DOI: 10.2352/ISSN.2470-1173.2016.16HVEI-125

Review on using physiology in quality of experience

Sebastian Arndt[1], *Kjell Brunnström*[2], *Eva Cheng*[3], *Ulrich Engelke*[4], *Sebastian Möller*[1], *Jan-Niklas Antons*[1]

[1] *Quality and Usability Lab, Technische Universität Berlin, Berlin, Germany*
[2] *ACREO Acreo Swedish ICT AB, Stockholm , Sweden Mid Sweden University, Sundsvall, Sweden*
[3] *Royal Melbourne Institute of Technology (RMIT), Melbourne, Australia*
[4] *Commonwealth Scientific and Industrial Research Organisation (CSIRO), Hobart, Australia*

Abstract

In the area of Quality of Experience (QoE), one challenge is to design test methodologies in order to evaluate the perceived quality of multimedia content delivered through technical systems. Traditionally, this evaluation is done using subjective opinion tests. However, sometimes it is difficult for observers to communicate the experienced quality through the given scale. Furthermore, those tests do not give insights into how the user is reacting on an internal physiological level. To overcome these issues, one approach is to use physiological measures, in order to derive a direct non-verbal response of the recipient. In this paper, we review studies that have been performed in the domain of QoE using physiological measures and we look into current activities in standardization bodies. We present challenges this research faces, and give an overview on what researchers should be aware of when they want to start working in this research area.

Introduction

Today's wide variety of online services leads to a huge amount of data which has to be delivered via the internet where bandwidth is one of the limiting factors. Especially, in the case of video streaming, these limitations become very obvious. However, due to the significant competition between different video streaming services, service providers cannot allow themselves to transmit their content at an unacceptable quality, as customers would abandon their services. In subjective quality tests, it can be determined which technical settings lead to an acceptable perceived quality. However, the ecological validity of the current subjective testing methodologies has been questioned [1]. Thus, methods that could be used in the user's normal environment, with as little interruptions as possible of the used service and preferable non-intrusive towards the user, would be a preferred testing methodology. Additionally, sometimes it is difficult for an observer to communicate the level of experienced quality through the rating scale presented to them, and consequently not all relevant responses can be collected using subjective ratings. This will require developing methods that can estimate quality with other means than explicitly asking the user. Here, (neuro)physiological measures may be helpful to overcome these challenges as they can be taken directly and non-verbally from the observer. Although, these measures are more difficult to gather, they measure directly a response of the observer which may even detect subtle differences that are not noticeable on the behavioral level, i.e. the level of the conscious opinion ratings.

Due to chemical and physical processes within the brain, electrical activity is being elicited that can be recorded from the scalp's surface using electrodes; for example, by an electroencephalogram (EEG). These electrical responses are directly due to neural activity and can be recorded at a very high temporal resolution. Thus, early responses can be detected [2], in comparison with hemodynamic measures, which analyze changes in blood flow and which take a few seconds until a response can be recorded [3]. Also, the apparatus' expenditure is much higher for those measures. Therefore, this paper focuses mostly on studies using EEG.

The analysis of EEG data can be performed in two different ways. Firstly, data can be analyzed concerning a short and distinct event that elicits an event-related potential (ERP). Here, the amplitude of the ERP's component can vary with the level of quality perceived by the user. Secondly, data can be studied using an analysis of the frequency band power. This is especially interesting, when drawing conclusions about the mental state of participants, or to describe the change in the mental state between conditions.

The topic of using neurophysiological assessment in the domain of quality of experience (QoE) is rather young, and only a very limited number of research has been conducted in this area. Still, there are some commonalities that will be brought together and summarized within this paper.

The basis for analyzing neurophysiological reactions towards changes in the experienced quality in audio and speech signals has been performed in studies from Miettinen [4] and Antons [5]. Based on these fundamentals, studies that have been performed in the area of video and image quality assessment using mostly measures of EEG, but also different other physiological techniques (such as near-infrared spectroscopy or electrocardiogram) will be described in detail.

This paper will give a general introduction and an overview into the topic of (neuro)physiology in QoE including, studies that have been performed. It will summarize their results and based on previous given recommendations, present tips and tricks for researchers who are new in this field. The paper is structured as follows. First, an introduction to traditional approaches in quality of experience is given, followed by an introduction into different aspects of physiology. Afterward, an extensive review on existing work combining QoE and methods of physiology is given. Following this, a section on tips and tricks when working in this area are given. At the end, an outlook towards future work concerning research and standardization bodies is given.

DOI: 10.2352/ISSN.2470-1173.2016.16HVEI-125

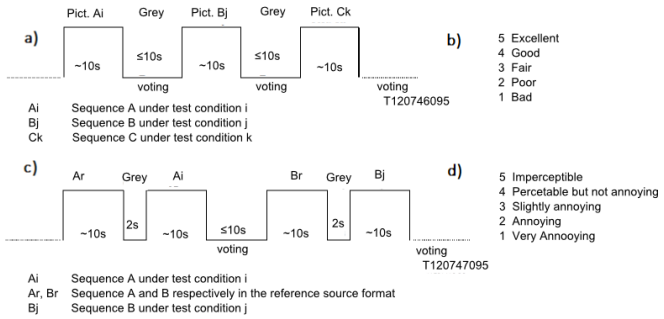

Figure 1: Illustration of (a) ACR test procedure and (b) corresponding rating scale, as well as sequence of (c) DCR test procedure and (d) their corresponding scale. Figures taken from ITU-T Rec. P.910 [6].

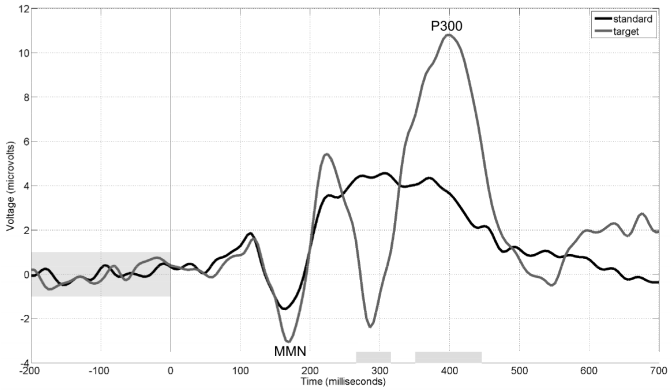

Figure 2: Example of an ERP; taken from [8]

Quality of Experience

According to [7], *Quality* is the result of an internal comparison between desired and perceived quality features. Whereby, "a *quality feature* is the perceived characteristics of an entity that is relevant to the entity's quality." To understand this relation various methods have been developed. Most of them are based on the presentation of stimuli and asking the observer to give a response, e.g. on a rating scale. In standard procedures recommended by the International Telecommunication Unit (ITU), stimuli are usually presented individually, one by one (Absolute Category Rating, ACR), or in pairs, two subsequent stimuli (Degradation Category Rating, DCR), to the test participant. After presentation, the observers have to judge the quality on a scale depending on the paradigm used, as shown in Fig. 1. The averaged judgment for one condition over all participants, in ACR tests is the so-called Mean Opinion Score (MOS), and in DCR tests the DMOS.

Unfortunately, MOS or DMOS values obtained in these tests usually do not give any information about the insights into the observers' state. To overcome this issue, measures which can be taken directly from the participant receive more attention. In this case, without explicitly asking the participants, information can be obtained which is potentially related to the perceived quality.

Neurophysiology

In the case of subjective opinion tests, the internal judgment of the observer somehow has to be encoded onto the scale that is given to the participant. It is a challenge in the design of these tests to select a good scale, and to make the observer understand it through the instructions given. On the other hand, physiological measures are derived directly from the participant and do not directly undergo this encoding process; thus, they may give a less biased result which is free e.g. of personal preferences or misunderstanding of the used scale.

The *electroencephalogram (EEG)* may be used for assessing quality related processes [5]. The *EEG* measures voltage variation due to neuron activity in the brain. It can be recorded by attaching electrodes to the scalp of a subject. Since its discovery by Berger in 1929, it has become a widely used method for investigating physiological correlation between perceptual and attentional processes [9] [10]. This measure has a rather limited spatial resolution – based on the fact that the brain is a wet conductor the signal recorded by one electrode is a mixture of all existent sources – but at excellent temporal resolution with a precision of milliseconds. The corresponding data can mainly be analyzed in two different ways: on the one hand by having a closer look at the spectrogram of spontaneous activity, and on the other hand by analyzing so called *Event-Related-Potentials (ERP)* which are a time-locked reaction to an external stimulus in terms of a voltage change [11] (see Fig. 2 for an example ERP). The latter approach can be used to analyze cortical potentials as well as voltage differences evoked in the brain stem. The focus of this review will be on the cortical brain activity because research on brain stem level is not yet fully usable in QoE-research and only very limited work has been performed here.

In addition to the relevant information – brain activity – other unwanted information is recorded as well, e.g. voltage changes due to eye-movement, body movement and other unrelated signal sources. Due to high noise present in the EEG signal, it is important to create highly controlled experimental setups. Clinical research guidelines for experimental designs already exist and carry important implications for research in the domain of Quality of Experience based on them [10].

The equipment that is used in EEG research typically is very expensive and in case of using wet electrodes also challenging to attach to the participant. Lately, new low-cost EEG devices have appeared on the market, such as the Emotiv-EPOC [1] and NeuroSky MindWave [2] headsets. While these consumer products are comparably inexpensive and easy to attach, the data quality, i.e. precision and noisiness of the signal, using those products is expectedly less reliable to the devices used in clinical applications. However, these products have shown to capture useful information in the context of QoE-related research.

For the analysis of ERPs, a small set of electrodes can be sufficient, usually up to 8 electrodes; they should be distributed along the central line following the *10/20 system* [12], and for hemispheric differences equally distributed electrodes over the right and left hemisphere are advisable. More electrodes are needed for the analysis of more complex patterns e.g. spatial pattern distribution.

As evoked potentials depend on an exact timing, it is important that triggers are exactly synchronized to be able to average the signal while keeping the temporal information intact. *ERPs*

[1] http://www.emotiv.com/
[2] http://www.neurosky.com/

DOI: 10.2352/ISSN.2470-1173.2016.16HVEI-125

cannot be observed in the raw *EEG* as they are overshadowed by other, unrelated activity, which is smoothed out when averaging several trials of single *ERP* recordings.

Usually 20–30 trials at minimum are needed for an average ERP per stimulus class; a baseline corrected signal uses the average value of the voltage in the interval of up to 200 ms before the stimulus. This rather high number of trials compared to standard quality tests also explains the typically small number of subjects used for EEG-studies.

The aforementioned averaging methods are performed offline and as an average over a group of subjects. This average over all subjects is the grand average, and is the result which is often plotted in these studies, see Fig. 2. Using classification techniques this can be transferred to online analysis of incoming physiological signals, such as deciding whether the brain activity of the proceeding stimulus was evoked by one special class of stimuli [13]. To address this in the case of Quality of Experience an exemplary class of degradation can be used. With classification as a measure of separability, it can be distinguished between perceived stimulus classes. For a tutorial on single-trial *ERP* classifications see [14].

In the *continuous EEG*, five main different frequency ranges are ascribed to specific states of the brain: *delta band* (1–4 Hz), *theta band* (4–8 Hz), *alpha band* (8–13 Hz), *beta band* (13–30 Hz) and the *gamma band* (36–44 Hz) [15]. The *delta band* is present during deep sleep, the *theta band* occurs during light sleep and is an indicator for decreased alertness. Activity in the *alpha band* is related to relaxed wakefulness with eyes closed and decrease in alertness. *Beta and gamma band* are ascribed to high arousal and focused attention [11].

Analyzing the power in the before mentioned frequency bands is widely done for assessing the *cognitive state* of car drivers. Lal *et al.* for example showed that fatigued drivers had an elevated power in the delta and theta bands [16]. Correlation between weighted combinations of the power in different frequency bands with subjective fatigue ratings was shown in [17].

Another reason to use frequency bands is to estimate the emotional state of subjects. Therefore, alpha values from frontal electrodes are being extracted. The asymmetry index is one way to obtain this information. It shows that higher values in the asymmetry index are the result of higher left frontal activity which is due to rather negative emotional processing [18].

Neurophysiology in Multimedia Quality Perception

In all areas where the classical QoE research is active, studies have been conducted which use a variety of physiological methodologies. In this review mainly neurophysiological measures are being focused on. First studies by Mietnnen have used audio stimuli and MEG [4]. Later, Antons *et al.* used audio stimuli and EEG [19], and from there on several different groups investigated independently different multimedia material using mostly EEG. An overview on the different studies can be seen in Table 1.

In the following, we review work related to the two main paradigms as defined earlier: ERP and spectral analysis.

ERP

A first study using classes of degradations that are of interest for research in the telecommunication industry was conducted by Miettinen *et al.* (2010) using *magnetoencephalography (MEG)*, where they could show a significant increase in the measured amplitudes for distorted stimuli [37].

The following sections are based on the work of [5]. One of the first studies using EEG for *quality assessment* were conducted by Antons *et al.* in the auditory domain, where signal-correlated noise was introduced into the stimuli and the signal-to-noise ratio was the independent and scalable variable [19]. Here, participants had to judge after each presentation whether they perceived a distortion in the current stimulus or not. The first paradigm using *EEG* in a *QoE* context was derived starting with meaningless syllables and developing the stimulus up to random words. In each of the experiments it could be shown that the elicited *P300* becomes significantly higher the more distorted the stimulus is [21]. Additionally, the *P300* occurs earlier with stronger distortions. Furthermore, it could be shown that stimuli that were perceived as undistorted by the participants, but were distorted on the signal level, had a similar trend in the *ERP* as trials rated by the participant as distorted. Thus, high machine-learning classification rates for these trials could be obtained and it was concluded that these degradations are presumably processed non-consciously as they do not penetrate up to the subjective behavior [19]. Aim of classification was to identify trials in which the participant was not able to detect a degraded stimulus, although an activation pattern similar to conscious detection was present. Linear Discriminant Analysis (LDA) with automatic regularization of the estimated covariance matrix was applied.

A series of studies using short (audio)visual stimuli snippets was presented by Arndt *et al.* [38] which consisted of five individual experiments. In this series the same video was used for all studies. The video consisted of a speaker uttering the syllable /pa/. The video was presented with different levels of quality, and while test participants were watching those videos an EEG was recorded. The quality degradation was achieved by using different levels of inserted artificial blockiness, ranging from hardly noticeable up to very annoying. The procedure for one trial was that first the standard (undistorted) video was shown, followed directly by a possibly distorted variant of the video. In the first study only the video was presented to the participants (*Video Experiment*). In a follow-up study half the trials were presented with accompanying audio, and all the remaining studies were presented with an audio track. Here, either audio, video or both modalities were distorted. For the distortion, artificial blockiness was introduced for the video which was generated as described in ITU-T Recommendation P.930 [39]. For the audio distortion, signal-correlated noise was generated using the modulated noise reference unit (MNRU) that is described in ITU-T Rec. P.810 [40]. In the final experiment, an actual codec, namely the H.264 in the x264 implementation, was used as realistic degradations. For analysis of the recorded EEG data, time-locked epochs around the beginning of the video snippet were extracted. The main component which was being investigated was the P300. In these experiments, it was shown that the P300 was larger for low quality stimuli (i.e. larger blockiness artifacts) compared to higher quality stimuli where the degradation was less visible. In the paper, it is argued that this is due to higher cognitive processing in case of low-quality presentations.

In work conducted by Lindemann *et al.* [32], static JPEG compressed images were presented to the subjects. Four dif-

DOI: 10.2352/ISSN.2470-1173.2016.16HVEI-125

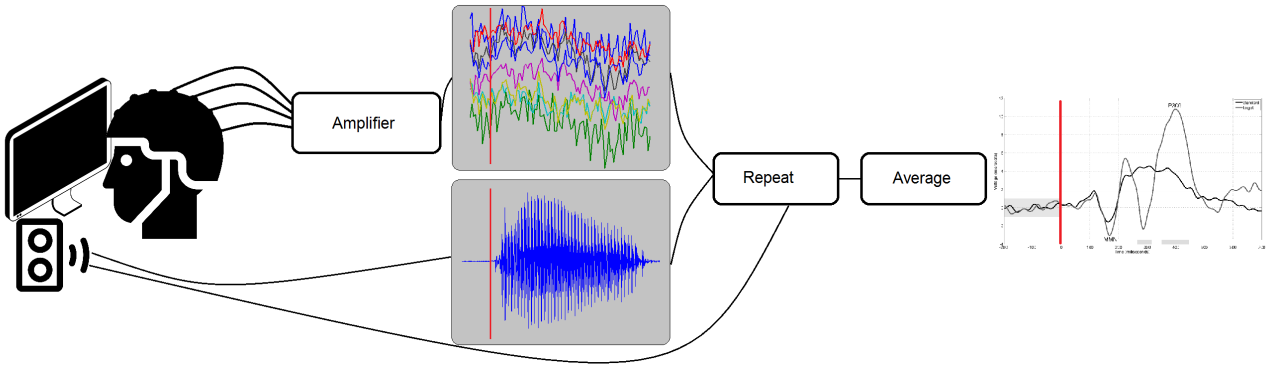

Figure 3: Path of stimulus presentation and EEG data analysis. The subject is listening to audio samples, the EEG is recorded simultaneously (the red line indicates stimulus onset). Stimulus presentation and EEG recording are repeated several times within one session, then the recorded EEG data is averaged and plotted (description from left to right).

Author	Study	Media Type	Artifact (#Distortion Levels)	#subjects	EEG Feature
Acqualagna	Texture Images [20]	Image	HM 10.0 (7)	16	SSVEP
Antons	Phoneme [21]	Speech	MNRU (4)	10	P300
	Word [19]	Speech	G.722.2 (4)	9	P300
	Sentence [5]	Speech	reverberation (3)	22	P300
	Audiobook I [22]	Speech	G.722.2 (2)	18	alpha
	Audiobook II [23]	Speech	G.722.2 (4)	12	alpha
Arndt	Video Experiment [24]	Video	artificial blockiness (6)	10	P300
	Audio /on/off [25]	(Audio)visual	artificial blockiness (4)	10	P300
	Audiovisual Exp 1 [26]	Audiovisual	artificial blockiness/MNRU (8)	13	P300
	Audiovisual Exp 2	Audiovisual	artificial blockiness/MNRU (8)	12	P300
	x264 Experiment	Audiovisual	x264 codec (4)	10	P300
	AV-Documentary - Audio [27]	Audiovisual	x264 codec (2)	12	alpha, theta
	AV-Documentary - Speech [28]	Audiovisual	x264 codec/GSM1.60 (4)	24	alpha, theta
	TTS [29]	Speech	synthetic speech (4)	14	alpha, P300
Beyer	Cloud Gaming [30]	Games	x264 (2)	32	alpha
Kroupi	2D vs 3D [31]	Video	2D vs 3D (4)	16	frequency
Lindemann	compressed images [32]	Image	JPEG compression (7)	10	P300
	zooming [33]	Image	different colored block (3)	8	P300
Moon	HDR videos [34]	Video	HDR vs LDR (2)	5	frequency
Mustafa	Video Artifacts [35] [13]	Video	popping/blurring/ghosting (6)	8	P300, alpha
Nunez Castellar	Flow Experience	Games	game level (3)	22	P300
Scholler	Chess Grid [36]	Video	HEVC (6)	9	P300

Table 1: Overview on studies analyzing quality aspects using measures of EEG. The table shows a summary of the media types used, induced artifacts, distortion levels, number of participants, and considered EEG feature.

DOI: 10.2352/ISSN.2470-1173.2016.16HVEI-125

ferent images were presented at seven different levels of JPEG-compression. In each trial, first the undistorted image was shown (standard stimulus) and subsequently the possible distorted image (target stimulus). Event-related potentials around the presentation of the target stimulus were extracted from the EEG recordings. The recorded ERP responses showed that lower quality compressed images elicited a stronger P300 component compared to higher quality images. In a second study [33], a different set of three images was used. To avoid sudden onset effects, zooming into the picture's center was emulated. During this zoom-in, two possible artifacts were inserted, one rather obvious (a magenta block) and one less obvious (a blurred block) distraction. For the EEG data, ERPs were extracted around the onset of the artifact. The data shows a stronger P300 peak for the obvious artifacts and a lower deflection for the less obvious one; this is in line with the previously reported results. In addition, classification on the EEG data was performed using support vector machine (SVM). Classification between trials with artifacts and trials without artifacts yield a high accuracy of up to 80% while classification between all trials show good classification rate for no-artifacts (75%) but poor for trials with artifacts (54%).

A study performed by Scholler *et al.* [36] used synthetic visual stimuli (with a duration of 8 s), which consisted of a chess grid with water rings on top. The video was distorted using a codec similar to the HEVC and introduced six different levels of distortion. One trial consisted of one video with the inserted distortion starting randomly between 2 s and 6 s. The participant's task was to report after each trial whether they perceived a degradation or not. The ERP was extracted around the onset of the inserted distortion. It was shown that stronger degradations lead to a larger P300 amplitude. Subsequently, classification was performed on the ERPs using LDA (linear discrimination analysis). In the case of trials which were correctly identified by the test participants (i.e. of either containing degradation or not), the strongest distortion levels were classified almost perfectly with an AUC (area under the curve) of close to 1. In the case of less obvious levels of degradation, the AUC decreases. For trials that were not identified correctly by the participants (i.e. videos that contained an impairment which was not recognized) classification was only for a few participants above chance.

In Acqualagna *et al.* [20], a set of six different images with gray level texture was selected. Each image was presented at seven different compression levels produced by the HM10.0 HVEC codec (including one reference condition). The images were presented using 3Hz flickering between the uncompressed and the distorted image. In contrast to the other studies presented in this section, the authors evoked a steady state visual evoked potential (SSVEP) which is a response being elicited by the visual cortex when presenting flickering visual stimuli [41]. The results show that the obvious compression levels lead to an increase in the modulated frequency power (i.e. 3 Hz). When applying LDA classification, the classification accuracy AUC is well above chance for the obvious compression levels, but not for the less distorted stimuli.

The overview of studies shows that they were carried out at different labs, using different video material as well as different EEG equipment (all clinical, though). In most studies, a comparison between stimuli was the task for the test participant, to decide whether they perceived a distortion between the standard and a possibly degraded stimulus. This is a modification of the oddball paradigm which is one of the standard paradigms in neuroscience in order to elicit a P300 response, but is much more difficult to implement in the domain of quality research, and thus may bring less obvious results than it would in the original paradigm. It can be seen that the P300 component is most variable with differently degraded visual stimuli. In Arndt's studies both modalities were presented to the subject and were varied in their quality. Here, it was shown that the presented effect is stable across modalities. Additionally, all of these studies have in common that they use very short video stimuli for their experimental paradigms, as well as lots of repetitions (which is mostly due to the nature of an ERP) and thus do not conform with standard quality recommendations [6] [42] nor with realistic settings. In the study of Acqualagna [20], the SSVEP, a different feature of the EEG signal, was used and could confirm prior results. As for SSVEPs, a lower number of repetitions necessary for this paradigm might be better suited for quality assessment of still images.

Spectral Analysis

Due to the possibility that more natural stimuli in terms of stimulus length can be used, it is possible to examine the effect of longer duration media stimuli (>10 min) on the recipients. In this section studies that used longer lasting stimuli are mentioned.

In studies by Antons *et al.*, participants were exposed to high quality and low quality sequences of longer auditory material. Their only task was to rate the content on a scale every few minutes, and in the rest of the time they should focus on the presented content. Higher values in the *alpha band* power were observed when being exposed to low quality stimuli compared to higher quality stimuli, which is ascribed to fatigue and impaired information processing [22]. In an additional study, results assessed the impact of a high quality audio segment (5 min) inserted in a low quality audio stimulus (15 min). Participants were less fatigued due to the better audio quality as indicated by a lower *alpha band* power [23].

Arndt *et al.* [43] conducted two studies in which longer stimuli that were equivalent to documentaries have been used; the documentaries showed sea life scenes. In the first study, for the audio only scene-related background noise was present. One half of the audiovisual material was presented in high-quality video and the other half in low-quality video. The second study had an accompanying background narrator that was constantly talking. Here, both modalities were presented in their original quality as well as in reduced quality, either only one or both modalities were distorted, leading to four different quality levels. Within both studies, the state of the test participants and how it changed with presenting different quality levels was analyzed. For the EEG data, a frequency power band analysis was performed. In both studies, it could be shown that an increase in alpha and theta level is the result of a reduction in quality. Compared with standard literature in neuroscience this increase is due to an increased level of fatigue and/or drowsiness. In both studies, several other physiological measures have been recorded additionally. For the first study, blink duration was one parameter which had been analyzed in more detail. It could be shown that for the low-quality sequence the blink duration was longer compared to the high-quality version [44]. This is an additional indicator of an higher level of fatigue. In the second study, several peripheral physiological mea-

DOI: 10.2352/ISSN.2470-1173.2016.16HVEI-125

sures were derived, amongst others electrocardiogram or the level of skin conductance. However, those measures did not show any significant change with the change of quality [28].

Mustafa *et al.* [13] used a 5.6 s long low-complexity video that showed a person walking. They introduced different kinds of distortion, such as partial freezing and blurring (on the person, and on a greater area), ghosting of the movement plus the original video, resulting into six different video conditions. For the EEG data, ERPs were extracted as well as performing frequency analysis. For analysis of the ERPs [35], the data was extracted around the onset of the distortion. It could be shown that for artifacts that are on the moving part of the video (i.e. the person) the P300 is larger compared to distortions that contain a greater area. In addition, power frequency analysis was performed [13] on the same data set. When analyzing the frequency range of 10 Hz to 20 Hz, they could show that artifacts on the person evoke again the greatest change compared to the reference condition. Classification on the EEG data using a SVM approach show high classification results for differentiating between trials containing artifacts and reference trials (85%). During single-trial classification on a per artifact basis, classification results only reach up to 70%.

Kroupi *et al.* [31] analyzed degradation of 2D and 3D videos where they used seven different one minute long sequences from a music festival. Each sequence was presented in a high-quality (HQ) and low-quality (LQ) version in 2D and 3D rendering. The different power frequency bands from the EEG data were extracted and correlated with the behavioral quality judgments. The main finding was that results of the EEG recordings show high frontal asymmetry in the alpha power band, which reflects emotional affect towards the two different quality levels.

Moon *et al.* [34] used four different sequences of HDR (high dynamic range) and LDR (low dynamic range) video content. The sequences showed different scenes and had a duration of 20 s to 50 s. The frequency power band was extracted from the EEG data and used for classification, with an accuracy of almost 70% if only EEG features were used, and improved up to almost 80% if other peripheral measures were used for classification as well (e.g. GSR, respiration, Plet, skin temperature). For the EEG, the gamma band seemed to give the most discriminative results between conditions.

Moldovan *et al.* used the features provided by the Emotiv EPOC System to infer the level of frustration from the human observer caused by the quality of the played audiovisual excerpt. This level was determined by using a metric predefined by the headset manufacturer. In their study, videos with different levels of quality were used. The level of quality was controlled through manipulation of the bitrate, frame rate as well as resolution of the presented video clips [45]. Perez *et al.* used the NeuroSky MindWave headset to measure brain activity and used the recorded data to classify the trials into high and low quality pictures [46].

This section shows that there is more variability between studies compared to the section on ERPs. One approach is to design setups closer towards realistic stimuli and therefore, rather research the impact of longer degraded low-quality sequences on the participant's state. Other approaches still use rather shorter stimuli and examine the difference in the cognitive state due to different media stimuli (2D vs 3D). Also consumer grade products were in use and provided insights into the change of emotional responses in the observer.

Discussion

The number of different studies and the variety of different labs conducting these kinds of studies show that there is big interest in this emerging research area. Furthermore, the presented results are in line with each other and with standard neurophysiological literature, which suggests that the measured components in the recorded EEG data are useful for such an approach. Obviously, the identified components do not solely represent quality related features, but make use of standard EEG paradigms and their components coming from psychology. The P300, which is a measure of difference detection, was used to be elicited when there was a difference between stimuli which in the context of this paper were related to quality features. The other major component used was the alpha band power that is indicative for higher levels of fatigue or relaxed wakefulness. In the context of the presented studies, the manipulation of the material was due to quality impairments.

As shown in this paper, two completely different approaches were used throughout a number of studies to analyze influences of quality degradation on the observer. It can be seen that if short stimuli are used, analysis of event-related potentials is being applied. Here, a relatively high number of repetitions has to be performed in order to obtain a smooth and averaged ERP signal. But if this is achieved, single-trial classification can be applied and subtle differences can be analyzed much better and in some cases even more sensitive than in solely behavioral studies. Especially in these cases the advantage of neurophysiological measures is striking, as technical settings could be improved using such methods.

How these short reactions may influence the observer in longer scenarios was analyzed in the second part of this paper. Here, the analysis of frequency components of the EEG data, and with this the approximation of change of the observer's cognitive state, was presented. It could be shown that due to quality reduction a change in the analyzed frequency components was visible. This change can be ascribed to a reduction of the cognitive state as mentioned in [43].

The studies mentioned in this paper all were passive experiments; thus, no direct interaction of the test participant was required. Recently there also have been efforts to use EEG in the area of gaming QoE. Especially, while games are emerging that are run over the Internet, classical QoE problems (paired with additional problems) evolve. In [30], the authors measured varying alpha activity with different levels of video compression. In this study participants played a first person shooter game in a cloud gaming setup with varying levels of video quality caused by different video compression bitrates. It was found that the video quality influenced the perceived quality, player experience, the subjective ratings and the alpha frequency band power. It is shown that physiological measures capture the influence on the player in terms of a reduced cognitive state.

Neural Correlates of the Subjective Flow Experience During Game Play are analyzed in [47]. Brain activity associated with this subjective experience of the attenional flow state has sparkled interest recently. A response locked laplacian transformed EEG data analyses revealed increased activity at fronto-central electrodes (activity maximal at FCz - Cz) around 250ms following the response-onset to oddball sounds in the flow condition. It is hypothesized that the medial frontal activity locked to the re-

DOI: 10.2352/ISSN.2470-1173.2016.16HVEI-125

sponse onset could reflect increased cognitive control, prompted by the high attentional demands in the flow condition. Likewise increased activity elicited after the detection of the novelty sounds (P300) might be an index of the to re-allocation attentional resources to the primary task.

Current activities and Look-out

Currently no standards on using physiology in the area of QoE exists. Therefore, this topic has been brought to the ITU where several contributions on using physiological measures in QoE have been made [48] [49] [50] [51]. Furthermore, activities have been initiated to form a recommendation to use physiological measures in addition to subjective tests [itupphysio], under the working item *P.PHYSIO*. Additionally, the topic was presented to the European Telecommunications standards institute (ETSI) to receive more attention[52].

Within the video quality experts group (VQEG), the RICE project (Real-Time Interactive Communications Evaluation) [53] picked up this topic. Current activities consist of a study (which currently is in the planning phase) in which several labs will be involved. Here, the goal is to test the methodology for its robustness. Within the different labs not only different kind of equipment will be used, but also the effects of experimenters which may be new to this assessment technique and different cultural backgrounds will be analyzed.

References

[1] K. De Moor, M. Fiedler, P. Reichl, and M. Varela, "Quality of Experience: From Assessment to Application (Dagstuhl Seminar 15022)," 2015.

[2] D. Purves, E. M Brannon, R. Cabeza, S. A Huettel, K. S LaBar, M. L Platt, and M. G Woldorff, *Principles of cognitive neuroscience*, vol. 83, Sinauer Associates Sunderland, MA, 2008.

[3] T. Wager, L. Hernandez, J. Jonides, and M. Lindquist, "2 elements of functional neuroimaging," *Handbook of psychophysiology*, p. 19, 2007.

[4] I. Miettinen, H. Tiitinen, P. Alku, and P. JC May, "Sensitivity of the human auditory cortex to acoustic degradation of speech and non-speech sounds," *BMC neuroscience*, vol. 11, no. 1, pp. 24, 2010.

[5] J.-N. Antons, *Neural Correlates of Quality Perception for Complex Speech Signals*, Springer, Cham, 2015.

[6] ITU-T Recommendation P.910, "Subjective video quality assessment methods for multimedia applications," *International Telecommunication Union, Geneva*, 2008.

[7] P. Le Callet, S. Möller, A. Perkis, et al., "Qualinet white paper on definitions of quality of experience," *European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003)*, 2012.

[8] J.-N. Antons, S. Arndt, R. Schleicher, and S. Möller, "Brain activity correlates of quality of experience," in *Quality of Experience*, S. Möller and A. Raake, Eds., chapter 8, pp. 109–119. Springer International Publishing, 2014.

[9] H. Berger, "Über das Elektrenkephalogramm des Menschen," *Arch f Pschiatr*, pp. 527–570, 1924.

[10] C. Duncan, R. Barry, J. Connolly, C. Fischer, P. Michie, R. Näätänen, J. Polich, I. Reinvang, and C. Petten, "Event-Related Potentials in Clinical Research: Guidelines for Eliciting, Recording, and Quantifying Mismatch Negativity, P300, and N400," *Clinical Neurophysiology*, pp. 1883–1903, 2009.

[11] M. S. Coles and M. Rugg, "Event-related brain potentials: an introduction," in *Electrophysiology of Mind: Event-Related Brain Potentials and Cognition*, M. S. Coles and M. Rugg, Eds. Oxford University Press, Oxford, 1995.

[12] American Clinical Neurophysiology Society, "Guideline 5: Guidelines for Standard Electrode Position Nomenclature," *Journal of Clinical Neurophysiology*, vol. 23, no. 2, 2006.

[13] M. Mustafa, S. Guthe, and M. Magnor, "Single Trial EEG Classification of Artifacts in Videos," *ACM Transactions on Applied Perception (TAP)*, vol. 9, no. 3, pp. 12:1–12:15, July 2012.

[14] B. Blankertz, S. Lemm, M. S. Treder, S. Haufe, and K.-R. Müller, "Single-Trial Analysis and Classification of ERP Components - A Tutorial," *Neuroimage*, vol. 56, pp. 814–825, 2011.

[15] D. A. Pizzagalli, "Electroencephalography and high-density electrophysiological source localization," in *Handbook of Psychophysiology*, G. Berntson J. Cacippio, L. Tassinary, Ed. Cambridge University Press, Cambridge, 1995.

[16] S. Lal and A. Craig, "Reproducibility of the Spectral Components of the Electroencephalogram During Driver Fatigue," *International Journal of Psychophysiology*, pp. 137–143, 2005.

[17] Y. Punsawad, S. Aempedchr, Y. Wongsawat, and M. Panichkun, "Weighted-Frequency Index for EEG Based Mental Fatigue Alarm System," *International Journal of Applied Biomedical Engineering*, pp. 36–41, 2011.

[18] J. A. Coan and J. Allen, "Frontal EEG asymmetry as a moderator and mediator of emotion," *Biological Psychology*, pp. 7–50, 2004.

[19] J.-N. Antons, R. Schleicher, S. Arndt, S. Möller, A.K. Porbadnigk, and G. Curio, "Analyzing Speech Quality Perception Using Electroencephalography," *J. Select. Topics Signal Proc.*, pp. 721–731, 2012.

[20] L. Acqualagna, S. Bosse, A. K Porbadnigk, G. Curio, K.-R. Müller, T. Wiegand, and B. Blankertz, "EEG-based classification of video quality perception using steady state visual evoked potentials (SSVEPs)," *Journal of neural engineering*, vol. 12, no. 2, pp. 026012, 2015.

[21] J.-N. Antons, A. K. Porbadnigk, R. Schleicher, B. Blankertz, S. Möller, and G. Curio, "Subjective listening tests and neural correlates of speech degradation in case of signal-correlated noise," in *Audio Engineering Society (AES) 129th Convention*. 2010, Curran Associates, Inc.

[22] J.-N. Antons, R. Schleicher, S. Arndt, S. Möller, and G. Curio, "Too tired for calling? a physiological measure of fatigue caused by bandwidth limitations," in *Proc. Quality of Multimedia Experience (QoMEX)*, 2012.

[23] J.-N. Antons, F. Köster, S. Arndt, S. Möller, and R. Schleicher, "Changes of vigilance caused by varying bit rate conditions," in *Proc. Quality of Multimedia Experience (QoMEX)*, 2013.

[24] S. Arndt, J.-N. Antons, R. Schleicher, S. Möller, S. Scholler, and G. Curio, "A physiological approach to determine video quality," in *Proc. 2011 IEEE International Symposium on*

DOI: 10.2352/ISSN.2470-1173.2016.16HVEI-125

Multimedia, 2011, pp. 518–523.

[25] S. Arndt, J. Bürglen, J.-N. Antons, R. Schleicher, and S. Möller, "Einfluss der Audiomodalität auf die Wahrnehmung und Qualitätsbeurteilung (audio-)visueller Stimuli," in *Proc. 2013 Berliner Werkstatt Mensch Maschine Systeme*, 2013.

[26] S. Arndt, J.-N. Antons, R. Schleicher, S. Möller, and G. Curio, "Percpetion of low-quality videos analyzed by means of electroencephalography," in *Proc. 2012 IEEE Quality of Multimedia Experience*, 2012.

[27] S. Arndt, R. Schleicher, and J.-N. Antons, "Does Low Quality Audiovisual Content Increase Fatigue of Viewers?," in *4th International Workshop on Perceptual Quality of Systems (PQS)*, 2013.

[28] S. Arndt, J.-N. Antons, and S. Möller, "Is low quality media affecting the level of fatigue?," in *Sixth Int. Workshop on Quality of Multimedia Experience (QoMEX 2014)*, 2014, pp. 47–48.

[29] S. Arndt, J.-N. Antons, R. Gupta, R. Schleicher, S. Möller, and T. H Falk, "Subjective quality ratings and physiological correlates of synthesized speech," in *Quality of Multimedia Experience (QoMEX), 2013 Fifth International Workshop on*. IEEE, 2013, pp. 152–157.

[30] J. Beyer, R. Varbelow, J.-N. Antons, and S. Möller, "Using electroencephalography and subjective self-assessment to measure the influence of quality variations in cloud gaming," in *Proc. 2015 IEEE Quality of Multimedia Experience*, 2015, pp. 1–6.

[31] E. Kroupi, P. Hanhart, J.-S. Lee, M. Rerabek, and T. Ebrahimi, "Eeg correlates during video quality perception," in *Signal Processing Conference (EUSIPCO), 2014 Proceedings of the 22nd European*. IEEE, 2014, pp. 2135–2139.

[32] L. Lindemann and M. Magnor, "Assessing the quality of compressed images using EEG," in *Image Processing (ICIP), 2011 18th IEEE International Conference on*. IEEE, 2011, pp. 3109–3112.

[33] L. Lindemann, S. Wenger, and M. Magnor, "Evaluation of video artifact perception using event-related potentials," in *Proc. ACM Applied Perception in Computer Graphics and Visualization (APGV) 2011*, Aug. 2011, p. 5.

[34] S.-E. Moon and J.-S. Lee, "Perceptual experience analysis for tone-mapped hdr videos based on eeg and peripheral physiological signals," *Autonomous Mental Development, IEEE Transactions on*, 2015.

[35] M. Mustafa, L. Lindemann, and M. Magnor, "Eeg analysis of implicit human visual perception," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2012, pp. 513–516.

[36] S. Scholler, S. Bosse, M.S. Treder, B. Blankertz, G. Curio, K.-R. Müller, and T. Wiegand, "Towards a Direct Measure of Video Quality Perception using EEG," *IEEE Transactions on Image Processing : a publication of the IEEE Signal Processing Society*, vol. 21, no. 5, pp. 2619–2629, 2012.

[37] I. Miettinen, H. Tiitinen, P. Alku, and P. May, "Sensitivity of the Human Auditory Cortex to Acoustic Degradation of Speech and Non-Speech Sound," *BMC Neuroscience*, pp. 1471–2202, 2010.

[38] S. Arndt, J.-N. Antons, R. Schleicher, S. Möller, and G. Curio, "Using Electroencephalography to Measure Perceived Video Quality," *Selected Topics in Signal Processing, IEEE Journal of*, pp. 366–376, 2014.

[39] ITU-T Recommendation P.930, "Principles of a reference impairment system for video," *International Telecommunication Union, Geneva*, 1996.

[40] ITU-T Recommendation P.810, "Modulated noise reference unit (MNRU)," *International Telecommunication Union, Geneva*, 1996.

[41] Christoph S Herrmann, "Human eeg responses to 1–100 hz flicker: resonance phenomena in visual cortex and their potential correlation to cognitive phenomena," *Experimental brain research*, vol. 137, no. 3-4, pp. 346–353, 2001.

[42] ITU-T Recommendation P.911, "Subjective audiovisual quality assessment methods for multimedia applications," *International Telecommunication Union, Geneva*, 1998.

[43] S. Arndt, J.-N. Antons, R. Schleicher, and S. Möller, "Using electroencephalography to analyze fatigue due to low-quality audiovisual stimuli," *Accepted at Signal Processing: Image Communication*, 2016.

[44] R. Schleicher, S. Arndt, and J.-N. Antons, "Changes in blinking behavior while watching videos with reduced quality," in *ECEM 2013*, 2013, pp. 1–2.

[45] A.-N. Moldovan, I. Ghergulescu, S. Weibelzahl, and C.H. Muntean, "User-centered EEG-based Multimedia Quality Assessment," in *Proc. International Symposium on Broadband Multimedia Systems Broadcasting*, 2013.

[46] J. Perez and E. Deléchelle, "On the measurement of image quality perception using frontal EEG analysis," in *International Conference on Smart Communications in Network Technologies*, 2013.

[47] Elena Patricia Núñez Castellar, Jan-Niklas Antons, and Jan Van Looy, "Investigating the Neural Correlates of the Subjective Flow Experience During Game Play: an EEG Study," in *18th Annual Meeting of the Organization for Human Brain Mapping*, Honolulu, Hawai, 2015, p. 1.

[48] ITU-T Contribution COM 12-039, "Investigating the subjective judgment process using physiological data," *International Telecommunication Union, Geneva*, 2013.

[49] ITU-T Contribution COM 12-112, "Using physiological data for assessing variations of the cognitive state evoked by quality profiles," *International Telecommunication Union, Geneva*, 2013.

[50] ITU-T Contribution COM 12-103, "Using physiological data for assessing subjective video quality ratings," *International Telecommunication Union, Geneva*, 2013.

[51] ITU-T Contribution COM 12-202, "Using physiological data for assessing the audiovisual quality of longer stimuli," *International Telecommunication Union, Geneva*, 2014.

[52] S. Arndt, J.-N. Antons, and S. Möller, "Using electroencephalography to gain insights into the perception of different levels of quality," *Proceedings of the ETSI Workshop on Telecommunication Quality and beyond 2015*, 2015.

[53] "RICE project of VQEG," http://www.its.bldrdoc.gov/vqeg/projects/rice/rice.aspx.

Author Biography

Sebastian Arndt *is a researcher at the Quality and Usability Lab of the Telekom Innovation Laboratories, TU Berlin. He*

DOI: 10.2352/ISSN.2470-1173.2016.16HVEI-125

studied Computer Science at Technische Universität Berlin, and University of Oklahoma (USA), and received his diploma in 2010. He received his doctoral degree (Dr.-Ing.) in 2015 with the thesis title 'Neural Correlates of Quality During Perception of Audiovisual Stimuli'. His current research focus is on physiological changes during the perception of audiovisual quality.

Sebastian Möller *studied electrical engineering at the universities of Bochum, Orléans, and Bologna. He received the Doctor-of-Engineering degree in 1999 and the Venia Legendi with a book on the quality of telephone-based spoken dialog systems in 2004. In 2005, he joined Telekom Innovation Laboratories, TU Berlin, and in 2007, he was appointed Professor for Quality and Usability at TU Berlin. His primary interests are in speech signal processing, speech technology, and quality and usability evaluation.*

Ulrich Engelke *received a PhD degree in Telecommunications from the Blekinge Institute of Technology, Sweden, in 2010. He is currently a senior research scientist in Cognitive Informatics at the Commonwealth Scientific and Industrial Research Organisation (CSIRO) in Hobart, Australia. He is on the board of the Video Quality Experts Group (VQEG) and on the technical programme committee of the IS&T Human Vision and Electronic Imaging (HVEI) conference. His research interests include visual analytics, perceptual imaging, and cognitive informatics.*

Kjell Brunnström*, Ph.D., is a Senior Scientist at Acreo Swedish ICT AB and Adjunct Professor at Mid Sweden University. He is Co-chair of the Video Quality Experts Group (VQEG). He has written about 100 peer-reviewed scientific journal articles and international conference papers as well as served as reviewer for scientific journals and international conferences. His research interests are in Quality of Experience for visual media in particular video and display quality assessment.*

Jan-Niklas Antons *works at the Telekom Innovation Laboratories since 2014 as a senior research scientist. He received his diploma in psychology in 2008 from the Technische Universität Darmstadt, a Doctor-of-Engineering degree in 2014 from the Technische Universität Berlin and has been doing research at the Quality and Usability Lab at the Technische Universität (TU) Berlin, since. His research interests are in Quality-of-Experience evaluation and its physiological correlates with an emphasis on media transmissions.*