Analyzing Speech Quality Perception Using Electroencephalography

Jan-Niklas Antons, Robert Schleicher, Sebastian Arndt, Sebastian Möller, *Senior Member, IEEE*, Anne K. Porbadnigk, and Gabriel Curio

***Abstract*—Common speech quality evaluation methods rely on self-reported opinions after perceiving test stimuli. Whereas these methods—when carefully applied—provide valid and reliable quality indices, they provide little insight into the processes underlying perception and judgment. In this paper, we analyze the performance of electroencephalography (EEG) for indicating different types of degradations in speech stimuli. We show that a certain EEG technique, event-related-potentials (ERP) analysis, is a useful and valid tool in quality research. Three experiments are reported which show that quality degradations can be monitored in conscious and presumably non-conscious stages of processing. Potential and limitations of the approach are discussed and lines of future research are drawn.**

***Index Terms*—Electroencephalography (EEG), perception, processing, speech quality.**

I. Introduction

In telecommunication research, speech, audio, and audiovisual quality is typically assessed with behavioral tests where subjects provide a rating corresponding to their impression. Neurophysiological data can complement these ratings as a comprehensive and non-intrusive measure, potentially revealing neuronal differences in quality processing below the threshold of conscious perception that might affect a user's long-term satisfaction. In general, only a fraction of all perceptual processing enters consciousness and is as such available to the introspection required for self-report [1]. Still, also the non-conscious processing steps are accompanied by neuronal changes, and thus physiological measures may provide insight into these processes which eventually lead to a given rating. In addition it is possible to record continuously, and obtain data closer to the actual auditory event (with respect to time). In preliminary experiments, we could show that a certain *electroencephalogram (EEG)* technique, *event-related-potentials (ERP)* analysis, is a useful and valid tool in quality research [2] and [3]. In addition, this method was transferred to visual material [4], [5] and [6].

In the present paper, we will give an overview of the theoretical background of this study, which combines quality measurements (Section II) and electroencephalic recordings (Section III). Based on this, we will describe the scope and the results of three experiments which have been carried out to analyze the usefulness of this technique: One addressing signal-correlated noise to provide a baseline for the performance of the approach (Section IV), a second one extending the paradigm to longer stimuli and more practical listening conditions (Section V), and a third one addressing coding distortions as a practically relevant type of degradation (Section VI). The results are summarized and a perspective for future research is given in Section VII.

II. Quality Measurements

Following the considerations of Jekosch (2005) [7], the process from an auditory event to a quality judgment includes a perceptual and an assessment part. The incoming auditory stimulus is perceived and then compared to an internal reference. The outcome of this comparison results in a quality event which happens inside the human listener. In order to get access to this event, the listener has to describe it, for example in a quantitative way on a quality rating scale. The recommended and commonly used procedures for this type of assessment are subjective listening tests which are described, e.g., in ITU Recommendations [8] (for audio quality), [9] (for video quality) and [10] (for speech quality). These methods can be divided into two classes; with and without reference. Methods without reference (e.g., *Absolute Category Rating (ACR)*) result in a *Mean Opinion Score (MOS)* based on judgements of the test stimulus alone.[1] The MOS is a numerical value commonly expressed on a scale from 5 (excellent) to 1 (bad). For the ratings of high quality samples, methods are suitable which have a reference stimulus. During this tests subjects have to rate the quality of the experimental stimulus compared to the quality of a reference sample. The *Comparison Category Rating (CCR)* and the *Degradation Category Rating (DCR)* are examples of such reference-based methods used in speech

Manuscript received November 01, 2011; revised February 18, 2012; accepted February 25, 2012. Date of publication March 23, 2012; date of current version September 12, 2012. This work was supported by the Bernstein Focus: Neurotechnology—Berlin (BFNT-B) by the Federal Ministry of Education and Research (BMBF) grant FKZ 01GQ0850. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Philip Loizou.

J.-N. Antons, R. Schleicher, S. Arndt, and S. Möller are with the Quality and Usability Lab, Berlin Institute of Technology, 10587 Berlin, Germany (e-mail: jan-niklas.antons@telekom.de; robert.schleicher@telekom.de; sebastian.arndt@telekom.de; sebastian.moeller@telekom.de).

A. K. Porbadnigk is with the Machine Learning Laboratory, Berlin Institute of Technology, 10587 Berlin, Germany, and also with the graduate school "Sensory Computation in Neural Systems" (GRK 1589), Bernstein Center for Computational Neuroscience Berlin, 10099 Berlin, Germany (e-mail: anne.k.porbadnigk@tu-berlin.de).

G. Curio is with the Department of Neurology and Clinical Neurophysiology, Charité—University Medicine Berlin, 12200 Berlin, Germany (e-mail: gabriel.curio@charite.de).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTSP.2012.2191936

[1]When presenting several test stimuli in a row, context effects may provoke that the judgment is not only influenced by the test stimulus alone, but also by other stimuli which form part of the test set.

quality assessment [10]; similar paradigms are available for visual and audiovisual stimuli.

III. ELECTROENCEPHALOGRAM

In addition to the well-known and approved approaches for quality measurement, the *electroencephalogram (EEG)* has been proven to be a valid technique for quality research in the auditory and visual domain, which can provide additional information about the underlying processes [2]–[4] and [11]. The EEG is a widely used method for investigating physiological correlates of perceptual processes [12]. It measures voltage differences due to neural activity in the brain by placing electrodes on the scalp's surface. It has an excellent temporal, but a rather limited spatial resolution. In a continuously recorded EEG, one can distinguish between event-independent parts (spontaneous EEG) and differences in voltage in reaction to an external stimulus, so called *event-related-potentials (ERP)* [13]. The latter ones are of special interest to the neuroscientist and will be described in more detail in the following paragraphs. Fabiani *et al.* (2007) considers ERPs to be: "... one of the main tools available to cognitive neuroscientists." [14, p.110]. In the context of EEG-based *brain–computer interfaces (BCIs)* [15], machine learning methods play a crucial role in extracting relevant information from the high-dimensional data [16] and [17]. In recent years, there has been an increasing interest in nonmedical applications of BCI technologies [18] and [19].

A. Mismatch Negativity (MMN)

The mismatch negativity (MMN) is a measure of low-level visual and auditory memory [20]. It is an automatic process caused by differences between the currently processed stimulus and previously received stimuli which generated an internal sensory reference [21]. It is elicited during the range of 100–250 ms after stimulus onset [12]. This automatic process is not conscious and can also be shown in sleeping participants [22]. Näätänen *et al.* first described the MMN and explained it as follows: "The 'traditional' MMN is generated by the brain's automatic response to any change in auditory stimulation exceeding a certain limit roughly corresponding to the behavioral discrimination threshold" [23]. The review from Garrido *et al.* gives a recent overview of the MMN [21].

The components of an ERP are named after their amplitude's polarity ("N" for negative, "P" for positive) and latency in milliseconds, respectively. Especially the N100 and P200 appear to be very meaningful for audiovisual integration; Pilling [24] reasoned that the N1/P2 amplitude reduction due to audiovisual synchrony represents a marker of audiovisual integration. Folstein compares the advantage of cognitive control and mismatch in the N2 component [25].

B. P300

Components such as the P300 and later ones are ascribed to higher cognitive processes. The P300 component, also referred to as P3, is a positive peak approximately 300 ms after stimulus onset. An example for the spatial distribution on the scalp and for the time course can be found in Figs. 1 and 2, respectively. It

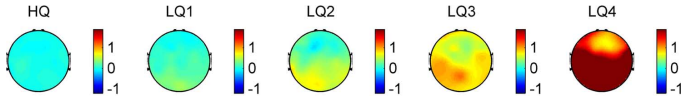

Fig. 1. Scalp topographies for all channels. Each circle depicts a top view of the head, with the nose pointing upwards. Colors code the mean voltage (microvolts) for the time interval from 300–1000 ms after stimulus onset. For LQ1-4, hits were used and for HQ, correctly rejected trials were used.

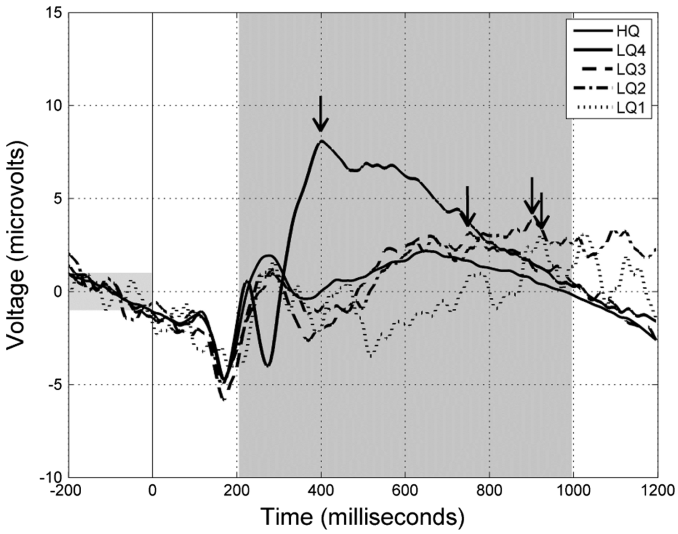

Fig. 2. Grand average ERP plots for HQ and LQ1-4 at channel Cz. For HQ correctly rejected trials (wherein no quality loss was perceived) and for LQ1-4 hits (wherein the quality loss was perceived) were used. Arrows denote P300 peak. Number of trials used for the grand average ERP plot per class: HQ $=$ 22832, LQ4 $=$ 3268, LQ3 $=$ 1332, LQ2 $=$ 610, and LQ1 $=$ 165.

is split into two parts: P3a and P3b. P3a is the result of a comparison between newly perceived information and internal memory copies, similar to the MMN. The P3b component is elicited by task-related attention. In general, the P300 is elicited when a deviant stimulus is presented among a series of more frequent "regular" stimuli, e.g., a high tone among a repeated series of low tones, which is one of the standard tests in ERP research called "oddball paradigm." The review from Polich gives background information on all relevant processes behind the P300, P3a, and P3b components [26]. Even later components, such as the N400, are associated with semantic processing of stimuli, e.g., on a sentence level. The guideline from Duncan *et al.* gives practical advice for the procedure of measuring MMN, P300, and N400 [12].

Using this methodological approach, recent neurophysiological studies of auditory processing led to a model on auditory processing and the conscious perception of stimulus features [27]. Koelsch, who investigated early components of music processing, gives an overview on the connection of music processing and MMN, which are both early stimulus processing stages but are triggered differently [28]. Furthermore, a first study using classes of degradations, which are of interest for research in telecommunication industry, was conducted by Miettinen *et al.* in context of magnetoencephalography (MEG), where they could show a significant increase in the measured amplitudes for distorted stimuli [29].

C. Experimental Preview

This paper is based on the results of three experiments. Experiment I was conducted as an exemplary application of a clinical method to a stimulus set which is relevant in quality research. Experiment II gives an insight on how comparable results for stimuli with varying length are. In addition, we tested the influence of different headphones on quality judgment. Experiment III was performed with even more realistic stimuli in terms of length and class of degradation.

IV. Experiment I

A. Introduction

In Experiment I, standard and deviant stimuli in terms of the oddball paradigm (see Section III-A) were a standard /a/, uttered by a male speaker (high quality, HQ), versus a disturbed version of that phoneme. As distortion, signal-correlated noise generated by a Modulated Noise Reference Unit (MNRU) [30], was chosen. The extent of the distortion was varied in four levels, i.e., from LQ1 to LQ4, where LQ1 (low quality 1, LQ1) referred to the weakest distortion. We hypothesized that the P300 would vary as a consequence of distortion intensity. In addition to the distorted /a/, a second deviant (/i/) was presented as a control stimulus or "sanity check." This stimulus should cause a P300 under any circumstances.

B. Methods

1) Participants: Ten right-handed students and university staff of Technical University of Berlin participated in Experiment I (six female, four male; mean age $= 28.20$ years; SD $= 8.49$; range $= 19$–51), all of them native German speakers. The data from one additional subject was excluded because the experimental task was not accomplished as instructed. All participants reported normal auditory acuity and no medical problems. Handedness was assessed using an inventory from Oldfield [31]. Participants gave informed consent and received monetary compensation. The experiments were conducted in accordance with ethical principles that have their origin in the Declaration of Helsinki.

2) Material: Fourteen vowel phonemes were used: /a/ undisturbed, /i/ undisturbed, and 12 disturbed versions of /a/ impaired with signal-correlated noise. None of these phonemes have lexical meaning in German. To account for possible individual differences in hearing sensitivity, a set of stimuli was selected for each subject individually, based on her/his detection rate. Out of the overall stimulus set, an individual set of four stimuli was selected for each participant, based on the results of an individual pretest. We aimed for detection rates of 100%, 75%, 25%, and 0% for the four selected stimulus levels. The *signal-to-signal-correlated noise ratio (SNR)* for the complete stimulus set were set to: 14, 16, 18, 20, 21, 22, 23, 24, 25, 26, 28, and 30 dB. Stimulus material was digitally recorded in a sound-attenuated experimental chamber with a 48-kHz sampling rate. The phonemes were articulated numerous times by a male speaker. To keep the acoustic variability minimal, we selected only one version of each phoneme. Intensities were normalized using the root mean square of the speech period of the sound file with the software Adobe Audition®. The duration of each stimulus was set to 200 ms. The stimuli were degraded by a MNRU according to ITU-T Rec. P.810 in a controlled and scalable way [30]. The median SNR for the deviant stimuli and for all subjects can be found in Table I.

TABLE I
SNR (dB) for All Subjects and Overall Median SNR

	SNR(dB)				
Subject	HQ	LQ1	LQ2	LQ3	LQ4
1	100	28	24	21	5
2	100	25	20	17	5
3	100	28	24	22	5
4	100	24	22	21	5
5	100	30	26	22	5
6	100	35	28	26	5
7	100	30	26	22	5
8	100	22	20	18	5
9	100	28	25	22	5
10	100	30	25	20	5
median	**100**	**28**	**24**	**21**	**5**

3) Experimental Design and Procedure: In this experimental condition, oddball stimulus sequences of 300 trials in total were presented. In each sequence, the undisturbed phoneme /a/ served as the standard stimulus (70% of the trials), whereas the undisturbed phoneme /i/ and four selected disturbed versions of the phoneme /a/ served as deviants (6% of the trials each), delivered in a pseudo-randomized order, forcing at least one standard to be presented between successive deviants. Since the oddball paradigm has not been used for studies in quality telecommunication research so far, we initially included a control stimulus (/i/) as a sanity check using a well-established P300 event [26]. An exploration of the P300 evoked by the /i/ stimulus showed indeed that a regular novelty P300 was evoked in any subject. As we eventually found a P300 for at least one degradation condition for every subject, a further evaluation of the control stimulus (i.e., /i/) was not conducted. Per subject, eight to twelve sequences were recorded, resulting in a total of 107 sequences. Three additional sequences were run as a pretest: The 12 versions of disturbed phoneme /a/ were presented in separate sequences (four each) with pseudo-randomized order. All of these sequences contained six trials per degradation strength (SNR), respectively. Based on the behavioral results of each subject during the pretest, an individual set of four stimuli was chosen for the experiment. As mentioned, the four selected degradation levels that were selected should be detected with a rate of LQ4 $= 100\%$, LQ3 $= 75\%$, LQ2 $= 25\%$, and LQ1 $= 0\%$. Stimulus sequences were presented with an inter-stimulus-interval varying from 1000 to 1500 ms. Participants were seated comfortably and were instructed to press a button, whenever they detected one of the deviants or the control stimulus (identification task, LQ1-4, and /i/). Stimuli were presented binaurally at the individual preferred listening level through Sennheiser® in-ear headphones. After the physiological measurement, subjects had to rate all 12 stimuli on a CCR scale. An experimental session lasted approximately 3 hours (plus additional time for electrode application and removal), including breaks to avoid participants' fatigue.

4) Electrophysiological Recordings: The EEG (Ag/AgCl electrodes, Brain Products GmbH, Garching, Germany) was

recorded continuously from 64 standard scalp locations according to the extended 10-20 system (AF3-4, 7-8; FAF1-2; Fz, 3-10; Fp1-2; FFC1-2, 5-8; FT7-10; FCz, 1-6; CFC5-8; Cz, 3-6; CCP7-8; CP1-2, 5-6; T7-8; TP7-10; P3-4, Pz, 7-8; POz; O1-2 and the right mastoid) [32]. The reference electrode was placed on the tip of the nose. Electroocular activity was recorded with two bipolar electrode pairs. Impedances were kept below 10 kΩ. The signal was digitized with a 16-bit resolution and a sampling rate of 1000 Hz.

C. Data Analysis

1) Behavioral Data: As behavioral data during the EEG measurement two parameters were extracted. First, the reaction time for the different stimuli, and second, the psychometric functions. The reaction time for each stimulus class is measured in milliseconds, as the duration between the onset of stimulus presentation and the reaction of the subject (received button click). The psychometric function is the result of the detection rate as a function of SNR. A logistic function was fitted to the detection rates of the stimulus levels with the MATLAB® toolbox *psignifit*, approximating the data points in a least-squares sense [33]. After the EEG measurement, subjects had to complete a CCR opinion test, namely rate all LQ level on the scale from excellent to bad. The slider of the Audio Research Lab software STEP set by the participants according to the scale from excellent to bad was converted to a value from 100 to 0.

2) ERP Data: Offline signal processing was carried out using the MATLAB® toolbox EEGLAB [34]. The raw EEG data were low-pass filtered with a finite impulse response filter (low-pass filter with a critical frequency of 40 Hz). EEG epochs with a length of 1400 ms, time locked to the onset of the stimuli, including a 200 ms pre stimulus baseline, were extracted and averaged separately for each condition (HQ, LQ1-4, and C) and for each participant. Epochs (-200 ms to 1200 ms around stimulus onset) showing an amplitude change exceeding 100 μV at any of the recording channels were rejected as artifacts, as this voltage change is unlikely to be produced by neuronal activity. Grand averages were subsequently computed from the individual participant averages. To quantify the deviance-related effects of P300, we measured mean amplitudes, peak latency, peak amplitude and the area below the curve in a fixed time window relative to the pre-stimulus baseline. The time window for P300 quantification was set from 200 to 1000 ms after stimulus onset. The maximal positive amplitude in this time window was automatically determined and its voltage and latency were extracted for further analysis.

3) Classification: The aim of classification was to identify trials in which the subject was not able to detect a degraded stimulus, while notwithstanding an activation pattern similar to conscious detection was present. The detailed selection of classes can be found in Section IV-D3. The classification was done using the MATLAB® toolbox BCILAB [35]. The comparison of ERP data with classification is usually done by comparing the HQ versus the LQ ERPs. Features were the averaged voltage for the time windows, 200–400 ms, 400–600 ms, 600–800 ms, and 800–1000 ms, for all EEG channels. In case of equal covariance matrices for both classes and Gaussian distributions, *Linear discriminant analysis (LDA)* is the optimal classifier [36]. For ERP signals LDA is most suitable for classifying (for detailed information on single-trial classification of EEG data; see [37]). We used a LDA with automatic regularization of the estimated covariance matrix using shrinkage.

D. Statistical Analysis

1) Behavioral Data: A Milton–Friedman-Test with a post-hoc comparison was calculated for the reaction times [38]. For the CCR an *analysis of variance (ANOVA)* with *degradation intensity* as the independent variable and the *mean opinion score (MOS)* as the dependent variable was calculated [39].

2) ERP Data: Deviance-related effects, namely the presence and amplitude of P300 responses, were analyzed on the basis of data from Cz electrode where P300 is typically maximal. While the present pilot study used 64 electrodes, the long-term goal is to identify a minimal electrode placement providing a reliable response estimate in the majority of subjects. Accordingly, in a pre-analysis we calculated a grand average and identified the one single electrode exhibiting the mean maximal P300 amplitude. As this was found at the vertex, all further analysis were run using the Cz electrode. Fig. 2 shows exemplary ERPs for different stimuli classes. To test for the presence of P300 in the control condition (/i/), we compared the deviant responses and the corresponding standard responses to the undisturbed standard phoneme by means of dependent t-tests. The minimum number of epochs constituting the ERP was set to 25. The peak latency and peak amplitude of the P300 responses were analyzed by means of repeated measures ANOVA with the factor *Stimulus (HQ, LQ1-4, and C)*. Finally, pair-wise post-hoc comparisons between target types were drawn with Sidak adjusted alpha level.

3) Classification: Classification was done subject-wise using bandpass filtered raw data (0.2 to 7 Hz). Each LQ class was divided in two separate classes: hits (true positives) and misses (false negatives). Stimuli that were degraded and not detected by the subjects were labeled as misses. Detected degradations were labeled as hits. Two classifications were done: 1) training of a classifier to distinguish between hits and correctly reported HQ trials and test this classifier on the same events (HQ against hits of each class); and 2) again training of a classifier to distinguish between hits and correctly reported HQ trials, but then tested on misses against correctly reported HQ trials. For the training of the second classification one half of the HQ trials and the hit trials of each LQ class were used. Two separate sets of HQ trials (HQ1 and HQ2) were created for the second classification by selecting even and odd HQ trials and assign them to HQ1 and HQ2, respectively. For the testing the other half of the HQ trials and the missed trials were used. This approach was first introduced in [2]. Analysis one was done for all stimulus levels with a five-fold cross validation. Only if a minimum of 15 trials containing hits (classification 1) or containing 15 hits and 15 misses (classification 2) was available, classification was performed (minimal number of trials needed to train and test a classifier). The classification of hits of each target class against HQ demonstrates that the classification of neural reactions due

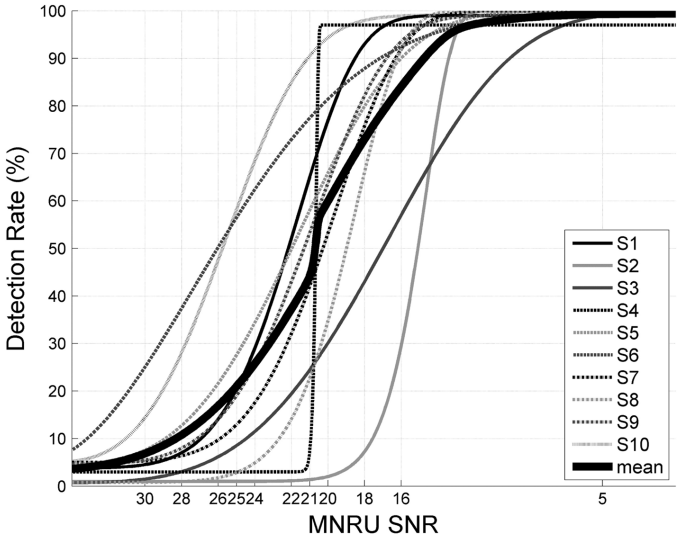

Fig. 3. Psychometric function fits from the psychophysical data; for all subject (Subject(S) 1–10) and the mean.

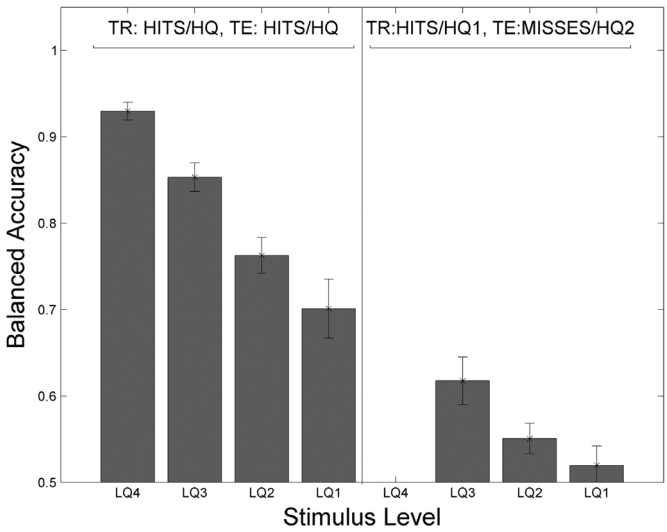

Fig. 4. Classification results. Bars show the average classification performance (balanced accuracy value). Left: Trained (TR) on hits against HQ and tested (TE) on hits against HQ; Right: Trained (TR) on hits against HQ1 and tested (TE) on misses against HQ2; for all stimuli LQ1-4. The bar for LQ4 is missing because no participant had enough misses for testing (classification 2). Number of subjects used for the average of first classification (left): LQ1 $= 4$, LQ2 $= 7$, LQ3 $= 9$, LQ4 $= 10$, and for the second classification (right): LQ1 $= 4$, LQ2 $= 7$, LQ3 $= 8$, LQ4 $= 0$. Whiskers denote standard errors.

TABLE II
MEAN REACTION TIMES FOR ALL LEVELS OF DEGRADATION

	milliseconds			
	LQ1	LQ2	LQ3	LQ4
mean	724.93	736.28	749.24	598.54

to the perception of degraded stimuli is feasible. The second analysis, classifying misses against HQ trials, would reveal differences of the EEG signal due to degradations which were not noticed on the behavioral level. Still, on a physiological level these two classes might differ because the degradation is still processed on a neuronal level. Classification performance was measured by the balanced accuracy, expressed as the *area under the curve (AUC)* of the *receiver operating characteristic (ROC)*, as follows[40]:

$$\text{AUCb} = \frac{\left(\frac{tp}{(tp+fn)} + \frac{tn}{(fp+tn)}\right)}{2}. \quad (1)$$

The balanced accuracy stands for the relationship of true positive (tp) rate and false positive (fp) rate of a 2-class problem including the true negative (tn) rate and the false negative (fn) rate. A value of AUCb $> .9$ reflects excellent classification and AUCb $= .5$ chance level.

E. Results

1) Behavioral Data: For the CCR, the ANOVA with *degradation intensity* as the independent variable and the *mean opinion score (MOS)* as the dependent variable on the CCR data revealed a main effect on the factor *Stimulus (strength of degradation)* $(F(131, 4404) = 306.64, p < .01, \eta^2 = .76)$. The post-hoc test (Sidak adjustment for pairwise comparisons) reached significance for a level of 21 dB $(p < .05)$ compared to the non-degraded stimulus. The mean reaction times for the different conditions can be found in Table II. The reaction time for LQ1-3 are on a similar level, but significantly different compared to LQ4 $(p < .05)$. For stimulus condition LQ4, the reaction time was shorter. The psychometric function fits for all subjects are plotted in Fig. 3.

2) ERP Data: The time window for P300 quantification was set from 200 to 1000 ms after stimulus onset. To test for the presence of P300 in the control condition (/i/), we compared those ERP amplitudes with the ERP amplitudes of the standards by means of two-tailed dependent t-tests. The t-test result was significant ($t = 6.37$, $p < .01$). Fig. 2 shows the grand average ERP, the arrows point the location of the P300 peak for each LQ.

The ANOVA for repeated measurements revealed a main effect for the factor *stimulus* $(F(9, 27) = 3.56, p < .01, \eta^2 = .54)$. For the dependent variable P300 peak amplitude a significant effect was found $(F(3, 9) = 11.34, p < .05, \eta^2 = .79)$, as well as for the dependent variable *P300 peak latency* $(F(3, 9) = 9.35, p < .05, \eta^2 = .75)$. The pairwise comparison (Sidak adjustment for pairwise comparisons) for the peak amplitude revealed a significant difference between LQ2 and LQ4 $(p < .05)$. For the latency a significant effect could be found between LQ2 and LQ3 $(p < .05)$, in addition to a significant effect between LQ2 and LQ4 $(p < .05)$. Fig. 1 shows the scalp distribution of voltage for the different stimulus conditions (hits and correct rejections for LQ1-4 and HQ, respectively). For LQ4, a broad reaction was detected. For the less disturbed stimuli, a reaction was provoked, but not as strong as for LQ4.

In addition, we found a correlation between the P300 amplitude for electrode Cz and the detection rate $(r = .42, p < .05)$. Within the ERP data, a negative correlation between the P300 amplitude and the P300 latency at electrode Pz could be observed $(r = -.33, p < .10)$.

3) Classification: The classification results can be found in Fig. 4. For the first classification level, trained on hits versus HQ and tested on hits versus HQ, the average AUCb value reached a high level for LQ4: AUCb $= .92$, LQ3: AUCb $= .85$, LQ2: AUCb $= .76$, and LQ1: AUCb $= .70$. The second classification level reached the following values; for LQ4: not enough misses,

LQ3: AUCb = .61, LQ2: AUCb = .55, and LQ1: AUCb = .51. It needs to be noted, that classification could not be performed for all subjects (due to a small number of hit/miss trials) such that the average values reported here are averages calculated over subsets of subjects (classification 1: LQ1 = 4, LQ2 = 7, LQ3 = 9, LQ4 = 10; classification 2: LQ1 = 4, LQ2 = 7, LQ3 = 8, LQ4 = 0).

F. Discussion

The analysis of the subjective CCR ratings revealed that the quality was rated as significantly lower from an SNR of 21 dB on. This point denotes the threshold, from where on the quality was perceived as significantly worse in comparison to the reference. The reaction time for the strongest degradation was shorter compared to the weaker ones, meaning that subjects were faster in detecting the degradation and providing the corresponding rating. The psychometric functions showed that the mean detection rate passes 50% at a 21-dB degradation level. This result is similar to the result of the CCR. For the ERP data, the significant P300 generated by the control stimulus (i.e., /i/) showed that the experimental setup is appropriate for our purposes. It turned out that in this stimulus design and given the present signal-to-noise ratio, no robust MMN could be identified. It might appear surprising that a residual P3 response, which is known to represent cognitive stimulus appraisal, was detected in trials for which no behavioral detection was reported. In the context of the present paradigm one could argue that minor physical stimulus differences were initially detected, yet an internal response criterion has not been met, such that an overt behavioral report was not initiated. The discovered effects of P300 peak latency showed that the harder it is to detect a degraded stimulus the later a P300 was evoked. This can be due to the fact that more cognitive effort is involved in detecting the degradation. The significant variation of the P300 peak amplitude is comparable to the variation of latency, but shows the opposite pattern of change: the stronger the degradation, the higher the P300 amplitude. This result was supported by the two correlations.

The P300 amplitude is varying with the detection rate: The higher the amplitude, the higher the detection rate. Comparing the amplitude with the latency of the P300, a negative correlation suggests that the smaller the amplitude the longer the latency.

Interestingly, the analysis of the grand mean data obtained as average over all subjects showed the strongest P3 response at Cz. Thus, in the present paradigm the most effective placement of a single electrode was in between the commonly reported places for novelty (P3a) and target (P3b) ERP which have been described at more frontal or more parietal sites [26]. With the first level of classification we could show that the brain reaction due to the processing of a degradation, in our case the difference between the undisturbed and disturbed stimulus, can be well detected. With our second classification we could show that the pattern of brain activation related to processing degradations consciously can also be detected in trials which are not reported as degraded on a subjective level. We conclude that these trials might have been processed non-consciously and had no measurable influence on the direct user rating. This processing might still lead to an influenced long-term quality judgement, due to increased cognitive load and fatigue when being exposed to small degradations for a long period of time (for measuring fatigue using EEG see [42]).

V. Experiment II

A. Introduction

The aim of Experiment II was to determine if 1) the length of a stimulus has an influence on the subjective rating and 2) the type of headphones has an influence on the subjective rating of one degradation class. The motivation for this test was based on the fact that the common length of stimuli used for tests in telecommunication research is around 8 seconds [10], which is much longer than the common length in ERP research (between 100 to 1500 ms). In addition, common speech quality tests are carried out with circumaural headphones rather than in-ear plugs, the latter being typical for EEG setups. The result should reveal if the headphone type was an unwanted influence on degradation perception.

B. Methods

1) Participants: Twenty volunteers (ten female, ten male; mean age = 24.32 years; SD = 3.54; range = 22–28; all right-handed), all native German speakers took part in Experiment II. All participants reported normal auditory acuity. They gave informed consent and received monetary compensation.

2) Material: For Experiment II we used stimuli of three different lengths: phoneme, word and sentence. The phoneme from Experiment I was used: /a/ (200 ms). The stimulus with the length of a word was the German translation of eye brow /Augenbraue/ (1200 ms) and a test sentence from the EUROM data base uttered by a male speaker were used (8000 ms) [41]. The two tested headphones were Sennheiser in-ear plugs and AKG over-ear headphones. Stimuli were degraded with signal-correlated noise at the following SNRs: 5, 10, 14, 16, 18, 20, 21–35 dB in one dB steps.

C. Experimental Design and Procedure

As in Experiment I subjects had to rate all stimuli on a CCR scale from excellent (100) to bad (0). The stimuli types (three different lengths) were judged on all levels of degradation and with both headphones (in-ear versus over-ear).

D. Statistical Analysis

The data was analyzed performing an ANOVA with *type of headphone* and *length of stimulus* as the independent variables and the *mean opinion score (MOS)* as the dependent variable.

E. Results

We found a main effect for the factor length of stimulus ($F(21, 4404) = 598.19, p < .01, \eta^2 = .15$). There was no main effect for the type of headphone ($F(1, 4404) = .58$, not significant). The post-hoc analysis (Sidak adjustment for pairwise comparisons) revealed a significant difference between the stimuli with the length of phonemes and words ($p < .01$), as well for the difference between phonemes and sentences

($p < .01$). Not significant was the difference between stimuli with the length of words and sentences.

F. Discussion

As expected from Experiment I, the level of degradation had an influence on the subjective judgement. The factor length of stimulus had a significant effect on the user rating. A significantly higher quality was assigned to stimuli with the length of phonemes compared to stimuli with the length of words and sentences. There was no difference between the judgement of word and sentence-long stimuli. This leads to the conclusion that stimuli for EEG experiments on quality should be at minimum of word length. As there was no difference between the rating of the two types of headphones the influence of the headphone type can be neglected. Experiment I had used short stimuli (vowels) which are a standard in ERP studies and thus allowed to use the established physiological knowledge to interpret the new findings and their implications for cerebral processing of stimulus quality. However, in quality research longer stimuli are employed for behavioral detection of stimulus degradation. Therefore, Experiment II directly compared the behavioral effects of stimuli differing in length (vowels, words, sentences). Indeed, longer stimuli (i.e., words or sentences) permit a better detection of minor stimulus degradation. According to this behavioral result, we added the ERP Experiment III which now used word stimuli thereby linking the present ERP results directly to the behavioral standards in quality research.

VI. Experiment III

The goal of Experiment III was to test even more realistic stimuli in terms of the length, and to extend our paradigm to another class of degradation. The stimuli were words [/Haus/ (English: house) and /Schild/ (English: sign)] and differed each with regard to the bandwidth of the codec ITU-T Recommendation G.722.2 [43]. The difference between the high quality (HQ; wide-band) and lower qualities (LQ1-4; subset chosen from the conditions: 6.6, 8.85, 12.65, 14.25, 15.85, 18.25, 19.85, 23.05 kbit/s) was expected to elicit an early difference pattern for conditions LQ2-4 and a P300 for at least the highest degradation.

A. Methods

1) Participants: Nine volunteers (four female, five male; mean age $= 25.22$ years; SD $= 1.20$; range $= 24$–27; all right-handed), all native German speakers took part in Experiment III. None of them had participated in Experiment I or II. All participants reported normal auditory acuity. They gave informed consent and received monetary compensation.

B. Material

We used different stimulus material as in Experiment I. Two words were chosen, each spoken by a female and male speaker. For all four LQ conditions we used the codec G.722.2 as degradation. The best quality was used as high quality (HQ: direct wide-band quality without any coding-decoding process) as standard stimulus and all lower bandwidth conditions as deviants (6.6, 8.85, 12.65, 14.25, 15.85, 18.25, 19.85, and 23.05 kbit/s). All combinations result in a stimulus set of 8 per word and speaker. As in Experiment I a subset of four stimuli were determined individually for each subject. The aimed detection rates were LQ4 $= 100\%$, LQ3 $= 60\%$, LQ2 $= 40\%$, and LQ1 $= 0\%$. The selected stimulus levels can be found in Table III.

TABLE III
Bandwidth (kbit/s) for All Subjects and Overall Median

Subject	kbit/sec				
	HQ	LQ1	LQ2	LQ3	LQ4
1	WB	14.25	12.65	8.85	6.6
2	WB	14.25	12.65	8.85	6.6
3	WB	14.25	12.65	8.85	6.6
4	WB	15.85	14.25	12.65	8.85
5	WB	23.05	19.85	18.25	15.85
6	WB	15.85	14.25	12.65	8.85
7	WB	15.85	14.25	12.65	8.85
8	WB	14.25	12.65	8.85	6.6
9	WB	14.25	12.65	8.85	6.6
median	**WB**	**14.25**	**12.65**	**12.65**	**6.6**

C. Experimental Design and Procedure

In a forced choice task, subjects had to rate whether a given word was of high quality (HQ) or degraded (LQ). Stimuli were presented either in wide-band quality or were impaired. Besides the aforementioned modifications, we used the same experimental setting as in Experiment I.

D. Electrophysiological Recordings

Settings for the electrophysiological recordings were the same as in Experiment I. Scalp locations: Fp1-2; AF3-4; Fz, 1-6,9-10; FCz, 1-8; T7-8; Cz, 1-6; TP7-8; CPz, 1-6; Pz, 1-10; Poz, 3-4, 7-8; Oz, 1-2; AF7-8 and the right mastoid were recorded.

E. Data Analysis

The data were analyzed in the same way as in Experiment I except the following changes.

1) ERP Data: The time window for P300 quantification was set from 400 to 900 ms after stimulus onset.

2) Classification: As time windows, we used: 400–500 ms, 500–600 ms, 600–700 ms, and 700–900 ms.

F. Statistical Analysis

Statistical analyses were performed in the same way as described for Experiment I.

G. Results

1) Behavioral Data: The ANOVA calculated on the subjective data, with *degradation intensity* as the independent variable and the *mean opinion score (MOS)* as the dependent variable, revealed a main effect on the factor *Stimulus (strength of degradation)* ($F(120, 4262) = 550.86, p < .05, \eta^2 = .64$). The post-hoc test (Sidak adjustment for pairwise comparisons) reached significance for a level of 8.85 kbit/s ($p < .05$). The mean reaction times for the different conditions can be found in Table IV. The reaction time for LQ1-3 are on a similar level, but significantly different compared to LQ4 ($p < .05$) and HQ ($p < .05$). For the stimulus condition LQ4 and HQ reaction time was shorter. The psychometric function fits for all subjects are plotted in Fig. 5.

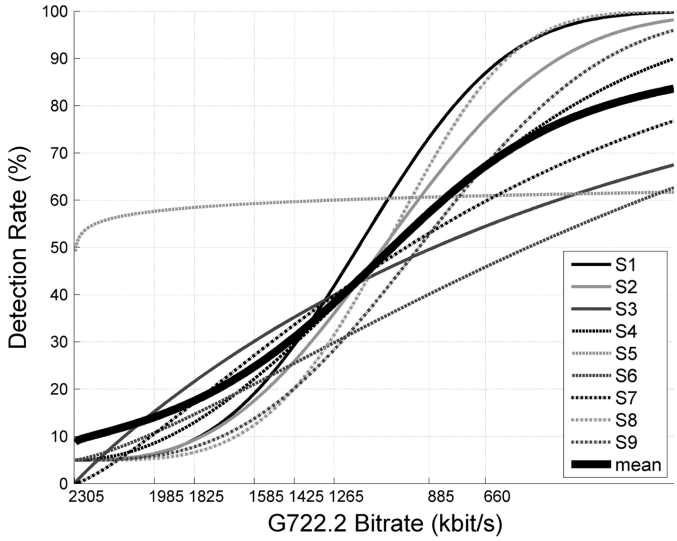

Fig. 5. Psychometric function fits from the psychophysical data; for all subject (Subject(S) 1–10) and the mean.

TABLE IV
MEAN REACTION TIMES FOR ALL CONDITIONS

	milliseconds				
	HQ	LQ1	LQ2	LQ3	LQ4
mean	609.17	724.93	736.28	749.24	598.54

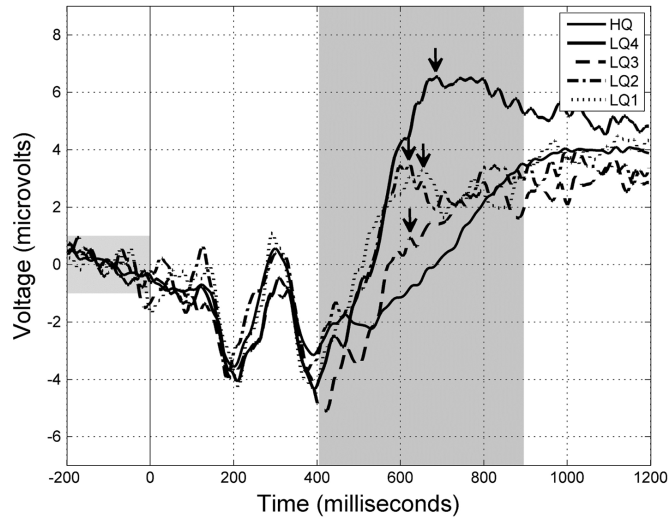

Fig. 6. Grand average ERP plots for HQ and LQ1-4 at channel Cz. For HQ correctly rejected trials (wherein no quality loss was perceived) and for LQ1-4 hits (wherein quality loss was perceived) were used. Arrows denote P300 peak. Number of trials used for the grand average ERP plot per class: HQ $= 11177$, LQ4 $= 1235$, LQ3 $= 826$, LQ2 $= 655$, and LQ1 $= 500$.

2) P300: The ANOVA for repeated measurements revealed a main effect for the factor stimulus ($F(4,3) = 3.36, p < .01, \eta^2 = .35$). The P300 area variation is significant ($F(3,18) = 5.83$, $p < .05$, $\eta^2 = .49$), as well the P300 mean ($F(3,18) = 5.84$, $p < .05$, $\eta^2 = .48$) and the peak amplitude ($F(3,18) = 10.10$, $p < .01$, $\eta^2 = .62$). For the dependent variable P300 peak latency no effect was found. The grand average for all stimulus classes can be found in Fig. 6. The pairwise comparison for the peak amplitude revealed a significant difference between LQ1 and LQ2 ($p < .05$) and LQ1 and LQ4 ($p < .05$). Fig. 7 shows the scalp distribution of

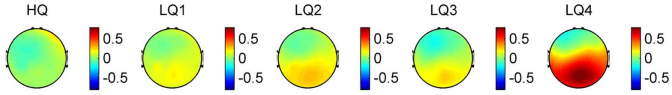

Fig. 7. Scalp topographies for all channels. Each circle depicts a top view of the head, with the nose pointing upwards. Colors code the mean voltage (microvolts) for the time interval from 500–1000 ms after stimulus onset. For LQ1-4, hits were used and for HQ, correctly rejected trials were used.

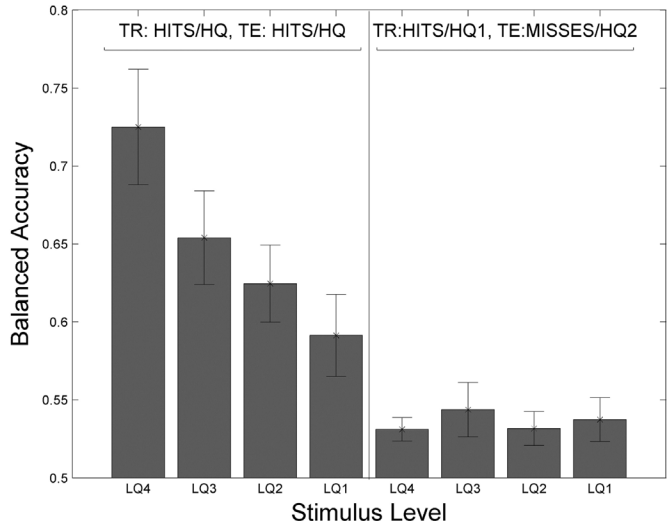

Fig. 8. Classification results. Bars show the average classification performance (balanced accuracy value). Left: Trained (TR) on hits against HQ and tested (TE) on hits against HQ; Right: Trained (TR) on hits against HQ1 and tested (TE) on misses against HQ2; for all stimuli LQ1-4. Number of subjects used for the average of first classification (left): LQ1 $= 9$, LQ2 $= 9$, LQ3 $= 9$, LQ4 $= 9$ and for the second classification (right): LQ1 $= 9$, LQ2 $= 9$, LQ3 $= 9$, LQ4 $= 7$. Whiskers denote standard errors.

voltage for the different stimulus conditions (hits and correct rejections for LQ1-4 and HQ, respectively). For LQ4 a broad reaction was detected. For the less disturbed stimuli a reaction was evoked but not as strong as for LQ4.

3) Classification: The results for the classification can be found in Fig. 8. For the first classification level, trained on hits versus HQ and tested on hits versus HQ, the average AUCb value reached a high level (for LQ1: AUCb $= .72$, LQ2: AUCb $= .66$, LQ3: AUCb $= .62$, and LQ4: AUCb $= .59$). The second classification level reached the following values; for LQ1: AUCb $= .53$, LQ2: AUCb $= .54$, LQ3: AUCb $= .53$, and LQ4: AUCb $= .53$. As for Experiment I, it needs to be considered that the average values reported here are averages calculated over subsets of subjects (classification 2: LQ4: 7 subjects; otherwise: all subjects).

H. Discussion

The analysis of the subjective CCR ratings revealed that the quality was rated as significant lower from a SNR of 8.85 kbit/s on.

Surprisingly, one of our subjects was more sensitive concerning the detection of the degradation (Fig. 5). Even after a detailed inspection of the data no irregularity of the ERP data was encountered and thus the data were included in the analysis. The reaction time for the strongest degradation and for the HQ stimulus was lower compared to the weaker degraded ones, meaning

that subjects were faster to detect the degradation and giving the corresponding rating. The psychometric functions showed that the mean detection rate passes 50% at 8.85 kbit/s degradation level. Compared with the psychometric functions of Experiment I the curves of Experiment III are smoother, meaning the detection rate is rising slower with the intensity of the degradation. This is due to the fact that participants can clearly identify the noise in Experiment I as a degradation, whereas the compression artifacts in this experiment, were harder to detect for some participants. The P300 area showed that the harder it is to detect a degraded stimulus, the smaller the P300 area. The significant variation of the P300 mean amplitude is comparable to the variation in area: the stronger the degradation, the higher the P300 mean amplitude.

With the first level of classification we could show that the brain reaction due to the processing of a degradation in our case the difference between the undisturbed and disturbed stimulus can be well detected. With our second classification we could show that the pattern of brain activation related to processing degradations consciously can also be detected in trials which are not reported as degraded on a subjective level. We conclude again that these trials might have been processed non-consciously and had no measurable influence on the direct user rating. Balanced accuracy for classification 2) show a huge similarity across quality levels, in contrast to what could be expected. This might be, because one stimulus had a surprisingly high number of hits for low quality levels for a considerable number of subjects (LQ2 and 1).

VII. General Discussion and Future Work

In this paper, we investigated the usefulness of event-related potentials (ERPs) for analyzing human speech quality perception. Our aim was to investigate whether ERPs specific to the detection of degradations can be identified, potentially also for non-conscious processing steps when listening to degraded speech files, and to test the applicability of the method in a realistic application scenario. Three experiments were carried out to perform the analysis. At present there is limited insight in the neuronal processes underlying quality appraisal. EEG can serve as a tool to learn about these processes, and may offer the benefit of potentially revealing non-conscious parts of the quality judgment process as well.

In the first experiment, we could show that the subjects' detection rate in the oddball paradigm experiment reached the 50% threshold at the same SNR level at which test participants also rated the quality significantly worse in the CCR test. The reaction time in the oddball paradigm was significantly higher for the high SNR condition, reflecting the cognitive effort required to process the (subtle) degradation. The P300 peak latency showed that the lower the degradation level, the later a P300 was evoked, which is most likely due to the same reason, namely the higher cognitive effort involved in detecting the degradation. In turn, the stronger the degradation, the higher the P300 peak amplitude. Using LDA classifiers on the EEG signals, we could show that patterns of brain activation which were similar to the ones for detected degradations could also be observed for trials where the participants did not report a degradation. It is likely that small degradations are non-consciously processed in a similar way as larger ones, although they do not result in the same conscious user rating.

The brain accomplishes the given experimental task mainly as expected. Strongly degraded stimuli were assessed faster compared to the weaker degradations. The stronger the degradation, the earlier is the maximum amplitude of P300 and the higher its amplitude. The expected scalp potential distribution and peak latency matches neatly with the literature [26]. One unusual feature is the detection of degradation-specific brain responses to weak degradations which were not reported at the behavioral level. Thus, ERP analysis might provide objective evidence for non-conscious engagement of brain processes by minor stimulus degradations which eventually could influence the users' appreciation of stimulus quality during long-term confrontation with this degraded material.

The applicability of the method was further analyzed in the second experiment. The results showed that stimulus length had a significant impact on the subjective responses, with stimuli of word or sentence length being rated significantly worse than phoneme stimuli. That implies that stimuli of at least word length should be used in subsequent experiments in order to reflect realistic usage scenarios. The type of headphone did not have a significant impact. Therefore in-ear headphones which are more common in EEG studies can be used.

For the third experiment, similar results could be found for coding distortions as for signal-correlated noise. It should be noted, however, that the subtlety of the degradation did not affect the latency of the P3 for coding distortions in words. This leads to the question to which stimulus modalities and types of degradations the ERP-based analysis can be applied to. Considering the initial investigations reported in [2] and [3], it is expected that the method is applicable to auditory, visual and audiovisual stimuli alike. With respect to the latter, we expect that further insight on the quality integration process of auditory and visual perception may be obtained by our method. For example, one could determine how the impact on MMN and P300 components adds up when a stimulus contains both auditory and visual degradations. We are interested in how the brain activity patterns sum up in such cases, and how this affects the classification rate.

A further point of future work will be to analyze the impact that the presumably non-conscious detection of degradations may have on perceived quality aspects. A missing response behaviorally does not necessarily mean that small degradations do not have an influence on quality perception neurally. For example, we assume that cognitive load and fatigue might increase when being exposed to small degradations for a long period of time, e.g., when viewing films with subtle degradations. In order to analyze this impact, EEG-based measurement techniques need to be combined with other physiological methods such as *Heart rate variability (HRV)* and non-physiological indices (self-reported load indices). On the long run, we expect that ERP-based analysis will be one of several methods providing insight into the quality perception and judgment processes which are still not well understood. In the future, a new compression or transmission method might be declared "subjectively lossless" only if the subject's brain activity shows no difference compared to the activity during perception of the original signal.

ACKNOWLEDGMENT

The authors would like to thank K.-R. Müller, B. Blankertz, I. Sturm, M. Treder, and S. Scholler from the Machine Learning group of the Berlin Institute of Technology for providing EEG hardware, expertise in data analysis (specifically of EEG data), and helpful discussions. They would also like to thank I. Sturm and A. Abbas for assisting us with the data collection. For helpful comments on the manuscript, they would also like to thank B. Belmudez and B. Blankertz. They would like to thank B. Blankertz for design and implementation of Exp. 1, for contributing the initial EEG analysis, and guiding the data analysis process developed by A. K. Porbadnigk, in particular classification.

REFERENCES

[1] A. R. Damasio, *Descartes' Error: Emotion, Reason, and the Human Brain*. New York: G.P. Putnam, 1994.

[2] A. K. Porbadnigk, J.-N. Antons, B. Blankertz, M. S. Treder, R. Schleicher, S. Möller, and G. Curio, "Using ERPs for assessing the (sub)conscious perception of noise," in *Proc. 32nd Annu. Int Conf. IEEE Eng. Med. Biol. Soc.*, 2010, pp. 2690–2693.

[3] A. K. Porbadnigk, J.-N. Antons, M. S. Treder, B. Blankertz, R. Schleicher, S. Möller, and G. Curio, "ERP assessment of word processing under broadcast bit rate limitations," *Neurosci. Lett.*, vol. 500, no. 1, p. e49, 2011.

[4] S. Arndt, J.-N. Antons, R. Schleicher, S. Möller, S. Scholler, and G. Curio, "A physiological approach to determine video quality," in *Proc. IEEE Int. Symp. Multimedia*, 2011, pp. 518–523.

[5] L. Lindemann, S. Wenger, and M. Magnor, "Evaluation of video artifact perception using event-related potentials," in *Proc. ACM Appl. Percept. Comput. Graph. Visualiz. (APGV) 2011*, 2011.

[6] H. Hayashi, H. Shirai, M. Kameda, S. Kunifuji, and M. Miyahara, "Assessment of extra high quality images using both EEG and assessment words on high order sensations," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, 2000, vol. 2, pp. 1289–1294, 2000.

[7] U. Jekosch, *Voice and Speech Quality Perception: Assessment and Evaluation*. Berlin, Germany: Springer, 2005.

[8] "Methodology for the subjective assessment of the quality of television pictures," Int. Telecomm. Union. Geneva, Switzerland, 2002, ITU-T Rec. BT.500-11.

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

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

[11] 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 Trans. Image Process.*, vol. 21, no. 5, pp. 2619–2629, May 2012.

[12] 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 Neurophysiol.*, vol. 120, pp. 1883–1903, 2009.

[13] M. S. Coles and M. Rugg, "Event-related brain potentials: An introduction," in *Electrophysiology of Mind: Event-Related Brain Potentials and Cognition*. New York: Oxford Univ. Press, 1995.

[14] M. Fabiani, G. Gratton, and K. D. Federmeier, "Event-related potentials: Methods, theory, and applications," in *Handbook of Psychophysiology*. Cambridge, U.K.: Cambridge Univ. Press, 2007.

[15] G. Dornhege, J. del R. Millán, T. Hinterberger, D. McFarland, and K. R. Müller, *Toward Brain-Computer Interfacing*. Cambridge, MA: MIT Press, 2007.

[16] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, and K. R. Müller, "Optimizing spatial filters for robust EEG single-trial analysis," *IEEE Signal Process. Mag.*, vol. 25, no. 1, pp. 41–56, Jan. 2008.

[17] K. R. Müller, S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf, "An introduction to kernel based learning algorithms," *IEEE Neural Netw.*, vol. 12, no. 2, pp. 181–201, Mar. 2001.

[18] K. R. Müller, M. Tangermann, G. Dornhege, M. Krauledat, G. Curio, and B. Blankertz, "Machine learning for real-time single-trial EEG-analysis: From brain-computer interfacing to mental state monitoring," *J. Neurosci. Meth.*, vol. 167, no. 1, pp. 82–90, 2008.

[19] B. Blankertz, M. Tangermann, C. Vidaurre, S. Fazil, C. Sannelli, S. Haufe, C. Maeder, L. E. Ramsey, I. Sturm, G. Curio, K. R. Müller, and T. Wiegand, "The Berlin brain-computer interface: Non-medical uses of BCI technology," *Frontiers in Neurosci.*, vol. 4, p. 198, 2010.

[20] R. Näätänen, "Mismatch negativity (MMN) as an index of central auditory system plasticity," *Proc. Int. J. Audiol.*, vol. 47, pp. 16–20, 2008.

[21] M. Garrido, J. Kilner, K. Stephan, and K. Friston, "The mismatch negativity: A review of underlying mechanisms," *Clinical Neurophysiol.*, vol. 120, pp. 453–463, 2009.

[22] L. Sculthorpe, D. Ouellet, and K. Campbell, "MMN elicitation during natural sleep to violations of an auditory pattern," *Brain Res.*, vol. 1390, pp. 52–62, 2009.

[23] R. Näätänen, P. Paavilainen, T. Rinne, and K. Alho, "The mismatch negativity (MMN) in basic research of central auditory processing: A review," *Clinical Neurophysiol.*, vol. 118, pp. 2544–2590, 2007.

[24] M. Pilling, "Auditory event-related potentials (ERPs) in audiovisual speech perception," *J. Speech, Lang., Hear. Res.*, vol. 52, pp. 1073–1081, 2009.

[25] J. R. Folstein and C. Petten, "Influence of cognitive control and mismatch on the N2 component of the ERP: A review," *Psychophysiol.*, vol. 45, no. 1, pp. 152–170, 2008.

[26] J. Polich, "Updating P300: An integrative theory of P3a and P3b," *Clinical Neurophysiol.*, vol. 118, no. 10, pp. 2128–2148, 2007.

[27] R. Näätänen, T. Kujala, and I. Winkler, "Auditory processing that leads to conscious perception: A unique window to central auditory processing opened by the mismatch negativity and related responses," *Psychophysiol.*, vol. 48, pp. 4–22, 2011.

[28] S. Koelsch, "Music-syntactic processing and auditory memory: Similarities and differences between ERAN and MMN," *Psychophysiol.*, vol. 46, pp. 179–190, 2009.

[29] 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 Neurosci.*, vol. 11, no. 24, pp. 1471–2202, 2010.

[30] "Modulated noise reference unit (MNRU)," Int. Telecomm. Union. Geneva, Switzerland, 1996, ITU-T Rec. P.810.

[31] R. C. Oldfield, "The assessment and analysis of handedness: The Edinburgh Inventory," *Neuropsychologia*, vol. 9, pp. 97–113, 1971.

[32] American Clinical Neurophysiology Society, "Guideline 5: Guidelines for standard electrode position nomenclature," *J. Clinical Neurophysiol.*, vol. 23, no. 2, pp. 107–110, 2006.

[33] F. Wichmann and N. Hill, "The psychometric function: I. fitting, sampling, and goodness of fit," *Percept. Psychophys.*, vol. 63, no. 8, p. 1293, 2001.

[34] A. Delorme and S. Makeig, "EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics," *J. Neurosci. Meth.*, vol. 134, pp. 9–21, 2004.

[35] A. Delorme, T. Mullen, C. Kothe, A. Z. Akalin, N. Bigdely-Shamlo, A. Vankov, and S. Makeig, "EEGLAB, SIFT, NFT, BCILAB, and ERICA: New tools for advanced EEG processing," *Comput. Intell. Neurosci*, 2011:130714, 12 pp., doi: 10.115/2011/130714, 2011.

[36] R. Duda, P. Hart, and D. Stork, *Pattern Classification*. New York: Wiley, 2001.

[37] 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.

[38] J. Bortz, G. A. Lienert, and K. Boehnke, *Methoden in der Biostatistik*. Berlin, Germany: Springer, 2008.

[39] J. Bortz, *Statistik: Für Human- und Sozialwissenschaftler*. Berlin, Germany: Springer, 2005.

[40] M. Sokolova, N. Japkowicz, and S. Szpakowicz, "Beyond accuracy, F-score and ROC: A falmily of discriminant measures for performance evaluation," in *Proc. 21st National Conf. Artif. Intell.*, 2006, pp. 781–787.

[41] D. Gibbon, EUROM.1 German Speech Database, Univ. Bielefeld, Bielefeld, Germany, ESPRIT Project 2589 Report, 1992, (SAM, Multi-Lingual Speech Input/Output Assessment, Methodology, and Standardization).

[42] R. Schleicher, N. Galley, S. Briest, and L. Galley, "Blinks and saccades as indicators of fatigue in sleepiness warnings: Looking tired?," *Ergonomics*, vol. 51, no. 7, pp. 982–1010, 2006.

[43] "Wideband coding of speech at around 16 kbit/s using adaptive multi-rate wideband (AMR-WB)," Int. Telecomm. Union. Geneva, Switzerland, 2002, ITU-T Recommendation G.722.2.

Jan-Niklas Antons received the German Diplom in psychology from the Darmstadt Institute of Technology, Darmstadt, Germany, in 2008.

He is with the Quality and Usability Laboratories, Telekom Innovation Lab, Berlin Institute of Technology, Berlin, Germany. At Telekom Innovation Laboratories, he works as a Research Associate in the domain of neurotechnologies for man–machine-interaction. Further, he worked on physiological models for quality estimation.

Robert Schleicher received the diploma degree in psychology from the University of Bonn, Bonn, Germany, in 2003 and the Ph.D. degree from the University of Cologne, Cologne, Germany, in 2008 with a Ph.D. thesis on psychophysiology, eye movements, and emotions.

After the diploma degree, he worked at the Department of Clinical Psychology, University of Cologne, in the workgroup biological psychology. He joined the Telekom Innovation Laboratories, Berlin Institute of Technology, Berlin, Germany, as a Senior Research Scientist in November 2006. His research interests are applied psychophysiology with an emphasis on emotions and eye movements in human–machine interaction.

Sebastian Arndt received the diploma degree in computer science from the Berlin Institute of Technology, Berlin, Germany, in 2010.

He is a Research Assistant in the Quality and Usability Lab, Telekom Innovation Laboratories, Berlin Institute of Technology. During his studies, Sebastian worked as a Student Assistant in the area of mobile and physical interaction at Telekom Innovation Laboratories. Also, from 2008 to 2009, he was studying at the University of Oklahoma, Norman.

Sebastian Möller (M'05–SM'11) studied electrical engineering at the University of Bochum, Bochum, Germany, University of Orleans, Orleans, France, and University of Bologna, Bologna, Italy, and received the Dr. Eng. degree from Ruhr-Universität Bochum, Bochum, Germany, in 1999, and the Venia Legendi with a book on the quality of telephone-based spoken dialogue systems in 2004.

In 2005, he joined Telekom Innovation Laboratories, Berlin Institute of Technology, Berlin, Germany, and in 2007, he was appointed Professor for Quality and Usability at the Berlin Institute of Technology. His primary interests are in speech signal processing, speech technology, and quality and usability evaluation. Since 1997, he has taken part in ITU-T Study Group 12, where he is currently Co-Rapporteur of Question Q.8/12.

Anne K. Porbadnigk received the M.S. degree in computer science from the University of Michigan, Ann Arbor, in 2006, where she studied on a Fulbright scholarship, and the German Diplom in computer science from the Karlsruhe Institute of Technology, Karlsruhe, Germany, in 2009.

She spent the academic year 2008/2009 as a Visiting Researcher at Carnegie Mellon University, Pittsburgh, PA. She joined the Machine Learning Laboratory, Berlin Institute of Technology, Berlin, Germany, in 2009, working as a Research Assistant in the Berlin Brain Computer Interface (BBCI) group. She is also affiliated with the Bernstein Center for Computational Neuroscience Berlin.

Gabriel Curio received the Dr. med. degree from Freie Universität (FU), Berlin, Germany, with a thesis about attentional influences on smooth pursuit eye movements.

Since 1991, he has been leading the Neurophysics Group at the FU Berlin. Currently, he is a Professor of Neurology and Deputy Director at the Department of Neurology and Clinical Neurophysiology, Campus Benjamin Franklin. He is Founding Co-Director of the Bernstein Center for Computational Neuroscience Berlin and the Berlin NeuroImaging Center, Founding Member of the Bernstein Focus Neurotechnology Berlin, and Faculty member of the Berlin Excellence School Mind and Brain. His current research interests include integrating the neurophysics of noninvasive electromagnetic brain monitoring with both basic and clinical neuroscience concepts.