, Chap. 6, the reported results will be embedded in the general research area, and an outlook for future work will be given.

Neural Correlates of Quality During Perception of
Audiovisual Stimuli

Arndt, S.

2016, XIV, 88 p. 31 illus., 19 illus. in color., Hardcover

ISBN: 978-981-10-0247-2