Signal Processing: *Image Communication* 42 (2016) 120–129

Contents lists available at ScienceDirect

Signal Processing: *Image Communication*

journal homepage: www.elsevier.com/locate/image

Using electroencephalography to analyze sleepiness due to low-quality audiovisual stimuli

Sebastian Arndt*, Jan-Niklas Antons, Robert Schleicher, Sebastian Möller

Quality and Usability Lab, Telekom Innovation Laboratories, TU Berlin

A R T I C L E I N F O

Article history:
Received 9 September 2014
Received in revised form
21 January 2016
Accepted 21 January 2016
Available online 12 February 2016

Keywords:
Quality of experience
Audiovisual
Physiology
Cognitive state
Alpha frequency band power

A B S T R A C T

Standardized methods to assess the quality of experience (QoE) of multimedia focus typically on short sequences of approximately 10 s and a subjective judgment of the test participant. Two main problems occur when using this methodology: On the one hand these short sequences do not represent the typical media usage, and on the other hand it is still not completely understood how these subjective ratings are formed within the test participant. To overcome the second issue and to gain insight into the internal processing, electroencephalography (EEG) has been introduced successfully to the domain of QoE. As it is possible to use EEG to assess the quality perception continuously, we present two studies using standard documentary audiovisual clips with a length of 40 min and 60 min as stimuli. During the presentation of these quality-wise manipulated test sequences, we record an electroencephalogram and other physiological measures. We show that features of the EEG recordings indicate a change in the cognitive state of the test participant during the exposure to low-quality compared to high-quality sequences.

1. Introduction

The quality of multimedia services has been enhanced substantially in the past decade. At the beginning of the millennium standard definition (SD, resolution: 720×576) resolution was the standard for multimedia playback, today the standard is full HD (1920×1080), and soon to come 4k (3840×2160). An increase in resolution brings an increase in transmitted bit rate that is necessary to deliver the media to the user. Although the bit rate of broadband connections has been increased as well, there still is a mismatch, leading to the situation that bandwidth is a restricting variable. Thus, it is important for service providers to find the optimal tradeoff between transmitted bit rate and perceived quality at the customer side, in order to keep costs low and still deliver a satisfactory quality service. Especially, for popular live reporting events not the best possible video quality can be transmitted, as the required hardware setup would be too complex. In order to determine perceived quality, subjective tests have to be conducted. How these tests should be carried out is specified in several recommendations released by the International Telecommunication Unit (ITU).

For most subjective studies in the area of QoE, short sequences with a length of around 10 s are used as stimulus material, as recommended e.g. by ITU-T P.910 [1]. These rather short stimuli are not representing the typical use of media, therefore other recommendations have been released with a scope of longer videos, with up to 30 min. These are specified e.g. in ITU-R BT.500 [2]. Here, participants are asked to give a continuous quality judgment with the help of a slider with which they continuously can give a rating of the experienced quality, and therefore can react to spontaneous changes in quality. However, this setup has the disadvantage that test participants have to avert their gaze from time to time from the screen to the slider, as they need to check the adjustment of the slider.

* Corresponding author.
E-mail address: sebastian.arndt@mailbox.alumni.tu-berlin.de (S. Arndt).

http://dx.doi.org/10.1016/j.image.2016.01.011
0923-5965/

Thus, the flow experience might suffer. In order to overcome these problems, the use of different modified input devices has been investigated. Here, a steering wheel, joystick [3], or a glove [4] have been proposed. Still, the advantageous use could not been shown compared to the slider method. A different approach has been proposed by Borowiak et al. [5], which aims not at a direct quality evaluation, but the momentary quality can be adjusted by the test participant. This is done as the participant turns a knob until he or she perceives a perfect quality of the shown material.

However, using any of the above mentioned methods does not bring any information about underlying cognitive processes, and how these are affecting the mental state of the media consumer. In order to get some insights in what happens inside a consumer of multimedia content, physiological measures can be helpful.

Physiological measures, such as electroencephalography (EEG), have shown to bring additional results to subjective MOS (mean opinion score) values. This could be shown in studies using rather short stimuli, with a length of words or syllables. Using audio [6], visual [7], and audiovisual [8] stimuli, it was shown that the stronger the degradation is, the greater the cognitive response. Thus, a higher P300 peak amplitude was present when exposing test participants to low quality sequences compared to better quality. The P300, a positive deflection about 300 ms after stimulus onset, is part of the event-related potential (ERP) which is elicited when a participant is exposed to a stimulus. Commonly, the latency and amplitude of the P300 depends on several factors, such as e.g. the intensity or irregularity of a stimulus [9].

When analyzing long-term changes (i.e. several minutes) in the brain, an analysis of frequency bands is often performed. The band power can be divided into different sub-bands, delta (0.5–3.5 Hz), theta (3.5–7.5 Hz), alpha (7.5–12.5 Hz), beta (12.5–30 Hz), and gamma (30–60 Hz) [10]. The alpha band is most interesting when it comes to draw conclusions concerning cognitive states, such as fatigue,[1] as it is shown in [12] in the context of driver's fatigue, and in [13] for effects of mental fatigue. However, research has shown that identifying individual alpha bands per subject may be the more accurate way to analyze this [14], due to individual difference of the human brain.

First studies in the context of QoE using frequency band power analysis could already give some insights into the perception process during the presentation of multimedia content. In two studies, an audiobook was played to the test participants [15,16]. It could be shown that when confronting participants with a low-quality version of the audiobook, they tended to be more fatigued compared to the high-quality version. Alpha and theta portions were larger for low-quality conditions compared to high-quality conditions.

[1] The terms sleepiness and fatigue have been used interchangeably in the literature. In the literature, sleepiness is defined more detailed and is referred to changes in the neurobiological processes, whereas the term fatigue is indicating to not being able to sustain on a certain level of task performance [11]. Therefore, we will use the two terms, *sleepiness* and *fatigue*, interchangeable within this paper.

Using physiological measures assumptions about the cognitive state and its parameters, such as e.g. work load and fatigue, can be drawn.

Work load is defined as the needed processing resources to fulfill a task [17]. To perform a task a certain amount of resources is demanded and these resources are limited. The amount of available resources can be influenced by numerous factors e.g. stable influences such as individual differences between humans or time of day effects [18] or fluctuating influences such as time on task effects or monetarily cognitive/physical condition. The cognitive state describes the level of free resources which a user of a system has. In an increased/high cognitive state almost all of the available resources are used and in a decreased/low cognitive state almost all resources are available. A sustained period or high workload can lead to a lower cognitive state as fewer resources are available and lead to being more sleepy; users have less available resources for task processing available. This goes along with an increase in alpha band power [19,20]. Fatigue refers to the feeling of reduced alertness that impairs the capabilities and willingness to perform a task [21].

In this paper, we will present two studies using audiovisual material with quality degradations in both modalities. We recorded an EEG and analyzed the frequency band power concerning portions of alpha and theta band power, thus indicators of fatigue and reduced attention.

The remainder of this paper is organized as follows: In Section 2, we report on a first study using a stimulus impaired only in the visual modality, and in Section 3 on a second study where the material was degraded in both modalities. In Section 4, we discuss the two studies, before we conclude with the consequences for physiological recordings applied in QoE in Section 5.

2. Study 1

Motivated by [15], a study has been carried out using a commercially available documentary about sea life which consists in the audio track only of content related noise, such as e.g. movement of the waves, and almost no speech. As we have learned from [8], superimposition effects can appear when analyzing both modalities at the same time for an initial study, thus we decided to start with a simple stimulus setup. Initial results of this study were already reported in [22].

2.1. Stimulus

We took the documentary ‘Ocean’ as a stimulus. This video shows mainly sea life scenes, and is easy to follow. It was taken from a BluRay with full HD resolution (1920×1080), and the original video bit rate of 40,000 kB/s. The video shown was a 40 min excerpt. The cutting did not affect the story of the video. For the test, we had two conditions, (a) the original video without any quality reduction (named HQ from here on), and (b) a video with reduced bit rate (named LQ). For quality reduction, we used the H.264 codec, in the $\times$264 implementation, and had a reduced bit rate of 2000 kB/s (constant bit rate settings).

The video was then divided into two parts, of which one part was shown in HQ and the other in LQ. Whether the first or second part was degraded was randomized between participants.

2.2. Experiment

The study was conducted in a room conform to ITU-T Rec. P.910 [1], with gray curtains around the room, and appropriate light settings. The video was shown on a 42" LCD screen, and subjects sat 1.8 m away from the screen (i.e. 4 times video height).

During the experiment the subjects had the task to rate the quality every 6 min, resulting in three ratings per quality condition. Therefore, a pop-up screen appeared with a discrete 11-point ACR scale with the corresponding labels recommended by ITU-T P.910 [1].

2.3. Physiological recordings

During the presentation of the stimulus we recorded EEG activity as well as eye movements, using EOG (electro-oculogram). For the EEG recording, six (passive) EEG channels were attached to the subject, using electrodes Fz, Cz, Pz, P3, P4, and O1 according to the 10–20 international standard system [23]. Mastoid was used as reference. For EOG, horizontal and vertical movements were recorded. Both recordings were done with a sampling frequency of 500 Hz.

Both recordings were set in relation to a baseline which consisted of the first two minutes of the EEG and EOG recording respectively. Thus, a level of 100 indicates no change compared to the baseline.

2.4. Results

Twelve test participants took part in the study (3 female, 9 male) with a mean age of 26 years. All of them had normal or corrected-to-normal vision, none reported any hearing impairments or health issues. A Snellen test for visual acuity and an Ishihara test for colorblindness were conducted before the experiment.

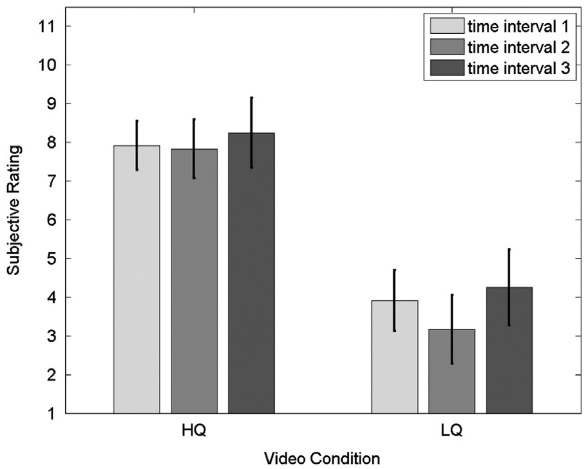


Fig. 1. Averaged subjective quality ratings for *Study* 1 for the three rating times and two conditions. Left side showing HQ, and right side LQ condition. Error bars indicate standard deviation.

2.4.1. Subjective quality data

The averaged subjective rating over all participants is represented in Fig. 1. The ratings show that the HQ sequence was evaluated significantly higher than the LQ part ($F(1,11) = 142.46$, $p \leq 0.01$, $\eta^2 = 0.93$). There is no significant difference between ratings of one shown quality during the three assessment time points ($F(2,22) = 2.84$, n.s.).

2.4.2. EEG data

When analyzing the alpha sub-band, we divided the signal temporally in the three parts. These correspond to the rating intervals. A significant effect can be observed for most of the posterior electrodes (Pz, P4, P3, O1, Cz), see exemplary electrode Pz in Fig. 2a ($F(10,1) = 8.76$, $p \leq 0.01$, $\eta^2 = 0.47$). Also generally a higher theta activity for LQ can be observed for most electrodes, though, the effect is not significant. In Fig. 2b, the distribution of alpha and theta activity over the whole scalp is plotted. As can be seen already during the first part, a big gap in alpha between the HQ and LQ condition is visible. To investigate this in more detail the signal was split into 1 min chunks, and the first period was analyzed in more depth, see Fig. 3. Calculating a running t-test over time between the two conditions, a significant difference can already be obtained after 4 min with $p \leq 0.05$.

2.4.3. EOG Data

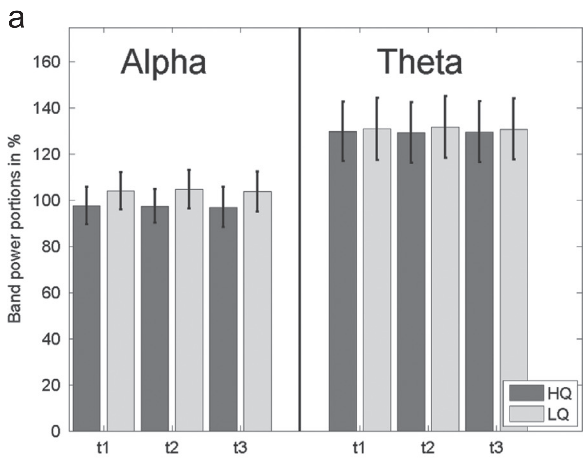
The data gathered by the EOG electrodes was first reported in [24]. Here, it was shown that the blink duration is longer when participants were watching the LQ video compared to the HQ video.

Blink detection was done offline using the freely available Matlab software eogui. After all events were extracted and parameterized, the parameters for each blink were normalized subject-wise by dividing it through the average value during the baseline (the first two minutes of each respective video clip) and then multiplied by 100, resulting in values that can be understood as 'percent baseline activity'. Baseline-corrected average blink rate (measured as blink interval, i.e. time between consecutive blinks) did not differ between the two videos ($F(1,55) = 0.125$, n.s.) or in relation to rating moment within one video ($F(2,54) = 1.048$, n.s.), and there was also no interaction of both factors video type*rating moment within video block ($F(2,56) = 0.006$, n.s.). Baseline-corrected average blink duration, however, showed a clear effect of video type ($F(1,70) = 14.098$, $p \leq 0.01$), a non-significant tendency to increase with watching/rating time ($F(2,57) = 2.743$, $p \leq 0.1$, see also Fig. 4), and no interaction of video type*rating moment within video block ($F(2,51) = 1.932$, n.s.).

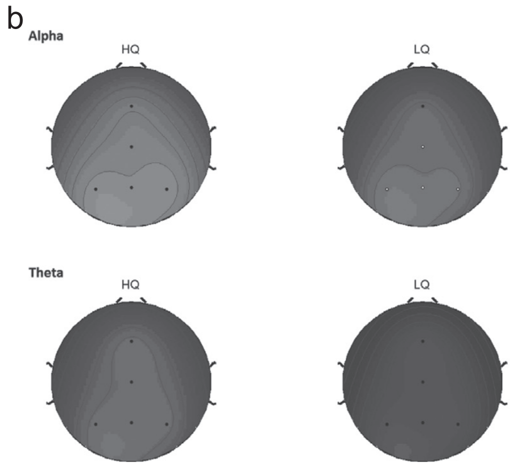
2.5. Discussion

In *Study* 1 we used subjective ratings as a sanity check, in order to ensure that we provoked the intended effect, thus, the difference between low- and high-quality media. We could confirm that sequences in the HQ condition were rated better than in the LQ condition.

Using the analysis of frequency bands, we could show that the band power of the alpha and theta band was higher for LQ sequences compared to HQ. Analyzing these



Alpha (left) and theta (right) band power portions for the three times (t1, t2, t3) and two conditions at electrode Pz. Error bars indicate standard deviation.



Topo plots for alpha (top) and theta (bottom) band power for each condition. White electrodes show a significant difference compared to the HQ condition.

Fig. 2. EEG data of *Study* 1. Data is averaged over all test participants.

portions in smaller time steps, we showed that a significant difference between conditions appeared already after a few minutes. Thus, it might be enough to provide stimuli with several conditions, each lasting only for a couple of minutes, to gain results comparable to those presented here, and thus reduce the time expenditure.

Prior research has shown that blink rate and blink duration are influenced by different task aspects: Blink rate is predominately affected by the visual demands of a task [25], and [26] could dissociate this from the mental workload imposed. On the other hand, we could show that blink duration is more sensitive to sleepiness than blink rate [27]. In this line, we interpret the unchanging blink rate as an indicator for comparable visual load of both videos as intended. The lower video quality however

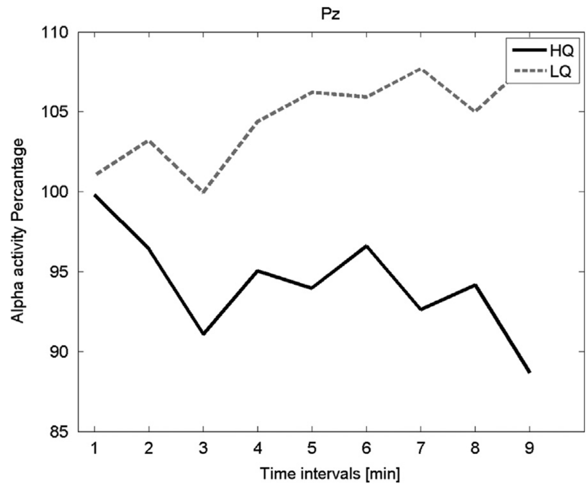


Fig. 3. *Study* 1: Alpha band power portions over the first minutes for electrode Pz.

appears to have a fatiguing effect on the subjects as indicated by the higher blink duration.

Both physiological measures show a significant difference between the two presented conditions. A higher level of alpha band power and longer blink duration are both indicators of a higher level of fatigue during the LQ condition compared to when watching the HQ sequence.

3. Study 2

In *Study* 1 we showed that the analysis of EEG frequency power bands is sensitive to long videos with varying quality and that this effect is already present after several minutes. *Study* 2 was designed to not only degrade one modality (the visual as in *Study* 1, or the auditory as in [15]), but both synchronously. Furthermore, it should be analyzed how the two modalities are influencing each other. Therefore, a documentary in which audio and video had a similar importance to transmit the content was used. The importance was estimated subjectively by the experimenter. In order to evaluate the validity of the EEG recordings, again, additional subjective ratings were captured. Initial results are reported in [28].

3.1. Stimulus

For the stimulus, we used a German documentary with a male narrator ('Giganten der Tiefe', Giants of the sea). The movie contained mainly scenes from below the water, no people were visible in the movie for the entire duration. The film was in HD resolution (1920×1080). The original material had a bit rate of 5250 kb/s and was encoded with the H.264 encoder. The manipulated video track had a reduced bit rate in average of 645 kB/s, applying the H.264 codec, in the $\times 264$ implementation, using the constant quantizer mode.

The audio track had a German male voice as a background narrator providing detailed information on the purpose of the depicted animal behavior, and some scene-related background noise (without any music). The original audio signal had a bit rate of 1411 kB/s and 44,100 Hz sampling rate, for the

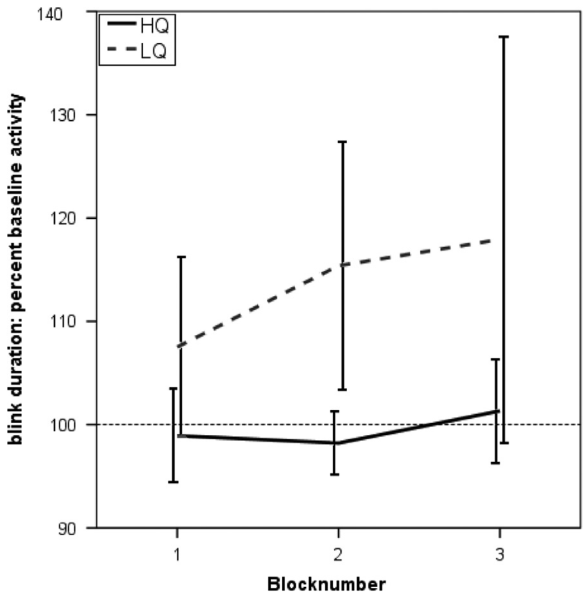


Fig. 4. Eye blink duration for the three blocks separated by condition, for *Study* 1.

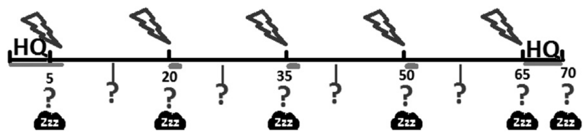


Fig. 5. General procedure of *Study* 2. Gray bars indicate reference sequences. Question-marks: subjective ratings. 'zzz': self-assessment on the sleepiness scale. 'Thunder': Introduction of a quality change.

track with reduced quality the GSM 1.60 codec was used with 8000 Hz sampling rate and 13 kB/s bit rate.

The video was divided into six parts. The first part was a 5 min baseline condition, where no degradation was introduced (called Prebaseline). The same was done for the last 5 mins of the video, where no modifications were applied to the signal (Postbaseline). Between these two anchors, all four degradation combinations were presented, thus HQ-Video/HQ-Audio, HQ-Video/LQ-Audio, LQ-Video/HQ-Audio, LQ-Video/LQ-Audio. The sequence in which these appeared was randomized between participants, resulting in 24 combinations.

3.2. Experiment

The general procedure of the experiment can be seen in Fig. 5. For each condition two subjective ratings were obtained, one in the middle and one at the end. The quality judgments were gathered separately for each modality (audio, video, and audiovisual). Therefore, an 11-point discrete ACR scale was used with the corresponding labels recommended by ITU-T Rec. P.910. Additionally, the subjective sleepiness was reported on the 9-point discrete Karolinska scale [29] with the corresponding German labels. The beginning and end of the movie was shown in reference condition. The next condition also always started with 90 s of the undistorted material (indicated by gray bars in Fig. 5).

Table 1
Overview of subjective quality ratings of *Study* 2 for all conditions including Pre- and Post-baseline. Containing separate judgments for both modalities and the overall rated quality as well as the self-reported sleepiness level. Numbers in brackets denote standard deviation.

Condition	Video	Audio	AV	Sleepiness
Prebaseline	6.42 (1.50)	7.96 (1.49)	6.96 (1.40)	3.71 (1.68)
Postbaseline	7.96 (1.38)	8.17 (1.09)	7.33 (1.05)	4.96 (1.97)
HQ-V HQ-A	6.65 (1.21)	8.08 (1.00)	7.06 (1.25)	4.75 (1.78)
HQ-V LQ-A	6.04 (1.67)	3.23 (1.39)	4.83 (1.43)	5.00 (1.93)
LQ-V HQ-A	3.58 (1.80)	7.52 (1.36)	5.29 (1.78)	4.79 (1.74)
LQ-V LQ-A	2.65 (1.29)	2.81 (1.30)	2.85 (1.27)	5.04 (1.78)

3.3. Physiological recordings

Next to the standard subjective ratings, several physiological measures were taken during the experiment. For EEG recordings, the active wet system from g.tec was used with 16 electrodes at Fp1, Fp2, F3, Fz, F4, T7, C3, CZ, C4, T8, P4, Pz, P3, O1, Oz, O2 according to the 10–20 standard system [23]. Recordings were done in 512 Hz and were sampled down to 200 Hz for analysis.

During the experiment, skin conductance was measured at the palm of the left hand with 200 Hz sampling frequency. The skin conductance, galvanic skin response, or electrodermal activity (EDA) can be used as a measure of emotional response, more specifically, as a measure of arousal. Arousal is the degree of intensity of an emotion [30]. When a person is aroused, the sweat gland activity is higher, which leads to an increased value of the measured skin conductance [31].

As an additional response of the physiological system, an electrocardiogram (ECG) was obtained from the test participant. Therefore, electrodes were placed at the two common carotid arteries and the vein of the left forearm. Activity was recorded with 200 Hz sampling frequency. The heartbeat can be extracted from the ECG, thus, the heart-rate or heart-rate-variability (HRV) can be calculated [32]. Besides giving information about the general physical condition of the test participant, these are indicators e.g. for stress, as in stressful conditions the heart has a more uniform beat, thus, the HRV is lower [33,34].

An electromyogram (EMG) was obtained during the showing of the video from two sides, namely from the cheek (Zygomaticus, called EMG-cheek from here on) and from above the eye brow (Corrugator supercilii, called EMG-forehead) with a sampling rate of 200 Hz. EMG in general reports on the muscle activity below the recorded area. EMG-cheek can be used to draw conclusions about positive emotions, as theses muscles are tensed while smiling, whereas EMG-forehead can be used to rather quantify negative emotions, as frowning produces increased activity in this area [35].

3.4. Results

Twenty-four participants took part in this study (8 female, 16 male) who all were German native speakers with an average age of 26.7 years. None reported any hearing impairment or health issues, all of them had normal or corrected-to-normal vision. This was reconfirmed with a Snellen test for visual acuity and Ishihara plates for color blindness.

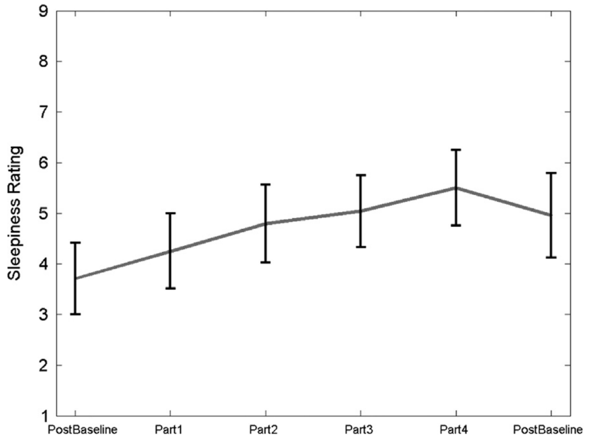


Fig. 6. Averaged self-reported sleepiness ratings, as they appeared in time. Error bars indicate standard deviation.

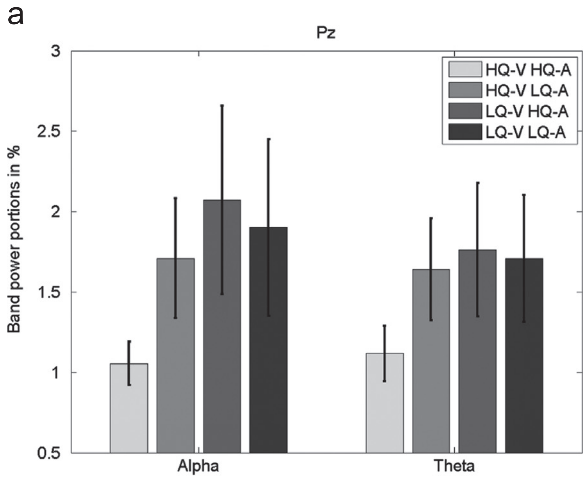
3.4.1. Subjective data

To test whether the different conditions led to the intended effect, such as that lower quality led to lower subjective ratings in quality, these ratings are analyzed first. It can be seen from Table 1 that the four conditions lead to significant differences on the three quality scales: Audio: $F(3,69)=150.43$, $p \le 0.01$; for video: $F(3,69)=52.61$, $p \le 0.01$; for audiovisual: $F(3,69)=44.83$, $p \le 0.01$. Bonferroni post-hoc comparisons show significant differences between HQ and LQ conditions for both modalities as well as significant differences in AV between HQ-V/HQ-A and all others as well as between LQ-V/LQ-A and all others. See Table 1 for detailed data inspection.

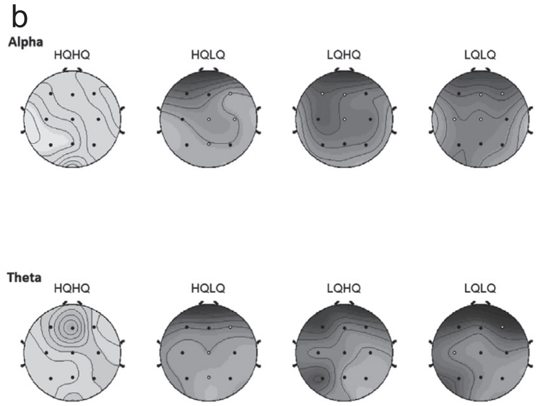
To evaluate the consistency of the participants' answers, we analyzed the MOS judgment of the Prebaseline, Postbaseline, and HQ-V/HQ-A sequences. Calculating an ANOVA with the three times as independent variable and the MOS for the three judgments (Audio, Video, and AV) as dependent variables, no significant difference between the sequences can be seen (audio: $F(2,46)=0.54$, n.s.; video: $F(2,46)=1.9$, n.s.; AV: $F(2,46)=1.55$, n.s.). However, from the descriptive analysis it turns out that the Prebasline was rated lowest and Postbaseline highest.

Checking also for the steadiness between the first rating and the second rating within each condition we observe that for Audio ($F(1,23)=0.75$, n.s.) and Video ($F(1,23)=1.47$, n.s.) there are no significant differences, but for the AV judgment a significant effect can be observed, as the second ratings tended to be higher than the first ratings ($F(1,23)=7.08$, $p \le 0.01$).

The averaged self-reported sleepiness ratings can be seen in Table 1. Based on the used Karolinska scale [29], lower values indicate a higher self-reported level of alertness, hence higher values indicate higher self-reported sleepiness. Here, it can be seen that there is no big difference between the conditions, which results in no significant difference in the ratings between conditions ($F(3,69)=0.47$, n.s.). However, when looking at the sleepiness ratings sequentially, thus, in the order they appeared in time an increase in the rating can be seen as time prolongs, see Fig. 6. This effect, frequently referred to as time on task, is also statistically significant ($F(5,115)=8.89$, $p \le 0.01$). Interestingly, the rating descents for the Postbaseline sequence which is not significant.



Alpha and theta band power level for each condition, in relation to Prebaseline alpha and theta level, at electrode Pz. Error bars indicate standard deviation.



Topo plots for alpha (top) and theta (bottom) band power level for each condition, in relation to Prebaseline alpha and theta level. White electrodes show a significant difference compared to the HQ condition.

Fig. 7. EEG Data of *Study* 2. Data is averaged over all test participants.

3.4.2. EEG data

For the EEG data, alpha band power proportions were calculated. Therefore, a frequency analysis was performed on the EEG data for each section, using Welch's method. Then, the so-called alpha values have been calculated, which are in the range of 8–13 Hz. These values then have been divided by the Prebaseline of each participant. The results are exemplarily plotted for electrode Pz in Fig. 7a. As can be seen, alpha portions are significantly higher for the low quality conditions compared to the conditions with higher quality ($F(3,69)=4.54$, $p \le 0.01$). No significant difference is observed between the different low-quality conditions.

In Fig. 7b, the overall distribution of alpha activity of the scalp for the different conditions is represented.

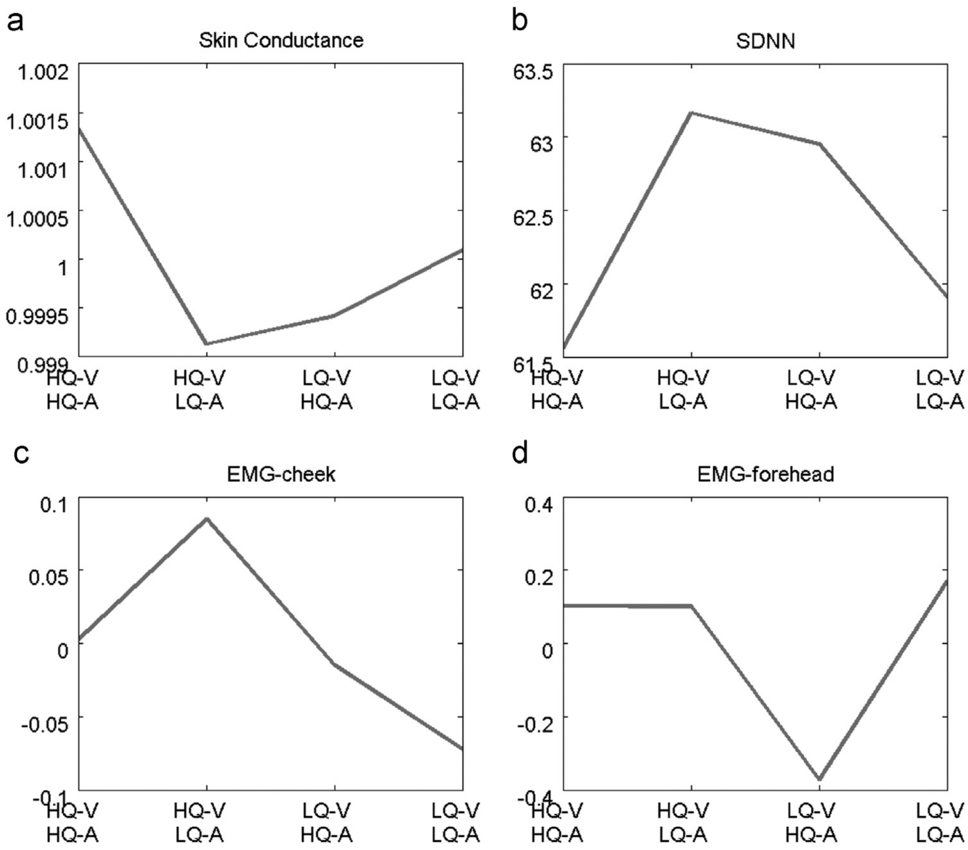


Fig. 8. Averaged physiological data for all four conditions of *Study* 2. (a) Skin conductance in relation to Prebaseline. (b) Standard deviation of inter-beat interval in ms. (c) EMG activity from the cheek from −1 s to 10 s after quality change, *z*-normalized within test participants. (d) EMG activity from above the eye brow from −1 s to 10 s after quality change, *z*-normalized within test participants.

Confirmatory, it can be seen that low-quality conditions produce higher alpha portions than the reference (HQ-V/HQ-A) condition.

Analyzing the theta portions (4–8 Hz) in the signal the same way as alpha reveals a significant difference between the shown conditions ($F(3,69)=3.23$, $p \leq 0.05$), see also Fig. 7a and b.

3.4.3. Other measures

In order to analyze the recorded skin conductance, the data was filtered from 0.0625 Hz to 5 Hz, and subsequently the amplitude was averaged over the time intervals [31], and as for the EEG data, set into relation to the Prebaseline. Looking at the obtained values for each condition, it can be seen that there is an increase in skin conductance from HQ-V/LQ-A, via LQ-V/HQ-A, and LQ-V/LQ-A, but the value is highest for the reference condition. Calculating an ANOVA also reveals no significant difference between the conditions ($F(3,69)=2.02$, n.s.), see Fig. 8a.

Additionally, we recorded an ECG (electrocardiogram) where we determined the heart-rate variability (HRV). HRV is indicative for stress for example [34]. Analyzing the data by looking at the standard deviation of the heart-rate interval (SDNN) no significant difference between the conditions was found, see Fig. 8b.

To examine the EMG data from both sides, the phasic response around the quality change was analyzed. Therefore, data from 1 s before until 10 s after the change was extracted. The data was filtered from 10 Hz to 45 Hz according to [36], and the mean amplitude was determined [35]. An ANOVA did not reveal any significant differences for both EMG sides between conditions (EMG-cheek: $F(3,69)=0.1$, n.s.; EMG-forehead: $F(3,69)=1.5$, n.s.), see Fig. 8c and d.

3.5. Discussion

In *Study* 2, subjective quality ratings of both modalities were obtained, and showed that the LQ condition led to lower quality ratings in the corresponding ratings. In situations where the reference condition was shown (Pre-, Post-baseline, and HQ-V/HQ-A) the ratings did not significantly differ, also the ratings did not differ between the two time points within one condition. Thus, we can conclude that the ratings are reliable over the entire experiment.

Furthermore, we asked for the self-judged level of sleepiness, which did not show any significant difference between condition, but with time. Thus, test participants reported to feel more fatigued as time proceeded. This shows, that the measure of self-reported sleepiness or alertness is working in general, and that it is a helpful tool for self-evaluation of the participant's current state. We tried to account for this time on task effect by randomizing the presentation order. No significant difference in the self-reports was found between conditions. This shows that the possible change of the cognitive state we are investigating may be very subtle, and may not be possible to detect using self-reports.

With help of the analysis of frequency bands we could show an increase in alpha and theta band power activity in the EEG, which are indicative for fatigue, for mixed and LQ conditions. Thus, the subjective judgment on sleepiness and the neurophysiological response of the test participant differed. Although, physiological measures reflecting the activity of the autonomous nervous system like EDA or ECG did not show significant differences between conditions, we can conclude that the analysis of EEG features can be an indicator of fatigue provoked by viewing audiovisual material with different quality levels.

4. Overall discussion

In both presented studies, we could show that the QoE reported by test participants is represented in the EEG data. More specifically, the calculated alpha portions were generally higher for low quality conditions compared to higher quality conditions.

In previous research, it was shown that a short speech/video clip in reduced quality leads to impaired information processing on the short term, i.e. increased P300, compared to presentation of non-degraded stimuli [6,8]. This leads to an increase of work load and therefore, to more dedication of resources, and thus, to a higher level of attention. When the presented stimulus material is in lower quality for a longer time this impaired information processing leads to an elevated level of work load over a longer time, which leads to a reduced cognitive state over time. A reduction of the cognitive state leads to an increase of portions in the alpha band power [37]. Thus, the participant is getting more fatigued in case of presentation of low quality sequences compared to sequences in high quality. Fatigue in general refers to a reduction of mental performance capabilities, we do not claim that low quality stimuli is leading to fatigue, but that the reduction in quality over longer time leads to an enduring high level of work load which leads to a reduced cognitive state that goes along with decreased attention. This state, we call being more fatigued compared to the condition where no reduction of quality was performed.

In standard EEG literature, an increase in alpha activity is associated with an increase in fatigue and low attention [38]. A review on EEG studies investigating the level of fatigue using EEG shows that in the majority of analyzed studies an increase in alpha and theta band power is present when a participant is fatiguing [19]. Thus, we argue that the increase in alpha is due to decreasing attention of the subject. The analyzed theta activity showed the trend in both studies to be higher for distorted versus non-distorted sequences. However, only in *Study* 2 this effect was also of statistical significance. An increase of theta activity being an additional indicator of drowsiness which results from impaired information processing [10].

In *Study* 1 we could show an increased blink duration for the LQ condition which is an additional indicator of fatigue [39]. Reference [40] shows in their work that higher work load leads to an increase in blink duration and an increase of fatigue. In the LQ conditions, participants were less immersed in the video content and their overall activity level decreased.

Additional recordings of peripheral physiological data did not show this clear trend. Especially, in *Study* 2 where different measures were obtained, such as EMG and EDA, no clear trend could be observed probably because the type of fatigue (i.e. fading attention) is predominately mediated by cognitive processing, less by somatic activity level. That those peripheral physiological measures might not be ideal in the context of QoE assessment was first shown in [41] who also used longer stimuli (ca. 30 s) and applied codecs in order to produce degraded stimulus material. However, in a prior study of [42], it could be shown that a strong reduction in frame rate, of 5 fps, resulted in a significant physiological response, namely an increase in galvanic skin response and heart rate, and a decrease in blood volume pulse (BVP) which are indicators of stress. The common ground of [41] and the here-reported studies is that both used codecs, resulting into spatial degradations, whereas [42] used a reduction in frame rate that leads to jerkiness which belongs to the class of temporal artifacts [43]. Thus, it would be interesting to investigate whether jerkiness in the video and interruptions in the audio lead to a similar response in the EEG. Furthermore, a within-study design of temporal and spatial artifacts can explore whether the results of the two studies can be reproduced.

5. Conclusions

The here-presented studies clearly show that low-quality long audiovisual stimulus material has a fatiguing effect to viewers. This effect was present in two frequency sub-bands of the EEG, namely in alpha in both studies, and in theta which reached only in *Study* 2 statistical significance. Data from the EEG was supported by data originating from blink duration. We saw that test participants did not feel more fatigued when being asked directly between the different conditions. Thus, EEG band power analysis might be more sensitive concerning quality changes which lead to reduced cognitive processing. Thus, the studies show a valid approach to look at physiological involvement in a test participant due to low quality media.

Generally, it can be assumed that increased fatigue leads to a decrease in attention [44]. Transferring this to our findings it suggests that presentation of low quality media leads to a decrease in attention (compared to high quality media). This is a relevant factor relevant for paying users as well as advertising partners of multimedia providers.

5.1. Future work

The audio book studies [15] as well as the here-presented studies, aim at showing the general feasibility to obtain such measures in the context of QoE. Subjective reactions towards quality reduction have been studied extensively. Here, the physiological response underlying the subjective evaluation is analyzed. At the moment this aims at looking at quality differences which are obvious. However, in the future it is aimed at conducting similar studies around the perceptual threshold and see whether the found results can be confirmed. Another interesting application area is to use these kind of measures for physiological adaptive streaming. Thus, based on the current evaluated state the quality can be adapted or advertisement can be placed when the consumer needs a break to mentally refresh.

Future work could concentrate more on the cause of why alpha is decreasing in *Study* 2 in the LQ-V/LQ-A condition and whether this is due to the previous mismatch of qualities in HQ-V/LQ-A and LQ-V/HQ-A conditions.

Another aspect which is interesting to look at is, whether rather subtle or subjectively non-perceivable quality degradations lead to a similar physiological response as shown. The just-noticeable-difference (JND) on the physiological level should be compared to the subjective JND. Based on these, codec or network parameters could be adjusted.

Another point is to look into the different perceptions of spatial versus temporal artifacts in coding or transmission. Thus, in a future experiment participants would be exposed to both classes of degradation where recordings of EEG as well as peripheral physiological data could be performed. It would be interesting to see whether EEG data would be affected by both types of degradation, and whether peripheral physiological data is only affected by temporal degradations and not by spatial, as suggested from prior work.

Acknowledgment

The authors would like to thank Steffen Zander and Christiana Manolaina for their help running the experiments. This work was co-funded by the German Research Foundation (DFG – 1013 Prospective Design of Human-Technology Interaction). We would like to thank the Bernstein Focus: Neurotechnology Berlin (BFNT-B) which was supported by the Federal Ministry of Education and Research (BMBF) FKZ 01GQ0850.

References

[1] ITU-T Recommendation P.910, Subjective Video Quality Assessment Methods for Multimedia Applications, International Telecommunication Union, Geneva.
[2] ITU-R Recommendation BT.500, Methodology for the Subjective Assessment of the Quality of Television Pictures, International Telecommunication Union, Geneva.
[3] O. Nemethova, M. Ries, A. Dantcheva, S. Fikar, M. Rupp, Test equipment of time-variant subjective perceptual video quality in mobile terminals, 2005.
[4] S. Buchinger, W. Robitza, P. Hummelbrunner, M. Nezveda, M. Sack, H. Hlavacs, Slider or glove? Proposing an alternative quality rating methodology, in: Fifth International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM), Scottsdale, Arizona, 2010.
[5] A. Borowiak, U. Reiter, U.P. Svensson, Audio quality requirements and comparison of multimodal vs. unimodal perception of impairments for long duration content, J. Signal Process. Syst. 74 (1) (2014) 79–89.
[6] J.-N. Antons, R. Schleicher, S. Arndt, S. Möller, A.K. Porbadnigk, G. Curio, Analyzing speech quality perception using electroencephalography, J. Sel. Top. Signal Process. 6 (6) (2012) 721–731.
[7] L. Lindemann, M. Magnor, Assessing the quality of compressed images using EEG, in: 2011 18th IEEE International Conference on Image Processing (ICIP), IEEE, Brussel, Belgium, 2011, pp. 3109–3112.
[8] S. Arndt, J. Antons, R. Schleicher, S. Möller, G. Curio, Using electroencephalography to measure perceived video quality, IEEE J. Sel. Top. Signal Process. (2014) 366–376.
[9] S. Luck, An Introduction to the Event-Related Potential Technique, The MIT Press, Cambridge, Massachusetts, USA, 2005.
[10] M.D. Rugg, M.G. Coles, Electrophysiology of Mind: Event-Related Brain Potentials and Cognition, Oxford University Press, Oxford, UK, 1995.
[11] S.K. Lal, A. Craig, Driver fatigue: electroencephalography and psychological assessment, Psychophysiology 39 (3) (2002) 313–321.
[12] S.K. Lal, A. Craig, A critical review of the psychophysiology of driver fatigue, Biol. Psychol. 55 (3) (2001) 173–194.
[13] Y. Punsawad, S. Aempedchr, Y. Wongsawat, M. Panichkun, Weighted-frequency index for EEG-based mental fatigue alarm system, Int. J. Appl. 4 (1) (2011) 37.
[14] W. Klimesch, M. Doppelmayr, H. Russegger, T. Pachinger, J. Schwaiger, Induced alpha band power changes in the human EEG and attention, Neurosci. Lett. 244 (2) (1998) 73–76.
[15] J.-N. Antons, R. Schleicher, S. Arndt, S. Möller, G. Curio, Too tired for calling? A physiological measure of fatigue caused by bandwidth limitations, in: the Fourth International Workshop on Quality of Multimedia Experience (QoMEx 2012), IEEE, Yara Valley, Australia, 2012, pp. 63–67.
[16] J.-N. Antons, F. Köster, S. Arndt, S. Möller, R. Schleicher, Changes of vigilance caused by varying bit rate conditions, in: the Fifth International Workshop on Quality of Multimedia Experience 2013 (QoMEX 2013), IEEE, Klagenfurt, Austria, 2013, pp. 148–151.
[17] C.D. Wickens, Processing Resources in Attention, Dual Task Performance, and Workload Assessment, Tech. rep., DTIC Document, 1981.
[18] M. Blake, Time of day effects on performance in a range of tasks, Psychon. Sci. 9 (6) (1967) 349–350.
[19] A. Craig, Y. Tran, N. Wijesuriya, H. Nguyen, Regional brain wave activity changes associated with fatigue, Psychophysiology 49 (4) (2012) 574–582.
[20] J.C. Shaw, The Brains Alpha Rhythms and the Mind, in: J.C. Shaw, Elsevier, Amsterdam, 2003.
[21] J. Shen, J. Barbera, C.M. Shapiro, Distinguishing sleepiness and fatigue: focus on definition and measurement, Sleep Med. Rev. 10 (1) (2006) 63–76.
[22] S. Arndt, R. Schleicher, J.-N. Antons, Does low quality audiovisual content increase fatigue of viewers? in: The fourth International Workshop on Perceptual Quality of Systems (PQS 2013), 2013, pp. 1–4.
[23] H. Jasper, The ten-twenty electrode system of the international federation, Electroencephalogr. Clin. Neurophysiol. 10 (2) (1958) 371–375.
[24] R. Schleicher, S. Arndt, J.-N. Antons, Changes in blinking behavior while watching videos with reduced quality, in: Book of Abstracts of the 17th European Conference on Eye Movements, 2013, p. 130.
[25] C. Fogarty, J.A. Stern, Eye movements and blinks: their relationship to higher cognitive processes, Int. J. Psychophysiol. 8 (1) (1989) 35–42.
[26] M.Á Recarte, E. Pérez, Á Conchillo, L.M. Nunes, Mental workload and visual impairment: differences between pupil, blink, and subjective rating, Span. J. Psychol. 11 (02) (2008) 374–385.
[27] R. Schleicher, N. Galley, S. Briest, L. Galley, Blinks and saccades as indicators of fatigue in sleepiness warnings: looking tired? Ergonomics 51 (7) (2008) 982–1010.
[28] S. Arndt, J.-N. Antons, S. Möller, Is low quality media affecting the level of fatigue? in: The Sixth International Workshop on Quality of Multimedia Experience (QoMEX 2014), 2014.
[29] T. Åkerstedt, M. Gillberg, Subjective and objective sleepiness in the active individual, Int. J. Neurosci. 52 (1–2) (1990) 29–37.
[30] D. Purves, E.M. Brannon, R. Cabeza, S.A. Huettel, K.S. LaBar, M.L. Platt, M.G. Woldorff, Principles of Cognitive Neuroscience, vol. 83, Sinauer Associates, Sunderland, MA, 2008.
[31] W. Boucsein, Electrodermal Activity, Springer, New York, USA, 2012.
[32] G.G. Berntson, K.S. Quigley, D. Lozano, Cardiovascular psychophysiology, in: Handbook of Psychophysiology, vol. 3.
[33] W.B. Cannon, The Wisdom of the Body, WW Norton & Co., New York, USA, 1932.
[34] T.G. Vrijkotte, L.J. van Doornen, E.J. de Geus, Effects of work stress on ambulatory blood pressure, heart rate, and heart rate variability, Hypertension 35 (4) (2000) 880–886.
[35] L.G. Tassinary, J.T. Cacioppo, E.J. Vanman, The skeletomotor system: surface electromyography, Handbook of Psychophysiology 2 (2000) 163–199.
[36] A. Van Boxtel, Optimal signal bandwidth for the recording of surface EMG activity of facial, jaw, oral, and neck muscles, Psychophysiology 38 (01) (2001) 22–34.

[37] M.A. Boksem, T.F. Meijman, M.M. Lorist, Effects of mental fatigue on attention: an erp study, Cogn. Brain Res. 25 (1) (2005) 107–116.
[38] W. Klimesch, EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis, Brain Res. Rev. 29 (2) (1999) 169–195.
[39] K. Holmqvist, M. Nyström, R. Andersson, R. Dewhurst, H. Jarodzka, J. Van de Weijer, Eye Tracking: A Comprehensive Guide to Methods and Measures, Oxford University Press, Oxford, UK, 2011.
[40] S. Benedetto, M. Pedrotti, L. Minin, T. Baccino, A. Re, R. Montanari, Driver workload and eye blink duration, Transp. Res. Part F: Traffic Psychol. Behav. 14 (3) (2011) 199–208.
[41] J. Lassalle, L. Gros, G. Coppin, Combination of physiological and subjective measures to assess quality of experience for audiovisual technologies, in: 2011 Third International Workshop on Quality of Multimedia Experience (QoMEX), IEEE, Mechelen, Belgium, 2011, pp. 13–18.
[42] G.M. Wilson, M.A. Sasse, Do users always know whats good for them? utilising physiological responses to assess media quality, in: People and Computers XIVUsability or Else!, Springer, London, UK, 2000, pp. 327–339.
[43] S. Möller, Quality Engineering. Qualität kommunikationstechnischer Systeme, Springer, Heidelberg, Germany, 2010.
[44] F.C. Bartlett, Ferrier lecture: fatigue following highly skilled work, Proc. R. Soc. London, Ser. B—Biol. Sci. 131 (864) (1943) 247–257.