

Chapter 2

ERPs and Quality Ratings Evoked by Phoneme Stimuli Under Varying SNR Conditions

In this chapter, the first and second contributions (see Sect. 1.4) of this book will be introduced. First, a test set-up combining neurophysiological and subjective quality assessment methods for speech quality perception testing will be presented. Secondly, the functionality of this set-up will be validated for short speech stimuli with the length of phonemes and a generic quality impairment, i.e., signal-correlated noise (for an overview see Fig. 2.1). It will be shown that a test set-up combining neurophysiological and subjective quality assessment methods for speech quality perception testing is suitable for measuring the perceived speech quality, and in some instances (trials) it is advanced in comparison to a standard subjective test set-up.

In this first experiment, the selection of short auditory stimuli is based on the fact that—in the majority of neurophysiological speech/audio research—stimuli are of short duration (up to approx. 2 s). In order to be able to follow standard neurophysiological recommendations (see Sect. 1.3.4) for ERP experiments, the duration of stimuli was restricted to the length of phonemes. The degradation class was selected on the basis that, in the first validation of the test set-up, a continuous degradation should be used. Signal-correlated noise, as used for comparisons of different transmission systems [90], was existent during the entire speech signal, but only when the speech signal was active (no noise in the pauses).

The test set-up and its initial validation will be described below. In Sect. 2.2, the experimental set-up including EEG, ERP measurement, and opinion tests will be explained, followed by a global analysis (Sect. 2.3), an introduction to the statistical tools (Sect. 2.4), and a presentation of the experimental results (Sect. 2.5). In the following chapters the test set-up will be extended to include stimuli with longer duration and other quality reductions (see Chap. 3 for stimuli with word length and bit rate reduction; Chap. 4 for stimuli with sentence length and reverberation).

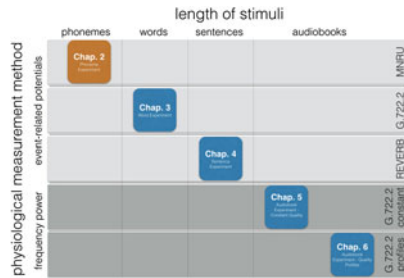


Fig. 2.1 Conducted experiments and structure of this book. Different lengths of stimuli on the x-axis (phonemes, words, sentences, and audiobooks). Physiological measurement techniques on the y-axis (EEG frequency band power and event-related potentials). Applied classes of degradations are color-coded (*grey bars*) and indicated on the *right* (signal-correlated noise introduced by a modulated noise regulation unit (MNRU), bit rate reductions introduced by using different settings of a speech codec in accordance with ITU-T Rec. G.722.2, and reverberation (REVERB) introduced by different room impulse responses. Current chapter is indicated in *orange*. Chapter 2, *Phonemes Experiment*

2.1 Introduction

In the *Phoneme Experiment*, *standard* and *deviant* stimuli are presented in terms of the *oddball paradigm* (see Sects. 1.3.4.4 and 2.2.3), in which the phoneme /a/, is uttered by a male speaker and presented continuously in a high-quality (HQ) version, interrupted by a disturbed version of that phoneme.¹ As distortion, signal-correlated noise generated by a Modulated Noise Reference Unit (MNRU) [19], was chosen. This degradation is well suited for an initial check of the test set-up, as noise is one of the most ubiquitous factors that hamper efficient communication [39] and will therefore most definitely have an impact on quality perception. In addition to its impact on quality perception, the MNRU is recommended by the International Telecommunication Union (ITU) and can be used to compare, e.g., different speech transmission systems. As a well known degradation factor with an almost guaranteed influence on quality perception, signal-correlated noise was selected to achieve the degradation effect. The extent of the distortion was varied at four levels, i.e., from LQ1 to LQ4, where LQ1 (low quality 1, LQ1) referred to the weakest distortion. It was hypothesized that the P300 peak amplitude and latency (see Sect. 1.3.4.3) 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 all circumstances.

¹ This chapter is based on a previous publication; text fragments, tables, and figures are based on Antons et al. [1]. Reprinted, with permission, from [1].

2.2 Methods

2.2.1 Participants

Ten right-handed students and personnel from the Technical University of Berlin participated in the experiment (six females, four males; average age = 28.20 years; SD = 8.49; range = 19–51 years old), all of them native German speakers. All participants reported normal auditory acuity and no medical problems. Handedness was assessed using an inventory from Oldfield (1971) [91]. Participants gave their informed consent and received monetary compensation. The experiments were conducted in accordance with ethical principles that have their origin in the Declaration of Helsinki.

2.2.2 Material

Fourteen vowel phonemes were used: /a/ undisturbed, /i/ undisturbed, and twelve disturbed versions of /a/ impaired with signal-correlated noise. None of these phonemes have lexical meaning in German. The vowel /a/ was selected due to the fact that it has an clear and strong onset, and in addition to this, a high energy expenditure. The vowel /i/ was selected, as it should be clearly distinguishable from the preceding vowel /a/. In order to account for possible individual differences in hearing sensitivity, a set of stimuli was selected for each participant individually, based on her/his detection rate. Out of the stimulus set as a whole, an individual sub-set of four stimuli was selected for each participant, based on the results of a pre-test. During the pre-test all fourteen stimuli were presented to the participants four times in the context of an opinion test. The task of the participants was to rate a stimulus as belonging to one of two classes, high quality (no degradation) and low quality (with degradation). A detection rate was calculated by dividing the number of correctly identified low-quality stimuli by the total number of stimuli presented in the corresponding category. Based on the resulting detection rate, a final stimulus selection of four stimuli was carried out. For every participant, those stimuli were selected that were closest to the targeted detection rates of 100, 75, 25, and 0% for the four selected stimulus levels. The correlated *signal-to-noise ratios* (SNR) for the complete stimulus set were set at: 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. In order to keep the acoustic variability minimal, only one version of each phoneme was selected. Intensities were normalized using the root mean square of the speech period in the sound file using the software Adobe Audition®. The duration of each stimulus was set at 200 ms. The stimuli were degraded by a MNRU according to ITU-T Rec. P.810 in a controlled and scalable way [90]. The median SNR for the deviant stimuli and for all participants can be found in Table 2.1.

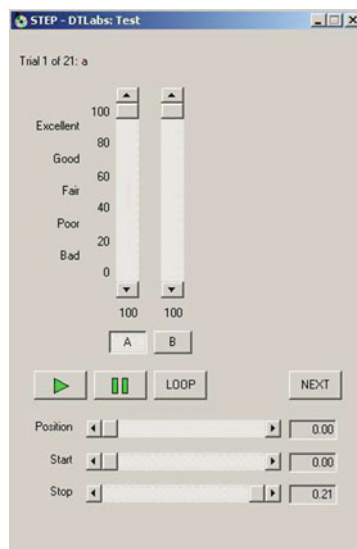
Table 2.1 SNR(db) for all participants and median SNRs

Participant	SNR(dB)				
	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

2.2.3 Experimental Design and Procedure

Under these experimental conditions, *oddball stimulus sequences* equalling 300 trials in total were presented (see Fig. 1.2 for a visualization of the oddball sequence). In each sequence, the undisturbed phoneme /a/ served as the *standard* stimulus (70 % of the trials), whereas the undisturbed phoneme /i/ as well as four selected disturbed versions of the phoneme /a/ served as *deviants* (6 % of the trials, respectively), delivered in a pseudo-randomized order, forcing at least one standard to be presented between successive deviants. As the oddball paradigm has not been frequently used for research studies in the telecommunications field concerned with quality so far, a control stimulus (/i/) was initially included as a sanity check, using a well-established P300 event [80]. An exploration of the P300 evoked by the /i/ stimulus showed indeed that a novelty P300 was consistently evoked in all participants. As a P300 for at least one degradation condition was eventually found for every participant, a further evaluation of the control stimulus (i.e. /i/) was not conducted. Per participant, eight to twelve sequences were recorded. All of these sequences contained six trials per degradation strength (SNR), respectively. Based on the behavioral results of each participant during the pre-test, an individual set of four stimuli was chosen for the experiment. As already mentioned, it was hypothesized that the four selected degradation levels that were selected would be detected with a rate of LQ4 = 100 %, LQ3 = 75 %, LQ2 = 25 %, and LQ1= 0 %. Stimuli were presented with an interstimulus interval varying from 1,000 to 1,500 ms (time between two consecutive stimuli). Participants were seated comfortably and 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 listening level preferred by the individual through Sennheiser® in-ear headphones. After the pre-test and physiological measurement, participants additionally had to rate all 12 stimuli—the complete stimulus set—in

Fig. 2.2 Graphical user interface of the STEP software. Used to collect the subjective quality ratings of participants on a continuous quality scale (CQS), ranging from bad (0) to excellent (100). An adapted version—without hidden references and anchors—of the MUlti Stimulus test with hidden reference and anchor (MUSHRA) methods



respect to their quality. For this test, the *MUlti Stimulus test with Hidden Reference and Anchor* (MUSHRA) techniques in accordance with ITU-R Rec. BS.1534-2 [33] was adapted to match as closely as possible the EEG experiment that was carried out. Therefore, no hidden reference and no anchor stimulus were used. Two stimuli were visually presented at the same time and had to be rated on a *continuous quality scale* (CQS) ranging from bad (0) to excellent (100). The Audio Research Lab software STEP (see Fig. 2.2) was used for collecting the data.

An experimental session lasted approximately 3 h (plus additional time for electrode application and removal), including breaks to avoid participant fatigue.

2.2.4 Electrophysiological Recordings

The EEG (Ag/AgCl electrodes, Brain Products GmbH, Garching, Germany) was recorded continuously using 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) [65]. 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 kOhm. The signal was digitized with a 16-bit resolution and a sampling rate of 1,000 Hz.

2.3 Data Analysis

2.3.1 Behavioral Data

Two parameters were derived as behavioral data during the EEG measurement. First, the reaction time (from tone onset to button press in milliseconds) for the different stimuli, and secondly, the psychometric functions. The reaction time for each stimulus class was measured in milliseconds, as the duration between the onset of stimulus presentation and the reaction of the participant (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 accordance with least-squares models [40]. After the EEG measurement, participants had to complete an opinion test, more specifically, they were asked to rate all LQ levels on a scale from excellent to bad. The slider of the Audio Research Lab software STEP (see Fig. 2.2) was set by the participants according to the *continuous quality scale (CQS)* in compliance with ITU-R Rec. BS.1534-2 [33] from excellent (100) to bad (0).

2.3.2 ERP Data

Off-line signal processing was carried out using the MATLAB® toolbox EEGLAB [92]. 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—the time interval around one stimulus, as well as one trial—with a length of 1,400 ms, time-locked to the onset of the stimuli, and 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 (−200–1,200 ms around stimulus onset) with an amplitude change exceeding 100 microvolt 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, the peak latency and peak amplitude were measured in a fixed time frame relative to the pre-stimulus baseline (see Fig. 1.3). The time frame for P300 quantification was set from 200 to 1,000 ms after stimulus onset. The maximal positive amplitude in this time frame was automatically determined; its voltage and latency were extracted for further analysis.

2.3.3 Classification

The 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. The detailed selection of classes can be found in Sect. 2.4.3 below. The classification was done using the MATLAB® toolbox BCILAB [93]. The comparison of ERP data with classifications is usually done by comparing the HQ versus the LQ ERPs. Features were the averaged voltages for the time windows 200–400, 400–600, 600–800, and 800–1,000 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 [94]. In respect to ERP signals, LDA is most suitable for classification purposes (for detailed information on single-trial classification of EEG data see [73]). An LDA with automatic regularization of the estimated covariance matrix and utilizing shrinkage procedures was applied.

2.4 Statistical Analysis

2.4.1 Behavioral Data

A Milton Friedman Test with a post-hoc comparison was calculated for reaction times [95]. As regards the opinion test, an analysis of variance (ANOVA)—with *degradation intensity* as the independent variable and the *mean opinion score (MOS)* as the dependent variable—was also calculated [96].

2.4.2 ERP Data

Deviance-related effects, namely, the presence and amplitude of P300 responses, were analyzed on the basis of data from the Cz electrode where P300 is typically at its maximum. Whereas in the present experiment 64 electrodes were used, the long-term goal for future research is to identify a minimal electrode placement providing a reliable response estimate in the majority of participants. Accordingly, in a pre-analysis a grand average was calculated and the one single electrode exhibiting the mean maximal P300 amplitude was identified. As this was found at the vertex,² all further analyses were carried out using the Cz electrode. Figure 2.4 shows exemplary ERPs for different stimuli classes. In order to test the presence of P300 under controlled conditions (/i/), deviant responses were compared to the corresponding standard responses, evoked by the undisturbed standard phoneme, by means of dependent t-tests. The minimum number of epochs constituting the ERP was set at 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, post-hoc comparisons were drawn between target types pairwise with Sidak-adjusted alpha levels.

² The uppermost surface of the head.

2.4.3 Classification

Classification was carried out participant-wise using bandpass-filtered raw data (0.2–7 Hz). Each LQ class was further divided into two separate subclasses: hits (true positives) and misses (false negatives). Stimuli that were degraded and not detected by the participants were labeled as misses. Detected degradations were labeled as hits. Two classifications were completed: (1) the training of a classifier to distinguish between hits and correctly reported HQ trials and testing this classifier on the same events (HQ against hits of each class), (2) once again the training of a classifier to distinguish between hits and correctly reported HQ trials, but alternatively, later testing on misses versus correctly reported HQ trials. For training purposes in the second classification, one half of the HQ trials as well as the hit trials of each LQ class were utilized. Two separate sets of HQ trials (HQ1 and HQ2) were created for the second classification by selecting even and odd HQ trials and assigning them to HQ1 and HQ2, respectively. For testing purposes, the other half of the HQ trials and the missed trials were used. This approach was first introduced in [42]. The analysis was carried out for all stimulus levels with a 5-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, would classification be carried out (minimal number of trials needed to train and test a classifier). The classification of hits within each target class versus HQ demonstrates that the classification of neural reactions based on the perception of degraded stimuli is feasible. The second analysis, classifying misses versus HQ trials, revealed differences in the EEG signals due to degradations which were not noticed at the behavioral level. Nonetheless, these two classes probably differ on a physiological level, as the degradation is still processed at the neural level. Classification performance was measured in terms of balanced accuracy (see also Sect. 1.3.4.5), expressed as the *Area Under the Curve* (AUC) of the *Receiver Operating Characteristic* (ROC), see Eq. 2.1 [84].

$$AUCb = \frac{\left(\frac{tp}{(tp+fn)} + \frac{tn}{(fp+tn)} \right)}{2} \quad (2.1)$$

Balanced accuracy stands for the relationship defined by 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 > 0.9 reflects excellent classification and AUCb = 0.5 chance level. The significance level of the classification outcome was tested—as to whether the classification result was significantly different to chance occurrence—using a Wilcoxon rank-sum test.

2.5 Results

2.5.1 Behavioral Data

For the opinion test, the ANOVA with *Degradation Intensity* as the independent variable, and the *Mean Opinion Score (MOS)* as the dependent variable in the opinion test data, proved to be a significant influence on the factor *Stimulus* (strength of degradation) ($F(131,4404) = 306.64, p < 0.01, \eta^2 = 0.76$). The post-hoc test (Sidak adjustment for pairwise comparisons) attained significance at a level of 21 dB ($p < 0.05$) compared to non-degraded stimuli. The psychometric function fits for all participants have been plotted in Fig. 2.3.

The mean reaction times for the different conditions can be found in Table 2.2.

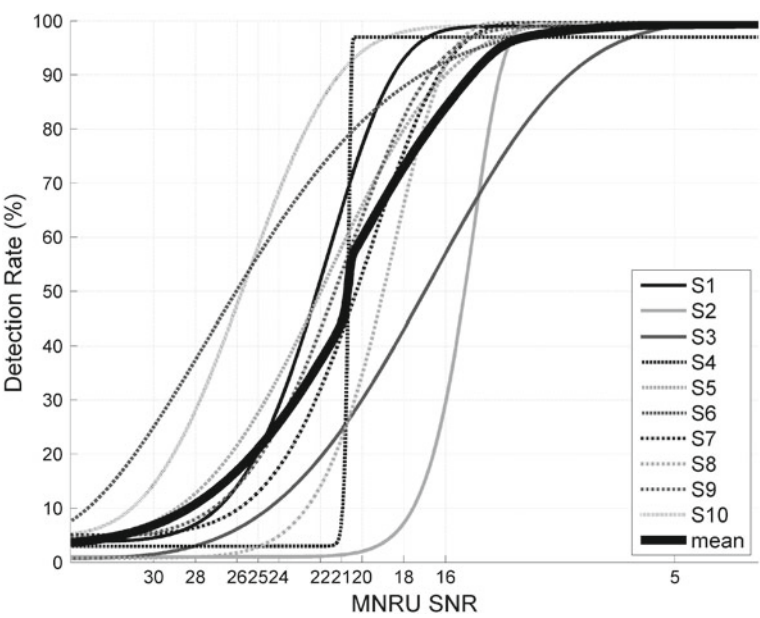


Fig. 2.3 Psychometric function fits from the psychophysical data; for all participants (participant (S) 1–10) and the average across all participants. Reprinted, with permission, from [1]

Table 2.2 Mean reaction times in milliseconds for all levels of degradation

	Milliseconds			
	LQ1	LQ2	LQ3	LQ4
Mean	724.93	736.28	749.24	598.54

LQ1 represents the weakest degradation and LQ4 the strongest

The reaction time for LQ1-3 are on a similar level, but significantly different compared to LQ4 ($p < 0.05$). For stimulus condition LQ4, the reaction time was shorter.

2.5.2 ERP Data

The time frame for *P300 peak* quantification was set at 200–1,000 ms after stimulus onset; within this interval, the maximum value of the ERP curve was identified and *peak latency and amplitude* were extracted as parameters. In order to test for the general presence of a P300 response evoked by the test set-up, it was checked whether the control stimulus (/i/) evoked a significantly different peak amplitude in comparison to the P300 peak amplitude of the standard (HQ, /a/ in high quality). As these two stimuli are clearly distinct on a physical level, and easily distinguishable if participants were asked to tell which stimulus had been presented, a clear neuronal response was anticipated. As regards testing to determine whether the difference defined by the *P300 peak amplitude* of the grand average was in fact significant, a two-tailed dependent t-test was performed. The t-test result was clearly significant ($t = 6.37$, $p < 0.01$). Due to the fact that the usage of the stimulus /i/ was only intended to check whether the test set-up was suitable for evoking a P300 response, the results connected with this stimulus will not be further pursued.

The ANOVA for repeated measurements revealed a significant main effect on the factor *Stimulus* ($F(9,27) = 3.56$, $p < 0.01$, $\eta^2 = 0.54$). Figure 2.4 shows the grand average ERPs, the arrows indicate the location of the *P300 peak* for each LQ. For dependent variable *P300 peak amplitudes*, a significant effect was found ($F(3,9) = 11.34$, $p < 0.05$, $\eta^2 = 0.79$), as well as for the dependent variable *P300 peak latency* ($F(3,9) = 9.35$, $p < 0.05$, $\eta^2 = 0.75$). The pairwise comparison (Sidak adjustment for pairwise comparisons) revealed a significant difference between LQ2 and LQ4 for the *peak amplitude* ($p < 0.05$). A significant effect could be found between LQ2 and LQ3 for the *peak latency* ($p < 0.05$), in addition to a significant effect between LQ2 and LQ4 for the *peak latency* ($p < 0.05$). Figure 2.5 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, a correlation between the *P300 amplitude* for electrode Cz and the detection rate was found ($r = 0.42$, $p < 0.05$). Within the ERP data, a negative correlation between the *P300 amplitude* and the *P300 latency* at electrode Pz could be observed ($r = -0.33$, $p < 0.10$).

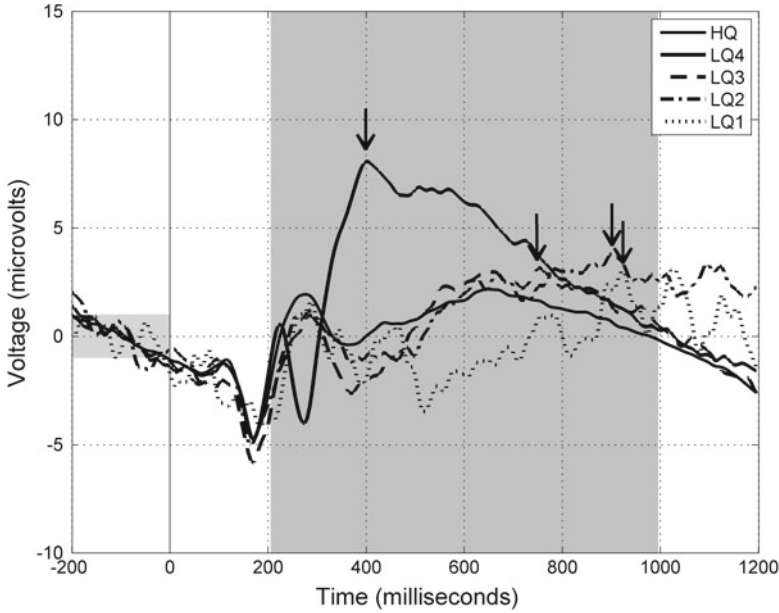


Fig. 2.4 Grand average of ERP plots for HQ and LQ1-4 at channel Cz. For HQ, correctly rejected trials (whereby no quality loss was perceived) and for LQ1-4 hits (whereby a quality loss was perceived) were utilized. *Arrows* denote P300 peaks. Number of trials used for the grand average of the ERP plots per class: HQ = 22,832, LQ4 = 3,268, LQ3 = 1,332, LQ2 = 610, and LQ1 = 165. Reprinted, with permission, from [1]

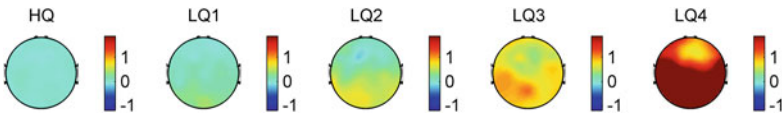


Fig. 2.5 Scalp topographies for all channels. Each *circle* depicts a top view of the head, with the nose pointing upwards. *Colors* code the mean voltage (microvolts) for the time interval from 300 to 1,000ms after stimulus onset. For LQ1-4, hits were used and for HQ, correctly rejected trials were used. Reprinted, with permission, from [1]

2.5.3 Classification

The classification results can be found in Fig. 2.6. At 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 = 0.92 ($p < 0.01$), LQ3: AUCb = 0.85 ($p < 0.01$), LQ2: AUCb = 0.76 ($p < 0.05$), and LQ1: AUCb = 0.70 (ns).

The second classification level reached the following values; for LQ4: not enough misses, LQ3: AUCb = 0.61 ($p < 0.05$), LQ2: AUCb = 0.55 (ns), and LQ1: AUCb = 0.51 (ns).

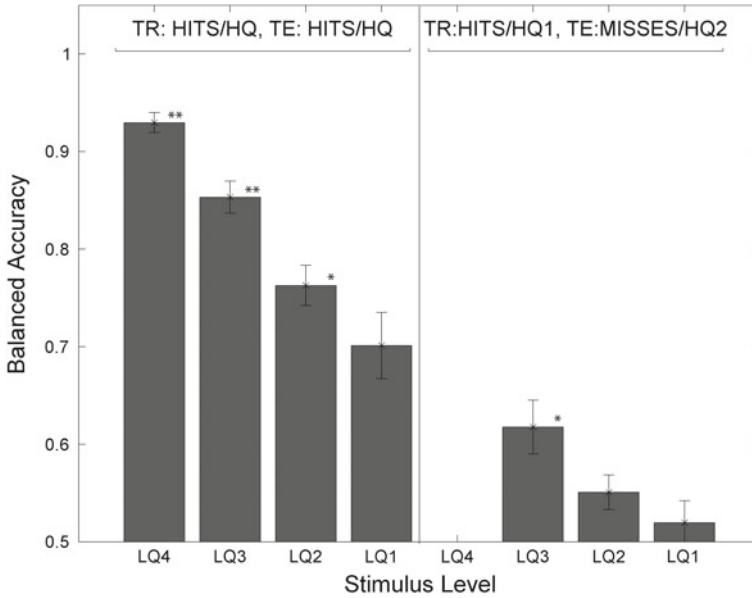


Fig. 2.6 Classification results. *Bars* show the average classification performance (balanced accuracy value). *Left* Trained (TR) on hits versus HQ and tested (TE) on hits versus HQ; *Right* Trained (TR) on hits versus HQ1 and tested (TE) on misses versus HQ2; for all stimuli LQ1-4. The bar for LQ4 is missing, as no participant had enough misses for testing (classification 2). Number of participants 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. *Asterisks* denote the significance level of the classification outcome in a Wilcoxon rank-sum test (one *asterisk* for $p < 0.05$, and two for $p < 0.01$). Reprinted, with permission, from [1]

It should be noted here that classification could not be carried out for all participants (due to a small number of hit/miss trials), so that the average values reported here are averages calculated over subsets of participants (classification 1: LQ1 = 4, LQ2 = 7, LQ3 = 9, LQ4 = 10; classification 2: LQ1 = 4, LQ2 = 7, LQ3 = 8, LQ4 = 0).

2.6 Discussion

An analysis of the *opinion test* ratings revealed that quality was rated as significantly lower from an SNR below 21 dB and downwards. This point denotes the threshold at which the quality was perceived as significantly worse in comparison to the reference value. The reaction time for the strongest degradation was shorter compared to the weaker ones, implying that participants were faster in detecting the degradation and providing the corresponding rating. The psychometric functions showed that the

mean detection rate surpassed 50 % at the 21 dB degradation level. This result is similar to the results of the opinion test.

For the *ERP data*, the significant P300 generated by the control stimulus (i.e. /i/) showed that the experimental set-up was appropriate for its purposes. 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 had been reported. In the context of the present paradigm, one could argue that minor physical stimulus differences were initially detected, yet an internal response criterion had not been met, so that an overt behavioral report was not initiated. The effects of *P300 peak latency* discovered here showed that the harder it was to detect a degraded stimulus, the later a P300 was evoked. This could 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 varies 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 an average across all participants 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) ERPs, which have been described at more frontal or more parietal sites [80]. At the first level of classification, it was demonstrated that the brain reaction due to the processing of a degradation—in this case the difference between the undisturbed and disturbed stimuli—can be well detected. With the second classification, it was shown that the pattern of brain activation related to consciously processed degradations can also be detected in trials which are not reported as degraded on a subjective level. It was concluded 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 exposed to small degradations over a long period of time (for measuring fatigue using EEG see [97]).

2.7 Length Influence Experiment

The developed test set-up cannot be implemented with every possible stimulus length and degradation class. In this remaining part of this chapter, it will be shown how the length of speech stimuli and the headphone type may influence the resulting subjective speech quality ratings. Therefore, three different lengths of stimuli will be analyzed (length of phonemes, words, and sentences).

2.7.1 Introduction

The aim of this experiment was to determine if (1) the length of a stimulus has an influence on the subjective rating and if (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 a stimulus used for tests in telecommunication research is around 8 s [19], which is much longer than the length common in ERP research (between 100 and 1,500 ms). In addition to this, common speech quality tests are carried out with circumaural headphones rather than in-ear headphones, the latter being typical for EEG set-ups. The result will reveal if the headphone type caused an unwanted influence on degradation perception.

2.7.2 Methods

2.7.2.1 Participants

Twenty volunteers (ten female, ten male; average age = 24.32 years; SD = 3.54; range = 22–28 years old; all right-handed), only native German speakers took part in this experiment. All participants reported normal auditory acuity. They provided their informed consent and received monetary compensation.

2.7.2.2 Material

For this experiment, stimuli of three different lengths were used: phonemes, words, and sentences. The phoneme from the first experiment (validation of test set-up, see Chap. 2) was used: /a/ (200 ms). The stimulus with the length of one word was the German translation of “eyebrow” /Augenbraue/ (1,200 ms), and a sentence shortened from the EUROM data base (following [98]) uttered by a male speaker was also used (8,000 ms) [99]. The two tested headphones were Sennheiser in-ear headphones and AKG over-ear headphones. Stimuli were degraded with the use of signal-correlated noise at the following SNRs: 5, 10, 14, 16, 18, 20, 21–35 dB in one-dB increments.

2.7.2.3 Experimental Design and Procedure

As in the first experiment (Chap. 2), participants had to rate all stimuli on a *continuous quality scale (CQS)* ranging from bad (0) to excellent (100). The stimuli types (three different lengths) were judged on all levels of degradation (four levels) and with both headphones (in-ear versus over-ear).

2.7.3 Statistical Analysis

The data was analyzed performing an ANOVA for repeated measures with *type of headphone* and *length of stimulus* as the independent variables and the *subjective quality rating* as the dependent variable.

2.7.4 Results

A significant main effect for the factor: *length of stimulus* was found ($F(21,4404) = 598.19, p < 0.01, \eta^2 = 0.15$).

As can be seen in Fig.2.7, ratings for the short phoneme stimulus (average rating = 82.26) were significantly higher compared to the stimuli in word (average

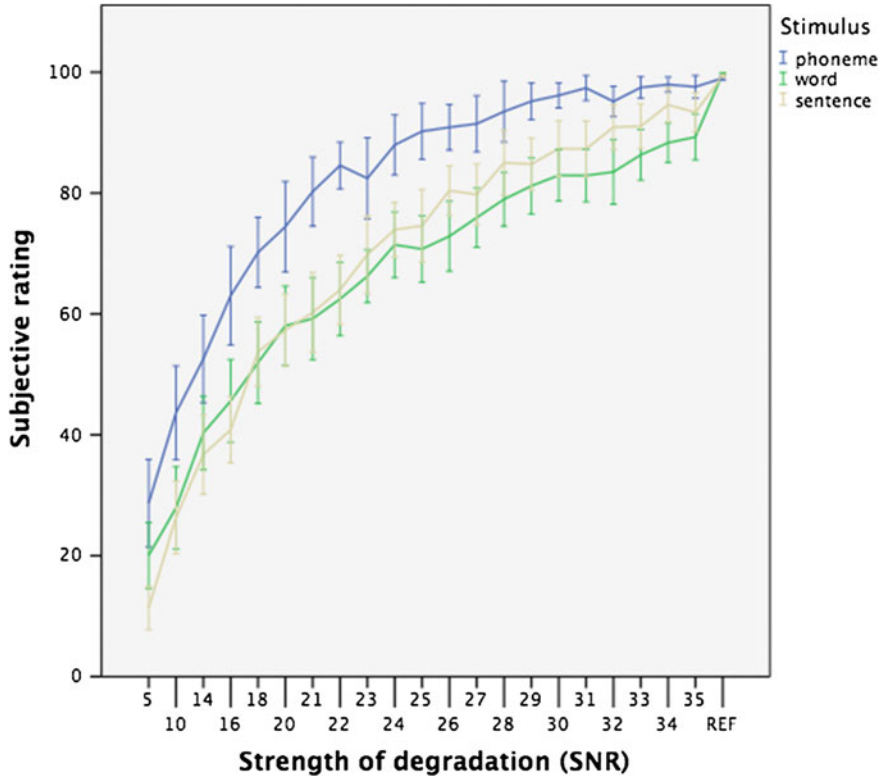


Fig. 2.7 Average rating from the *continuous quality scale (CQS)* across all participants. Subjective rating have been displayed as a function of degradation strength. REF denotes the clean reference stimulus. The lengths of stimuli have been color-coded. Whiskers denote the 95 % confidence intervals

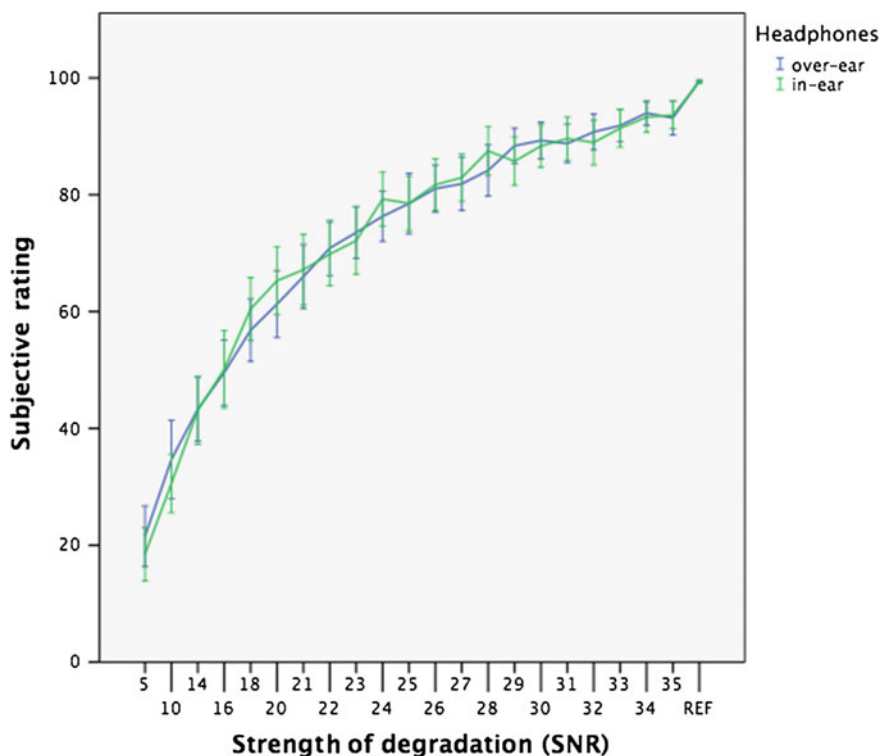


Fig. 2.8 Average rating from the *continuous quality scale (CQS)* across all participants. Subjective ratings have been displayed as a function of degradation strength. REF denotes the clean reference stimulus. The type of headphone used has been color-coded. Whiskers denote the 95 % confidence intervals

rating = 68.00) and sentence length (average rating = 70.13). There was no significant difference between stimuli with differing lengths of words and sentences.

There was no main effect for the *type of headphone* ($F(1,4404) = 0.58$, ns). As can be seen in Fig. 2.8, the confidence intervals overlap for stimuli played with in-ear (average rating = 73.52) and over-ear headphones (average rating = 73.41).

The post-hoc analysis (Sidak adjustment for pairwise comparisons) revealed a significant difference between the stimuli associated with the length of phonemes and words ($p < 0.01$), as well as for the difference between phonemes and sentences ($p < 0.01$). The difference between stimuli associated with the length of words and sentences was not significant.

2.7.5 Discussion

As expected from the *Phonemes Experiment* (Chap. 2), the *level of degradation* had an influence on subjective judgement. The factor *length of stimulus* had a significant effect on the subjective quality rating. A significantly higher quality was assigned to stimuli associated with the length of phonemes compared to stimuli associated with the length of words and sentences. There was no difference between the judgment of word-long and sentence-long stimuli. This leads to the conclusion that stimuli for subjective experiments on quality should consider stimuli minimum of word length. For EEG experiments concerning quality, it can be argued that stimuli should have word length as well. It still remains unclear whether the sensitivity of physiological measurement is higher, and therefore, here shorter stimuli can also possibly be valid for testing. As there was no difference between the ratings of *the two headphone types*, the influence may be negligible.

The *Phonemes Experiment* (Chap. 2) used short stimuli (vowels) which are a standard in ERP studies and this allowed the use of established physiological knowledge to interpret new findings and their implications for the cerebral processing of stimulus quality. However, in quality research, longer stimuli are employed for the behavioral detection of stimulus degradation. Correspondingly, the *Length Influence Experiment* directly compared the effects of stimuli differing in length (vowels, words, sentences) on the perceived quality in an opinion test. Indeed, longer stimuli (i.e., words or sentences) permit the better detection of minor stimulus degradation. In response to this behavioral result, the second combined experiment (the *Word Experiment*) will be introduced in the next chapter (Chap. 3), which makes use of word stimuli, thereby linking the ERP results presented here directly to the subjective standards in quality research.

2.8 Chapter Summary

In this chapter, a combined test set-up was introduced. During the stimulation of short speech stimuli (phonemes) of varying quality, EEG signals were measured. In addition to this, an opinion test was carried out. The results show that the physiological response, measured as parameters of an ERP component, can be used to gain insight into the perceived stimulus material. The P300 parameters vary with degradation strength, and therefore, can probably be used to estimate the quality of the presented stimulus material. As a second major contribution, the test set-up was validated using short speech stimuli and signal-correlated noise as degradation. This result is only true for the average response across a group of participants and can vary if applied to single participants. For the comparison with standard subjective tests, averaging methods using subjective, instrumental, and then physiological variables is a valid approach when it comes to standardization and technological developments intended for larger groups of listeners.

Neural Correlates of Quality Perception for Complex
Speech Signals

Antons, J.-N.

2015, XIV, 97 p. 33 illus., 22 illus. in color., Hardcover

ISBN: 978-3-319-15520-3