

ACCEPTED MANUSCRIPT

Neural correlates of speech quality dimensions analyzed using electroencephalography (EEG)

To cite this article before publication: Stefan Uhrig *et al* 2018 *J. Neural Eng.* in press <https://doi.org/10.1088/1741-2552/aaf122>

Manuscript version: Accepted Manuscript

Accepted Manuscript is "the version of the article accepted for publication including all changes made as a result of the peer review process, and which may also include the addition to the article by IOP Publishing of a header, an article ID, a cover sheet and/or an 'Accepted Manuscript' watermark, but excluding any other editing, typesetting or other changes made by IOP Publishing and/or its licensors"

This Accepted Manuscript is © 2018 IOP Publishing Ltd.

During the embargo period (the 12 month period from the publication of the Version of Record of this article), the Accepted Manuscript is fully protected by copyright and cannot be reused or reposted elsewhere.

As the Version of Record of this article is going to be / has been published on a subscription basis, this Accepted Manuscript is available for reuse under a CC BY-NC-ND 3.0 licence after the 12 month embargo period.

After the embargo period, everyone is permitted to use copy and redistribute this article for non-commercial purposes only, provided that they adhere to all the terms of the licence <https://creativecommons.org/licences/by-nc-nd/3.0>

Although reasonable endeavours have been taken to obtain all necessary permissions from third parties to include their copyrighted content within this article, their full citation and copyright line may not be present in this Accepted Manuscript version. Before using any content from this article, please refer to the Version of Record on IOPscience once published for full citation and copyright details, as permissions will likely be required. All third party content is fully copyright protected, unless specifically stated otherwise in the figure caption in the Version of Record.

View the [article online](#) for updates and enhancements.

Neural correlates of speech quality dimensions analyzed using electroencephalography (EEG)

Stefan Uhrig¹, Gabriel Mittag¹, Sebastian Möller^{1,2},
and Jan-Niklas Voigt-Antons¹

¹ Quality and Usability Lab, Technische Universität Berlin, 10587 Berlin, Germany

² German Research Center for Artificial Intelligence (DFKI), 10559 Berlin, Germany

E-mail: stefan.uhrig@qu.tu-berlin.de

Abstract. *Objective.* By means of subjective psychophysical methods, quality of transmitted speech has been decomposed into three perceptual dimensions named “discontinuity” (F), “noisiness” (N) and “coloration” (C). Previous studies using electroencephalography (EEG) already reported effects of perceived intensity of single quality dimensions on electrical brain activity. However, it has not been investigated so far, whether the dimensions themselves are dissociable on a electrophysiological level of analysis. *Approach.* Pursuing this goal in the present study, a high-quality (HQ) recording of a spoken word was degraded on each dimension at a time, resulting in three quality-impaired stimuli (F, N, C), which were on average described as being equal in perceived degradation intensity. Participants performed a three-stimulus oddball task, involving the serial presentation of different stimulus types: 1) HQ or degraded “standard” stimuli to establish sensory/perceptual quality references. 2) Degraded “oddball” stimuli to cause random, infrequent deviations from those references. EEG was employed to examine the neuro-electrical correlates of speech quality perception. *Main results.* Focus was put on modulations in the temporal and morphological characteristics of the P300 component of the event-related brain potential (ERP), whose subcomponents P3a and P3b are commonly linked to attentional orienting and task relevance categorization, respectively. Electrophysiological data analysis ($N = 28$) revealed significant modulations of P300 amplitude and latency by the perceptual dimensions underlying both quality references and oddball stimuli. *Significance.* The present study exemplifies the utility of physiological methods like EEG for dissociating speech degradations not only based on perceived intensity level, but also their distinctive quality dimension.

1. Introduction

Human listeners often perceive speech signals as possessing a certain transmission quality, which is usually communicated through free verbalizations or subjective ratings like in the context of a quality evaluation task [1]. Yet, such behavioral descriptions are only the end product of a hidden information processing chain within the experiencing person, being initiated by the occurrence of stimuli in the environment [2]. This presumed *quality formation process* includes all sensory, cognitive and motor processing

steps necessary for perceiving, evaluating and describing the quality of a given stimulus. According to current theoretical consensus, a “quality event” eventually results from the comparison of perceived and expected *quality features*, referring to a subset of stimulus features which are known to affect subjective quality experience. In other words, “quality” per se is identified with a specific cognitive judgment, resulting from a rather late evaluation process based on conscious reflection [3].

However, it appears plausible to assume that an immediate quality experience already arises from the perceptual analysis of an incoming sensory signal. The respective *percept* would then not only be influenced by present quality features, but also by the representation of previously perceived quality features stored in perceptual short-term memory, thus functioning as a *perceptual reference* (also termed “stimulus context” in the neurophysiology literature, see [4]): Each newly occurring stimulus is automatically compared to the current reference representation, presumably changing its perception and subsequent cognitive processing. How the quality of a given stimulus is perceived and judged should therefore always depend on the quality of its preceding stimuli.

1.1. *Quality assessment: Subjective versus Physiological*

The existence of reference effects in perception is of critical importance for quality assessment in applied settings and has guided the development and standardization of various quality rating paradigms and metrics by the International Telecommunication Union (ITU) [5]: In an absolute category rating (ACR) paradigm, participants judge the quality of single test stimuli. ACR ratings collected from a large sample are then averaged across participants, delivering mean opinion scores (MOS) for each stimulus. By contrast, the so called comparison category rating (CCR) paradigm requires the evaluation of each test stimulus’ quality relative to the quality of a previously presented reference stimulus. Afterwards, for each test stimulus another quality metric, the comparison mean opinion score (CMOS), is calculated.

Especially the combination of the ACR paradigm and MOS metric has been offering a reliable, easy-to-use and efficient way to quantify the average quality of various multimedia signals (audio/speech, visual, audio-visual). However, all these methods are essentially subjective and “black-box” in nature, i.e., they rely on active behavioral responses from human participants, who are instructed to retrospectively rate their private experiences in relation to test stimuli [6]. To also be able to capture internal processes leading up to perceptual representations and cognitive judgments—most of which are unconscious and therefore not accessible to introspection—physiological measures with high temporal resolution are needed [7]. Since perception and evaluation are realizations of brain activity, electroencephalography (EEG) proved to be the method of choice so far, referring to the non-invasive measurement of electrical fluctuations generated primarily in cortical regions of the brain from electrodes placed directly onto the scalp [8]. The EEG signal can be recorded continuously while human participants are exposed to defined test stimuli, picking up variations in neural activity at different

stages of internal information processing (e.g., sensation, perception and cognition) as well as changes in global brain states (e.g., arousal level, emotion and motivation). Features extracted from the time and frequency domains of the EEG signal are then providing neurophysiological indication concerning the impact of the perceived intensity and nature of quality degradations.

1.2. P300 event-related brain potential component: A neurophysiological indicator of perceived quality

Previous studies employing the EEG method to investigate the neural correlates of speech quality perception have focused primarily on the role of the late P300 event-related brain potential (ERP) component—a prominent positive deflection of the EEG signal triggered approx. 300 ms after the onset of behaviorally relevant events [4,9]. Different types of quality degradations have been shown to systematically affect P300 response characteristics: Increasing levels of signal-correlated noise induced on speech stimuli of different lengths (phonemes, words, sentences) had negative effects on the timing of the P300 relative to event onset (latency) as well as positive effects on its size (amplitude)—sometimes even if the magnitude of the induced degradation was below the threshold for conscious perception [10–14]. Furthermore, increases in bandwidth limitations and reverberation times correlated positively with the P300 amplitude [15,16]. Apart from the auditory modality, effects of quality impairments on event-related brain activity have also been demonstrated using visual and audio-visual stimuli (e.g., images, video clips) [17–22].

Decades of neurophysiological research have led to the conclusion that the “classic P300” waveform results from a superposition of two subcomponents, termed P3a and P3b, which are associated with separate perceptual processes [4,23]: Occurring with shorter latencies and most distinctively at fronto-central electrode positions, the P3a reflects enhanced attentional resource allocation after the presentation of novel stimuli, thus forming a part of the central component of the orienting response [24–26]. The P3b peaks later at centro-parietal electrode positions and is provoked by events categorized as relevant for the current behavioral task [4,23,27]. The two subcomponents of the P300 are commonly investigated using the *three-stimulus oddball paradigm* [4,28], during which stimuli are presented in random sequences and need to be categorized as instances of three stimulus types varying in presentation rate and task relevance (see Figure 1): A frequent “standard” (S) and two infrequent “oddballs”, namely the “target” (T) and the “distractor” (D). Only T is behaviorally relevant, since participants are instructed to respond to the occurrence of T (e.g., by pressing a push button) and to ignore the other stimulus types. While the P3a is triggered by presentation of D, both P3a and P3b are triggered by presentation of T, producing the superposed P300 waveform. In the auditory modality, the three-stimulus oddball paradigm usually evokes event-related neural responses with higher amplitudes for T than D, but equal latencies for both oddball types [29].

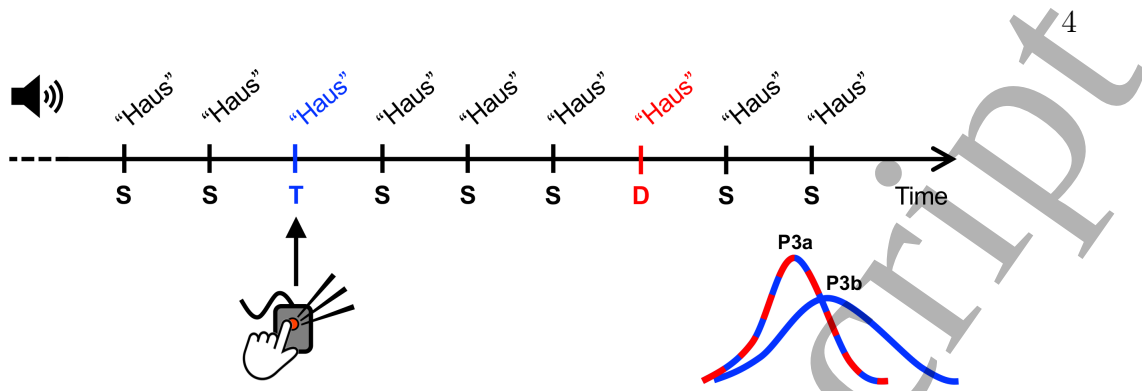


Figure 1: Active three-stimulus oddball paradigm. High-quality and quality-impaired word stimuli ("Haus") are repeatedly presented in random order. Each stimulus belongs to one of three stimulus types: Frequent standards (S = black), rare targets (T = red) and rare distractors (D = blue). In the main experiment, different assignments of stimuli to stimulus types were used (see Table 1). Participants are instructed to respond only to the occurrence of T by pressing a push button. Presentations of T trigger superposing P3a and P3b waveforms, presentations of D trigger only P3a waveforms. The schematic is based on Figure 7.1, Subfigure C, in [23].

1.3. Preliminary study: Critique

Prior applications of dimensionality reduction techniques to different types of subjective measures (e.g., factor analysis applied to semantic differential data, multidimensional scaling applied to stimulus pair comparisons) have revealed that the quality of speech stimuli, which are equalized in terms of loudness level, is decomposable into three independent perceptual dimensions [30–32]:

- "Discontinuity" (F), due to elimination of discrete portions of the speech signal (e.g., packet/frame loss, silence insertion).
- "Noisiness" (N), due to addition/multiplication of background/signal-correlated noise to the speech signal (e.g., environmental noise, circuit noise).
- "Coloration" (C) or timbre due to linear distortions (e.g., bandpass filtering, room acoustics).

It should be noted that in order to span a perceptual space for transmitted speech, other authors have discovered different kinds and numbers of dimensions by means of varying stimulus material and technical application conditions (see [32] for an extensive literature review).

A recent EEG study has adopted the three-stimulus oddball paradigm to examine passive listening situations in which the three quality dimensions described above would function as distinct perceptual references [33]. To this end, a high-quality (HQ) recording of a spoken word was degraded along each dimension at a time, resulting in three quality-impaired stimuli (F, N, C). Twelve stimulus sets were created, each containing a unique combination without repetition of four word stimuli (HQ, F, N, C) and three stimulus types of the three-stimulus oddball paradigm (S, T, D). By

comparison of selected stimulus sets, significant effects of perceptual references on P300 amplitude and latency were found predominantly within a P3a-related time window (350–550 ms). These results were in line with the assumption, that the P3a indicates the mobilization of attentional resources following the presentation of novel stimuli. However, less evidence was found for systematic neural modulations within a P3b-related time window (550–750 ms). Altogether, several critical points with regard to the design of this preliminary experiment need to be pointed out:

Loudness level. Prior to the oddball task, participants were instructed to adjust the audio output volume to their preferred loudness level. Despite creating a more ecologically valid listening situation, this procedure inevitably caused variation in loudness perception across participants, which is well-known to impact P300 characteristics [34,35]. Moreover, loudness is often considered to be a perceptual quality dimension of its own (e.g., like in another three-dimensional solution, “over-all clarity”, “distinction between signal distortion and background distortion”, “subjective loudness”, proposed by McDermott [36]; see also [32] for a detailed discussion). The individualized adjustments in loudness level might therefore have altered the perceived quality of all presented stimuli, adding further between-subject variability to the subjective and electrophysiological data.

Perceived degradation intensity. Participants judged all stimuli used in the experiment (HQ, F, N, C) before and after the oddball task in two different category rating paradigms (ACR, CCR). Both types of quality ratings revealed significant differences not only between HQ and each degraded stimulus, but also among the degraded stimuli themselves. Therefore, it could be argued that variation not in the nature, but in the perceived intensity of the induced degradations was the actual cause of the observed changes in P3a and P3b parameters.

Oddball stimulus probability. The relative frequencies of the two oddball types in each experimental block were not identical, with T always occurring in 20 % and D in 10 % of the total trial number per block. Since the P3b is only triggered by the occurrence of T, a higher presentation rate for T was expected to improve its signal-to-noise ratio (SNR) when averaging across trials. Although neural responses triggered by T are largely unaffected by changes in non-target presentation rates [37], stimulus probability normally exerts a clear influence on both subcomponents of the P300 [37–39]. As a consequence, separate statistical analyses had to be conducted for T and D, precluding the possibility to contrast both oddball types and test for effects of task relevance.

Conceptual framework of quality perception. Interpretation of the collected electrophysiological data was based on a conceptual framework of quality perception, which integrated earlier models from two independent lines of research in the fields of quality engineering and neurophysiology: The first model by Raake and Egger describes internal processes and representations leading to the formation of quality-related percepts and judgments in passively stimulated participants [2]. The second model by Polich explains the neuro-electrical phenomenon of the P300 as the updating of a stimulus context representation in working memory, as indicated by the P3b, which

is preceded by the mobilization of attentional resources as reflected by the P3a [4]. Hence, P3a and P3b were regarded as neurophysiological indicators of attentional and memory-related processing during the perception stage of the quality formation process, respectively. The results obtained from the experiment provided support for an influence of different perceptual references on attentional processing after the presentation of a degraded oddball, less so, however, for a modulation of subsequent memory-related processing. Besides, the original theoretical account by Polich lacked a sufficient explanation of why only the occurrence of T and not of D would cause an updating of the represented stimulus context, although *task relevance* constituted the only factor distinguishing between both oddball types. In past discussions, it has also been put into question as to whether the cumulative empirical evidence is actually supportive for the hypothesized connection between P3b and memory functions, but rather favors its association to mediating processes between perception and response initiation [40,41].

1.4. Present study: Extended replication

The present study realized an extended replication of the preliminary experiment [33]. Again, its primary goal was to clarify whether speech degradations induced on one of the three quality dimensions “discontinuity”, “noisiness” and “coloration” would modulate P300 characteristics. To examine the effects of sudden changes in quality dimension (*oddball stimulus*) and of different dimensions serving as perceptual references (*standard stimulus*), specific stimulus sets were compared against each other. Building upon the experimental design from the preliminary study, potential confounding factors described above (loudness level, perceived degradation intensity, oddball stimulus probability) were now controlled and the obtained results were interpreted within a revised conceptual framework of quality perception (to be described in subsection 2.2.9). Drawing valid inferences about underlying perceptual dimensions required that the degraded stimuli were on average perceived as possessing equal degradation intensities. For this purpose, a pretest was conducted prior to the main experiment to select three degraded stimuli, whose differences in subjective rating scores would not be statistically significant. Since the original study revealed a high congruency between MOS and CMOS values obtained from ACR and CCR tasks [33], the ACR was selected as the sole category rating paradigm.

In the following three sections, methods and results of both the pretest and main experiment will be presented and discussed. The final section 5 will review the most important findings and explicate possible directions for future studies.†

† A preliminary analysis of a subset of the collected experimental data (from a single stimulus set) was already presented at the poster session of the “10th International Conference on Quality of Multimedia Experience (QoMEX) 2018” and documented as a three-page-long article in the conference proceedings [42]. The present research article extends this preliminary analysis considerably (description of the pretest, analysis of all 12 stimulus sets, analysis of additional experimental factors, additional behavioral data analysis, additional ERP waveform analysis, change of electrophysiological data analysis method to mixed-effects modeling) thereby producing new results and beyond that provides an in-depth theoretical

2. Methods

2.1. Pretest

2.1.1. *Participants.* 16 healthy participants took part in the pretest (9 female; age: $M = 26.9$, $SD = 4.1$, $R = 15$ years; all but one right-handed). All participants were native German speakers without self-reported hearing deficits. At the end of the experiment, participants received a fixed monetary compensation. Data collection was conducted in accordance with ethical principles for medical research involving human subjects proposed by the World Medical Association (WMA) declaration of Helsinki.

2.1.2. *Stimuli.* As source material, the same clean speech file from the preliminary study was used, which consisted of a female voice uttering the German word /Haus/ (English: “house”), thereby ensuring a neutral meaning as well as clear and strong stimulus onset (duration: 500 ms, sampling rate: 48 kHz, bit rate: 16 bps). Its recording had originally been conducted in a sound-insulated environment with a high-quality condenser microphone (AKG Acoustics, C414 B-XLS/ST). Using MATLAB software (version R2017b), three methods were applied to degrade the clean, high-quality (HQ) audio file separately along each quality dimension (F = discontinuity, N = noisiness, C = coloration), producing three ranges of quality-impaired stimuli:

- Application of (multiplicative) signal-correlated noise generated by a modulated noise reference unit (MNRU) following ITU-T P.810 [43] produced ten N-impaired stimuli with distortion rates of $Q = 10, 12, 14, 16, 18, 20, 22, 24, 26, 28$ dB (Q : ratio of speech power to modulated noise power).
- Application of a bandpass filter with a low-cutoff frequency of 1.1 kHz and a high-cutoff frequency of 1.9 kHz produced ten C-impaired stimuli with lower transition bandwidths (i.e., width of rise on lower end of bandpass) of 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5, 0.55, 0.6 kHz.
- Application of random frame erasure (RFE) patterns generated following ITU-T G.191 [44] produced ten F-impaired stimuli with frame erasure rates (FER) of 11, 12, 13, 14, 15, 16, 17, 18, 19, 20 %.

Physical stimulus intensities were normalized to a target active speech level (ASL) of -26 dBov (dB relative to the overload point of the digital system) according to ITU-T P.56 [45]. First-hand listening by the authors confirmed good intelligibility of each word stimulus.

2.1.3. *Experimental procedure.* At the beginning of each testing session, participants were asked to report their age, gender, handedness, vision correction, known hearing problems and hours of sleep. Participants then performed an ACR task in accordance foundation (discussion of past theoretical approaches, description of a revised conceptual framework of quality perception, derivation of empirical hypotheses from this framework).

with ITU-T Rec. P.800 [5], during which they judged the quality of a given test stimulus on a discrete, five-point rating scale by pressing keys labeled with values “1”, “2”, “3”, “4”, “5” and corresponding German category labels “schlecht”, “dürftig”, “ordentlich”, “gut”, “ausgezeichnet” (English: “bad”, “poor”, “fair”, “good”, “excellent”). Each of the 31 stimuli (incl. 1 HQ, 10 F-impaired, 10 N-impaired, 10 C-impaired) was presented ten times, in randomized order.

2.1.4. Data collection. The word stimuli were presented binaurally through closed-ear headphones (Sennheiser HD 280) plugged into an audio interface (Cakewalk by Roland UA-25 EX), which was connected to a Windows PC running Psychophysics Toolbox Version 3 (PTB-3) for MATLAB. Prior to data collection, the sound pressure level (SPL) of the headphones was measured using a SPL meter (NTI Audio AL1 Acoustilyzer) and set to a constant comfortable listening level of approx. 65 dB SPL by adjusting the output volume control of the audio interface.

2.1.5. Data analysis. Using the “ez” package for R, a repeated-measures analysis of variance (ANOVA) was calculated with *stimulus* (31 test stimuli) as within-subject factor and ACR rating as dependent variable. For post-hoc pairwise comparisons, multiple paired t-tests with Holm correction were computed. Statistical significance level was set to $\alpha = 0.05$. As an effect size measure, generalized eta squared (η_G^2) was calculated [46].

2.2. Main experiment

The main experiment followed guidelines for the standardized employment of the EEG method in the context of speech quality assessment [14,47], specifically requiring:

- Use of minimal stimulus sets, including small numbers of stimuli and/or stimulus types with short presentation times and synchronous onsets.
- Use of well-established experimental paradigms with sufficiently high test-retest reliabilities and clear recommendations for application.

Involving the presentation of short word stimuli in the context of a three-stimulus oddball paradigm [4,9], the present experimental design fulfilled both requirements.

2.2.1. Participants. In total, 32 healthy participants took part in the main experiment, none of whom had attended the pretest. Four participants had to be excluded from data analysis due to high levels of noise in their electrophysiological recordings and poor behavioral task performance (exclusion criterion: zero hit rate in at least one block/stimulus set), leaving 28 participants (17 female; age: $M = 29.7$, $SD = 9$, $R = 38$ years; all but one right-handed; average 7.1 hours of sleep) for data analysis. Prior to giving his/her informed consent, each participant was handed an information about the

EEG method provided by the Deutsche Gesellschaft für Psychologie (DGPs; English: “German Psychological Society”).|| Apart from that, the procedure was identical to the pretest (see subsection 2.1.1).

2.2.2. Stimuli. Four stimuli, originally generated for the pretest (see subsection 2.1.2), were used in a three-stimulus oddball task (see Figure 1, described in section 1), one being of high quality (HQ), whereas the other three were degraded along each quality dimension at a time (F, N, C). It was assumed that the three degraded stimuli would on average be judged as possessing equal quality. Thus, they should be constant with regard to perceived degradation intensity and only vary in the nature of the induced quality impairment, i.e., the particular quality dimension.

2.2.3. Stimulus sets. Like in the preliminary study, twelve stimulus sets were generated by combining speech stimuli (HQ = high-quality, F = discontinuity-impaired, N = noisiness-impaired, C = coloration-impaired) and stimulus types of the three-stimulus oddball paradigm (S = standard, T = target, D = distractor) as listed in Table 1:

- Stimulus sets # 1–6 always assigned HQ to standard (S) and else comprised combinations without repetition of three degraded stimuli (F, N, C) and two oddball types (T, D).
- Stimulus sets # 7–12 consisted of combinations without repetition of degraded stimuli and stimulus types.

Considering natural limits in human sustained attention, no additional combinations with HQ assigned to either oddball type were included in the experimental design, which kept the duration of testing sessions at a tolerable level. Hereafter, stimulus sets will be named according to their specific stimulus to stimulus type assignments in the order of S-T-D, e.g., F-N-C standing for stimulus set # 7 with stimuli F, N and C assigned to stimulus types S, T and D, respectively.

2.2.4. Experimental procedure. Each testing session started with the collection of demographic data (same as in the pretest), which was succeeded by two experimental tasks (ACR, oddball), whose order was randomized across participants. The ACR task was conducted in the same manner as in the pretest, except that each test stimulus was presented only five times, in randomized order. The oddball task was composed of twelve blocks with each block only containing stimulus to stimulus type assignments of a particular stimulus set. Within each block, HQ and quality-impaired stimuli serving as the three stimulus types (S, T, D) were serially presented in pseudo-randomized order, such that at least one S fell in between two consecutive oddballs (T, D). Holding a push button in their dominant hand, participants were instructed to press it as fast and as accurately as possible whenever they detected T. Prior to the beginning of each

Table 1: Combinations of stimuli (HQ = high-quality, F = discontinuity-impaired, N = noisiness-impaired, C = coloration-impaired) and stimulus types (S = standard, T = target, D = distractor) and trial numbers (n) for twelve stimulus sets (#1–12) used in the main experiment.

#	S	T	D	n
1	HQ	N	C	100
2	HQ	C	N	100
3	HQ	F	C	100
4	HQ	C	F	100
5	HQ	F	N	100
6	HQ	N	F	100
7	F	N	C	200
8	F	C	N	200
9	N	F	C	200
10	N	C	F	200
11	C	F	N	200
12	C	N	F	200

new block, participants passively listened to four repetitions of the particular stimulus currently serving as T, after which they could manually start the test trials. Blocks with a degraded stimulus serving as S consisted of 200 trials, whereas those with HQ serving as S included only 100 trials in order to reduce overall testing time (see Table 1). The three stimulus types S, T and D always occurred in 70, 15 and 15 % of trials, respectively (in accordance with recommendations in [9]). A single trial comprised the presentation of a stimulus, followed by an inter-stimulus interval (ISI) of 1200 ms plus/minus a jitter of 200 ms. The order of blocks (i.e., stimulus sets) and trials within each block were randomized across participants. After completing a block, participants were offered to take a short break before starting the next one. All testing sessions were conducted in a quiet laboratory room according to ITU-R Rec. P.800 [5], with participants seated in front of a monitor, wearing an EEG electrode cap as well as closed-ear headphones with a constant SPL for stimulus presentation. The monitor screen was used to display task instructions and request quality judgments (in form of key presses) during the ACR task. It turned black with a central white fixation cross during blocks of the oddball task—with participants being instructed to keep their eyes open, to move as little as possible and look at the fixation cross. A single testing session lasted about two hours.

2.2.5. Data collection. The technical setup for controlling the experimental tasks and stimulus presentation was identical to the pretest (see subsection 2.1.4), except that now a MacBook Pro notebook computer was used for running MATLAB and sending event trigger to the EEG amplifier (g.USBamp by g.tec medical engineering GmbH). EEG signals were recorded via a medical-grade active electrode system (g.GAMMASys

with g.LADYbird electrodes) at 16 electrode positions (left hemisphere: Fp1, F3, C3, P3, PO7; right hemisphere: Fp2, F4, C4, P4, PO8; midline: Fz, FCz, Cz, CPz, Pz, Oz) in accordance with the international extended 10-20 system of electrode placement [48]. The reference and ground electrode were placed on the left and right mastoid, respectively. Electrode impedances were kept below 10 $k\Omega$. The sampling rate was set to 512 Hz. Three midline electrode positions, Fz, Cz and Pz, were of particular interest, across which the most prominent change in P300 amplitude known as “P300 scalp distribution” commonly occurs [4, 49].

2.2.6. Data processing. Using the EEGLAB toolbox (v13.6.5b) for MATLAB, (pre-)processing of the collected electrophysiological data followed a standard ERP processing pipeline based on the averaging method [50]:

- Down-sampling of the raw data to a sampling rate of 256 Hz.
- Application of a band-pass Hamming-windowed finite impulse response (FIR) filter with 0.1 and 40 Hz as low-and high-cutoff frequencies.
- Segmentation of the continuous EEG signal into stimulus-locked epochs with a pre-stimulus time of 200 ms and a post-stimulus time of 1000 ms.
- Baseline correction by subtracting the pre-stimulus baseline from each epoch time point.
- Rejection of epochs above/below maximum/minimum amplitude thresholds determined for each participant across all electrode positions, such that at most 10 % of epochs would be removed from further analysis.
- Computation of an independent component analysis (ICA) and subtraction of independent components associated with ocular activity (eye movements and blinks) from the segmented data.
- Single-subject averaging of single-trial epochs, resulting in single-subject average epochs for each participant, electrode position, stimulus set and stimulus type.
- Grand averaging of single-trial epochs, resulting in grand average epochs for each electrode position, stimulus set and stimulus type.

In addition, difference epochs were computed via the following processing steps: Since the number of single-trial epochs time-locked to S as eliciting stimulus (“standard epochs”) was always exceeding the number of epochs time-locked to either oddball type (“oddball epochs”), both numbers had to be equalized first by randomly selecting a subset of standard epochs. Afterwards, the standard epochs were subtracted from the oddball epochs, resulting in single-trial difference epochs. The single-trial difference epochs were then averaged separately for each participant, electrode position, stimulus set and oddball type, resulting in single-subject average T-S and D-S epochs. Final averaging of single-trial difference epochs across participants for each electrode position, stimulus set and oddball type resulted in grand average T-S and D-S epochs.

2.2.7. *Parameter extraction.* In order to quantify morphological and temporal characteristics of the P300, two ERP parameters were extracted from single-subject average difference epochs:

- *Peak amplitude*, defined as the maximum voltage value (“peak”) in a prespecified post-stimulus time window.
- *Peak latency*, defined as the time interval between stimulus onset and a peak within the post-stimulus time window.

Usually, extraction time windows are specified based on the typical timing of ERP components. Following an earlier study on P300 and speech quality [12], a post-stimulus time window ranging from 200 to 1000 ms was used for peak parameter extraction.¶ With regard to T and D, only trials classified as “hits” (i.e., response after presentation of T) and “correct rejections” (i.e., no response after presentation of S/D) were selected, respectively. Lastly, in order to quantify the task performance of participants within each block of the oddball task, two behavioral measures were calculated:

- *Hit response time*, defined as the time interval between the onset of T and the behavioral response, lasting at maximum 1 sec.
- *Hit rate*, defined as the relative frequency of hits (i.e., number of hits divided by the total number of Ts).

2.2.8. *Stimulus set comparisons.* Two experimental factors were varied via systematic comparisons of stimulus sets (see Table 1):

- *Oddball stimulus*: Variation of the degraded stimulus assigned to either oddball type (1: {F, N, C}, 2: {F, N}, 3: {F, C}, 4: {N, C}), while holding *standard stimulus* constant (1: HQ, 2: C, 3: N, 4: F). Realized by set comparisons listed in Table 2.
- *Standard stimulus*: Variation of the stimulus assigned to S (1: {HQ, F, N}, 2: {HQ, F, C}, 3: {HQ, N, C}), while holding *oddball stimulus* constant (1: C, 2: N, 3: F). Realized by set comparisons listed in Table 3.

For set comparisons # 1 and 5 listed in Table 2 and all set comparisons in Table 3, stimulus sets with identical stimuli assigned to S and either oddball type were merged beforehand, respectively, as symbolized by the “summation” of stimulus set indices.

2.2.9. *Revised conceptual framework of quality perception.* A revised version of the conceptual framework from the preliminary study was used to describe the internal

¶ Initial inspection of grand average ERP waveforms (see Figure 6) revealed a drastic shift in the succession of late ERP components for F versus N/C being presented as oddball stimulus. Because of this, it was concluded that separate fixed time windows for extraction of P3a and P3b peak parameters (like those used before in [33]) would not validly describe the P300 subcomponent structure in the present dataset.

Table 2: Set comparisons to analyze effects of *oddball stimulus* (HQ = high-quality, F = discontinuity-impaired, N = noisiness-impaired, C = coloration-impaired)

on P300 subcomponents for each *oddball type* (T = target, D = distractor).

#	Oddball type	Standard stimulus	Oddball stimulus	Set comparison
1	T	HQ	F vs. N vs. C	(3 + 5) vs. (1 + 6) vs. (2 + 4)
2	T	F	N vs. C	7 vs. 8
3	T	N	F vs. C	9 vs. 10
4	T	C	F vs. N	11 vs. 12
5	D	HQ	F vs. N vs. C	(4 + 6) vs. (2 + 5) vs. (1 + 3)
6	D	F	N vs. C	8 vs. 7
7	D	N	F vs. C	10 vs. 9
8	D	C	F vs. N	12 vs. 11

Table 3: Set comparisons to analyze effects of *standard stimulus* (HQ = high-quality, F = discontinuity-impaired, N = noisiness-impaired, C = coloration-impaired) on P300 subcomponents for each *oddball type* (T = target, D = distractor).

#	Oddball type	Oddball stimulus	Standard stimulus	Set comparison
1	T	F	HQ vs. N vs. C	(3 + 5) vs. 9 vs. 11
2	T	N	HQ vs. F vs. C	(1 + 6) vs. 7 vs. 12
3	T	C	HQ vs. F vs. N	(2 + 4) vs. 8 vs. 10
4	D	F	HQ vs. N vs. C	(4 + 6) vs. 10 vs. 12
5	D	N	HQ vs. F vs. C	(2 + 5) vs. 8 vs. 11
6	D	C	HQ vs. F vs. N	(1 + 3) vs. 7 vs. 9

processes and representations involved in oddball task performance as well as their associations with both behavioral and neurophysiological indicators:

As illustrated in Figure 2, a stimulus present in the external environment is transduced into a sensory signal and transmitted to the brain. During the initial sensory processing stage, the incoming sensory signal is compared with a sensory reference, i.e., a representation of previous stimulus features stored in sensory memory [51,52]. In awake persons, a significant mismatch between the sensory signal and the sensory reference immediately provokes the mobilization of attentional resources [24,25]. P3a amplitude and latency are positively and negatively related to the size of this mismatch, thus reflecting the amount and timing of early attentional orienting, respectively [53].

Based on the attended sensory difference pattern, various perceptual analysis steps (e.g., feature binding, grouping, object recognition) take place, ultimately leading up to the formation of a perceptual representation of the external environment with the *percept* as its experiential correlate. In compliance with oddball task requirements, the stimulus is then categorized as being either relevant (T) or irrelevant (D) for executive processing [27], which includes the planning, selection and generation of task-related actions [40, 54]. This categorization process requires a more or less effortful

discrimination between the perceived stimulus and the current perceptual reference. Working memory operations maintain the perceptual reference representation, but also delegate attentional resources to the discrimination process. The amount of attentional resources invested in discrimination (i.e., its cognitive load) directly depends on the difficulty to distinguish between the current stimulus and the perceptual reference. P3b amplitude and latency are negatively and positively related to this perceptual discriminability, indicating variation in cognitive load and processing time needed for discrimination/categorization, respectively [27, 53, 55, 56]. Finally, the execution of actions causes physical effects in the external environment (e.g., push button pressed), hence enabling their quantification in form of behavioral measures like response times and error rates.

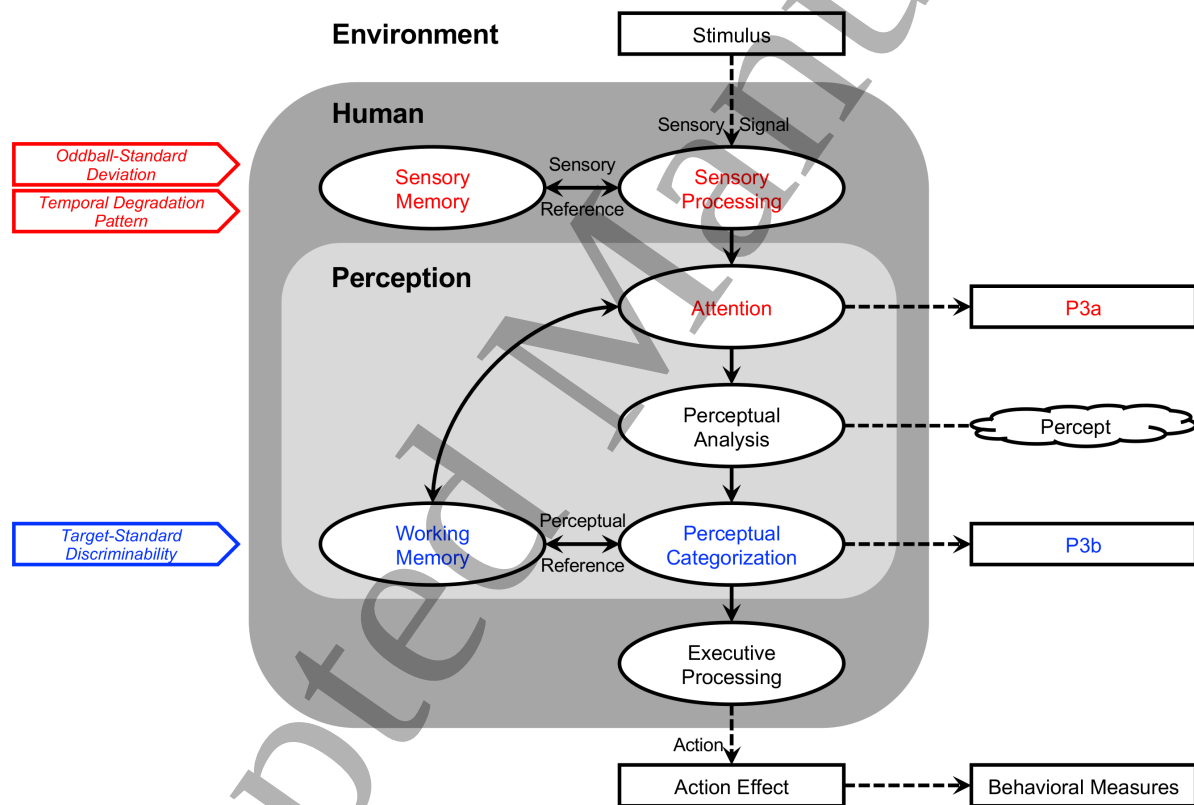


Figure 2: Revised conceptual framework of quality perception in the context of the three-stimulus oddball task (circle = internal process, rectangle = external event, pentagon = influencing factor related to the oddball task, cloud = subjective experience). The framework integrates earlier theoretical considerations about internal quality formation [2] and the functional role of P3a and P3b in attentional orienting [24], task relevance categorization [27] and response initiation [40].

2.2.10. Hypotheses. Hypotheses concerning the effects of *oddball stimulus* and *standard stimulus* on P300 peak parameters were derived from the revised conceptual framework

(see Figure 2, described in subsection 2.2.9 above). Three internal causes of electrophysiological effects were assumed:

Temporal degradation pattern: Over the course of the oddball task, rare oddball stimuli are randomly interspersed into the standard sequence. Sensory processing automatically matches the incoming sensory signals with current sensory references. In case HQ serves as S and therefore establishes a HQ sensory reference, the “discontinuous” degradation in F should be detected later than the “continuous” degradations in N and C: The earliest time point at which the presence of F could be detected was the first frame loss within the speech signal portions of the audio file, whereas N and C—due to being induced on the whole speech signal—were detectable already from the first speech sample onward. This lengthening of sensory processing would also delay subsequent attentional resource allocation. Thus, P3a latencies should be increased for F versus N/C serving as either oddball type (T, D).

Sensory oddball-standard deviation: P3a amplitude is positively related to the degree of deviation between the oddball and S. Since all quality-impaired stimuli were of equal sensory/perceived degradation intensity, it was expected that modulations of P3a amplitudes would occur only if HQ served as S (high oddball-standard deviation), but not if a quality-impaired stimulus served as S (low oddball-standard deviation).

Perceptual target-standard discriminability: P3b parameters are known to correlate with the difficulty to perceptually distinguish between T and S [53, 55, 56], which determines the expenditure of attentional resources to process T in working memory. In case HQ served as S (high T-S discriminability), the clear difference in perceived degradation intensities should permit an easy discrimination between both stimulus types. Hence, cognitive processing of T should be effortless and fast, causing an increase in P3b amplitude and a reduction in P3b latency [27]. On the contrary, when a degraded stimulus served as S (low T-S discriminability), T-S discrimination cannot be based on differences in perceived degradation intensity, but rather in the underlying quality dimension. In this case, more attentional resources should be consumed and processing of T be slowed down, as indicated by reduced P3b amplitudes and increased P3b latencies.

Taken together, the following effects of *oddball stimulus* and *standard stimulus* on P300 peak parameters (including post-hoc comparisons) were anticipated due to the three internal causes explicated above:

- (i) Effects of *oddball stimulus* on P3a peak latency ($F > N/C$) due to variation in temporal degradation patterns of quality-impaired oddball stimuli (see set comparisons # 1, 3–5, 7, 8 listed in Table 2).
- (ii) Effects of *standard stimulus* on P3a peak amplitude ($HQ > F/N/C$) due to variation in sensory degradation intensity between high-quality and quality-impaired oddball stimuli (see all set comparisons listed in Table 3).
- (iii) Effects of *standard stimulus* on P3a peak latency ($F > N/C$) due to variation in temporal degradation patterns of quality-impaired stimuli serving as perceptual

reference (see set comparisons # 2, 3, 5, 6 listed in Table 3).

- (iv) Effects of *standard stimulus* on P3b peak amplitude ($HQ > F/N/C$) due to variation in perceived degradation intensity between high-quality and quality-impaired target stimuli (see all set comparisons listed in Table 3).
- (v) Effects of *standard stimulus* on P3b peak latency ($HQ < F/N/C$) due to variation in perceived degradation intensity between high-quality and quality-impaired target stimuli (see all set comparisons listed in Table 3).

2.2.11. Data analysis. Using R, separate statistical analyses were performed on the collected subjective, behavioral and electrophysiological data:

- *Subjective data:* A repeated-measures ANOVA was calculated with *stimulus* (HQ, F, N, C) as within-subject factor and ACR rating as dependent variable.
- *Behavioral data:* Four repeated-measures ANOVAs were computed with either *block* (block 1–12) or *stimulus set* (see Table 1) as within-subject factor and either hit response time or hit rate as dependent variable.
- *Electrophysiological data – ERP parameters:* Linear mixed-effects models (LMEMs) assigning random intercepts to each level of the random effect were fitted to the electrophysiological data, thereby realizing stimulus set comparisons described in subsection 2.2.8:
 - Eight LMEMs were calculated with *oddball type* (T-S, D-S), *oddball stimulus* (1: {F, N, C}, 2: {C, F}, 3: {N, F}, 4: {N, C}) and *channel* (Fz, Cz, Pz) as fixed factors and *subject* (1–32) as random factor for each dependent variable (peak amplitude, peak latency) and constant *standard stimulus* (1: HQ, 2: N, 3: C, 4: F).
 - Six LMEMs were computed with *oddball type* and *standard stimulus* (1: {HQ, C, F}, 2: {HQ, N, F}, 3: {HQ, N, C}) as fixed factors and *subject* as random factor for each dependent variable (peak amplitude, peak latency) and constant *oddball stimulus* (1: N, 2: C, 3: F).
- *Electrophysiological data – ERP waveforms:* To examine whether grand average difference waveforms for varying factor levels deviated significantly from each other across the epoch time course—particularly within the time window used for P300 quantification (200–1000 ms)—multiple paired (“running”) t-tests were calculated across epoch time points.

Mixed-effects modeling was realized using the “nlme”⁺ package for R following recommendations on the statistical analysis of unbalanced ERP data by [57]. Otherwise, the procedure was identical to the pretest (see subsection 2.1.5). When violating the assumption of sphericity in ANOVAs, degrees of freedom were corrected with the Huynh-Feldt estimate of sphericity (ϵ_{HF}). Statistical significance levels of $\alpha = 0.05$ were

⁺ <https://cran.r-project.org/web/packages/nlme/>

assumed at each analysis step (subjective, behavioral, ERP parameter, ERP waveform) and Šidák-corrected (α_{SID}) for the number of computed tests (ANOVAs, LMEMs, running t-tests) to avoid alpha inflation.

3. Results

3.1. Pretest

Statistical analysis revealed a significant effect of *stimulus* on ACR rating ($F[30, 450] = 2.677, p < 0.001, \eta_G^2 = 0.151$). A discontinuity-impaired stimulus with 15 % FER (F), a noisiness-impaired stimulus with a distortion rate of $Q = 20$ dB (N) and a coloration-impaired stimulus (C) with transition bandwidth of 0.2 were chosen, whose MOS values deviated significantly from the high-quality (HQ) stimulus (all $p < 0.001$), but not among each other (all $p = 1$). Figure 3 shows MOS values of the four selected stimuli in the pretest.

3.2. Main experiment

3.2.1. Subjective analysis. Analysis of subjective data demonstrated a significant effect of *stimulus* on ACR rating ($F[3, 81] = 36.16, \epsilon_{HF} = 0.658, p < 0.001, \eta_G^2 = 0.413$). Post-hoc analysis revealed non-significant pairwise comparisons between the three degraded stimuli (F vs. N: $p = 0.91$, F vs. C: $p = 0.70$, N vs. C: $p = 0.70$), yet each degraded stimulus deviated significantly from HQ (all $p < 0.001$). Figure 3 depicts MOS values of four stimuli used in the main experiment.

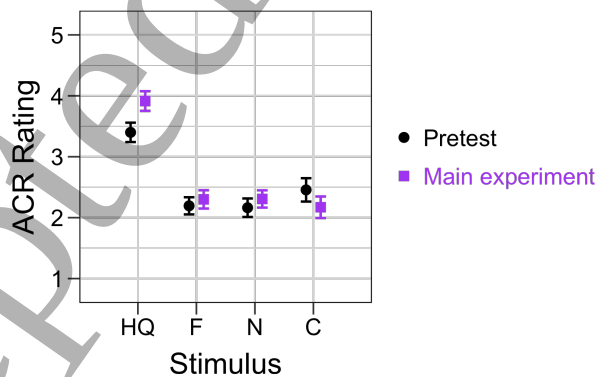


Figure 3: Effect of *stimulus* on ACR rating in the pretest ($N = 16$) and main experiment ($N = 28$). Error bars represent 95 % confidence intervals.

3.2.2. Behavioral analysis. Analysis of response measures revealed significant effects of *block* on hit rate ($F[11, 297] = 2.452, p = 0.006, \eta_G^2 = 0.083$) and of *stimulus set* on hit response time ($F[11, 297] = 19.683, p < 0.001, \eta_G^2 = 0.422$) as well as hit rate ($F[11, 297] = 2.357, p = 0.008, \eta_G^2 = 0.08$). However, only post-hoc pairwise comparisons of the effect of *stimulus set* on hit response time listed in Table 4 turned

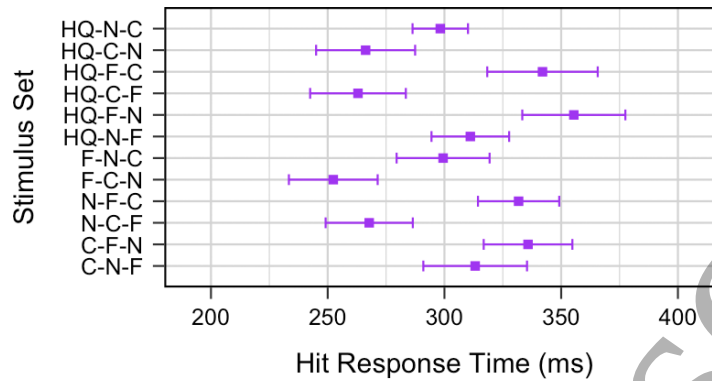


Figure 4: Effect of *stimulus set* on hit response time ($N = 28$). Plot axes were swapped to improve readability. Error bars represent 95 % confidence intervals.

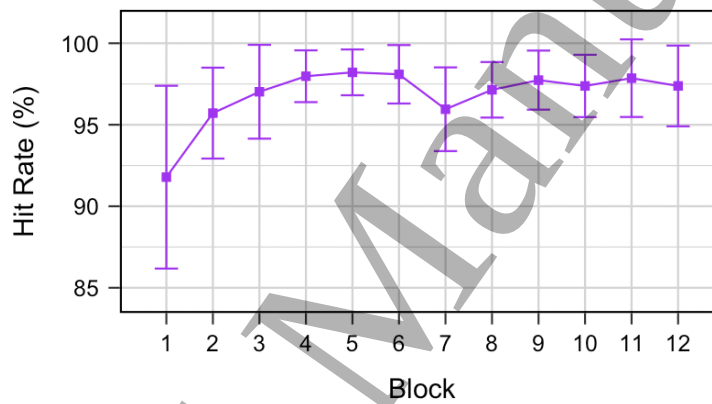


Figure 5: Effect of *block* on hit rate ($N = 28$). Error bars represent 95 % confidence intervals.

Table 4: Post-hoc analysis for the effect of *stimulus set* on hit response time ($N = 28$). Cells contain p values for each pairwise comparison. Shades of gray indicate statistical significance level (light gray: $p < 0.05$, medium gray: $p < 0.01$, dark gray: $p < 0.001$).

	HQ-N-C	HQ-C-N	HQ-F-C	HQ-C-F	HQ-F-N	HQ-N-F	F-N-C	F-C-N	N-F-C	N-C-F	C-F-N	C-N-F
HQ-C-N	0.249	-	-	-	-	-	-	-	-	-	-	-
HQ-F-C	0.0292	< 0.001	-	-	-	-	-	-	-	-	-	-
HQ-C-F	0.077	1	< 0.001	-	-	-	-	-	-	-	-	-
HQ-F-N	< 0.001	< 0.001	1	< 0.001	-	-	-	-	-	-	-	-
HQ-N-F	1	0.04	0.777	0.013	0.052	-	-	-	-	-	-	-
F-N-C	1	0.213	0.086	0.147	0.007	1	-	-	-	-	-	-
F-C-N	0.001	1	< 0.001	1	< 0.001	< 0.001	0.01	-	-	-	-	-
N-F-C	0.004	< 0.001	1	< 0.001	0.777	0.777	0.004	< 0.001	-	-	-	-
N-C-F	0.062	1	< 0.001	1	< 0.001	0.012	0.087	0.992	< 0.001	-	-	-
C-F-N	0.01	< 0.001	1	< 0.001	1	0.081	0.027	< 0.001	1	< 0.001	-	-
C-N-F	1	0.081	1	0.027	0.366	1	1	0.003	1	0.045	1	1

out significant. Figure 4 shows average hit response times as a function of *stimulus set*, Figure 5 depicts average hit rates as a function of *block*.

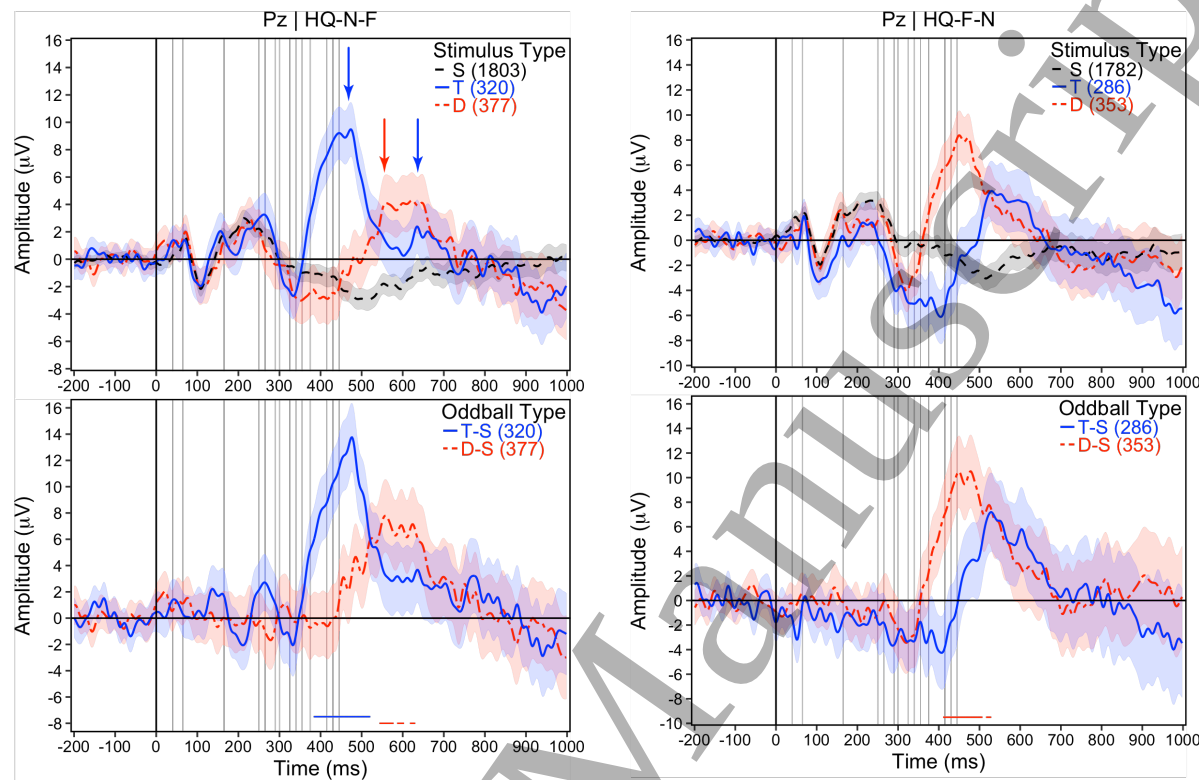


Figure 6: Grand average waveforms for stimulus sets HQ-N-F (left column) and HQ-F-N (right column) at electrode position Pz ($N = 28$). Top row: Grand average ERP waveforms per stimulus type (S, T, D), representing average neural responses to each stimulus type. Bottom row: Grand average difference waveforms per oddball type (T-S, D-S), representing average deviations of neural responses to each oddball type (T, D) from neural responses to S. Arrows point at presumed peaks of P3a and P3b subcomponents. Legends contain numbers of trials used for grand averaging. Vertical gray lines mark onsets of discrete frame losses during audio playback of the discontinuity-impaired stimulus (F). Lines of dots above time axis indicate, at which epoch time points each waveform significantly deviated from baseline. Error bands represent 95 % confidence intervals.

3.2.3. Electrophysiological analysis. Representative grand average ERP waveforms for stimulus sets HQ-F-N and HQ-N-F at electrode position Pz are depicted in Figure 6. Stimulus set comparisons were conducted to test modulations of P300 peak parameters within the post-stimulus time window from 200 to 1000 ms:

- Table 5 lists significant results obtained from the analysis of effects of *oddball type* and *oddball stimulus* ordered by standard stimulus and peak parameter. Exemplary grand average difference waveforms for set comparisons # 1 and 5 listed in Table 2, split up by *oddball type* (T-S, D-S) and *channel* (Fz, Cz, Pz), are plotted in Figure 7.

- Table 6 contains significant results for effects of *oddball type* and *standard stimulus* ordered by oddball stimulus and peak parameter. Figure 8 shows grand average difference waveforms for set comparisons listed # 1 and 4 in Table 3, split up by *oddball type* and *channel*.

Table 5: Effects of *oddball stimulus* on P300 peak parameters for constant standard stimulus ($N = 28$). All effects were statistically significant ($p < \alpha_{SID}$, $\alpha_{SID} = 0.0036$; α_{SID} : significance level $\alpha = 0.05$, Šidák-corrected for 14 LMEMs). Bold indices (#) point to correspondingly numbered subtables in Table A1 containing results from post-hoc analysis.

#	Oddball stimulus	Standard stimulus	Peak parameter	Effect	df_n	df_d	F
1	F vs. N	C	amplitude	oddball type	1	8196	19.266
2	N vs. C	F	amplitude	oddball type $\times$ oddball stimulus	1	8640	9.703
3	F vs. N vs. C	HQ	amplitude	oddball type	1	12843	11.915
4	F vs. C	N	amplitude	oddball type	1	8661	40.407
5	F vs. N	C	latency	oddball type $\times$ channel	2	8196	6.452
6	N vs. C	F	latency	oddball type	1	8640	15.85
7	N vs. C	F	latency	oddball stimulus $\times$ channel	2	8640	5.794
8	F vs. N vs. C	HQ	latency	oddball type	1	12843	22.81
9	F vs. N vs. C	HQ	latency	oddball stimulus	2	12843	13.801
10	F vs. N vs. C	HQ	latency	channel	2	12843	10.638
11	F vs. C	N	latency	oddball stimulus	1	8661	11.387
12	F vs. C	N	latency	oddball type $\times$ oddball stimulus	1	8661	10.092

Table 6: Effects of *standard stimulus* on P300 peak parameters for constant oddball stimulus ($N = 28$). All effects were statistically significant ($p < \alpha_{SID}$, $\alpha_{SID} = 0.0036$). Bold indices (#) point to correspondingly numbered subtables in Table A2 containing results from post-hoc analysis.

#	Standard stimulus	Oddball stimulus	Peak parameter	Effect	df_n	df_d	F
1	HQ vs. F vs. N	C	amplitude	oddball type	1	13440	41.478
2	HQ vs. F vs. N	C	amplitude	oddball type $\times$ standard stimulus	2	13440	6.27
3	HQ vs. N vs. C	F	amplitude	oddball type	1	12498	27.092
4	HQ vs. F vs. N	C	latency	oddball type	1	13440	25.429
5	HQ vs. F vs. N	C	latency	standard stimulus	2	13440	6.505
6	HQ vs. F vs. N	C	latency	channel	2	13440	7.921
7	HQ vs. N vs. C	F	latency	standard stimulus	2	12498	10.552
8	HQ vs. F vs. C	N	latency	oddball type	1	12429	12.999
9	HQ vs. F vs. C	N	latency	standard stimulus	2	12429	15.474
10	HQ vs. F vs. C	N	latency	oddball type $\times$ standard stimulus	2	12429	8.357

4. Discussion

4.1. Pretest

Results from the pretest demonstrated that applications of the degradation methods explicated in subsubsection 2.1.2 had indeed produced stimuli with lower quality compared to the HQ stimulus. Three stimuli were selected for the main experiment, which were quality-impaired on one perceptual dimension at a time (F, N, C) and equalized in terms of average perceived degradation intensity (see Figure 3).

4.2. Main experiment

Employing EEG in the context of an auditory oddball task, the main experiment investigated the influence of the perceptual dimensionality of speech quality on P300 characteristics. To this end, the experimental design of a preliminary study [33] was extended to control for potential confounding factors (individual loudness level, perceived degradation intensity, oddball probability; see subsection 1.3 and subsection 1.4). Statistically significant effects manifested on three levels of analysis (subjective, behavioral, electrophysiological) and will be discussed below with respect to the hypotheses originally derived in subsubsection 2.2.10.

4.2.1. Subjective results. Similar to the pretest, participants performed an ACR task to judge the quality of all stimuli presented in the (preceding or subsequent) oddball task. Due to the selection of three quality-impaired stimuli with equally strong perceived degradations (as ensured by the results of the pretest, see subsection 4.1) their MOS values should deviate only from the HQ stimulus, but not from each other. This prediction was confirmed by the subjective data depicted in Figure 3, thus replicating the results from the pretest and ruling out variation in perceived degradation intensity as an alternative explanation for the observed behavioral and electrophysiological differences between the quality-impaired stimuli.

4.2.2. Behavioral results. During the oddball task, participants actively responded to infrequent presentations of T, which required perceptual discrimination between successive stimuli serving as T and S. Analyzing only correct responses to T (“hits”), two behavioral measures (hit response time, hit rate) were calculated to quantify task performance. For healthy participants without self-reported hearing problems—out of which the present sample was exclusively composed—oddball task performance should on average be high, as indicated by fast hit response times and high hit rates. This assumption was confirmed with average hit response times and hit rates mostly falling into ranges of 250–350 ms (see Figure 4) and 90–100 % (see Figure 5), respectively.

On the other hand, increasing mental fatigue and/or drowsiness might have slowed down hit responses and reduced hit rates across blocks (time-on-task effects). Contrary to this assumption, no significant post-hoc pairwise comparisons of *block* were found

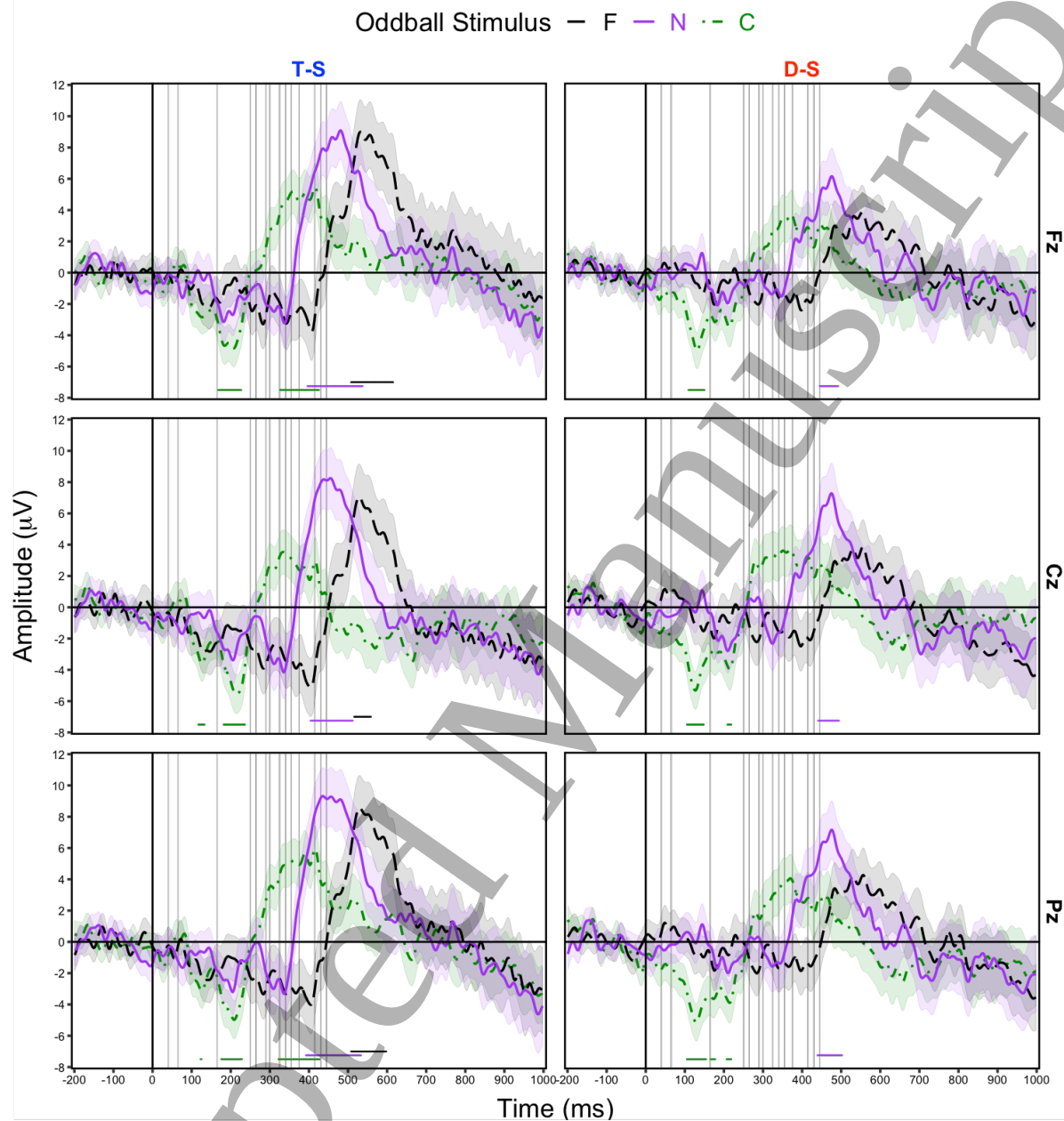


Figure 7: Grand average difference waveforms for each oddball type (T-S = left column, D-S = right column) and channel (Fz = top row, Cz = medium row, Pz = bottom row), showing variation in neural responses to *standard stimulus* (F, N, C) with HQ serving as constant oddball stimulus ($N = 28$). Vertical gray lines mark onsets of discrete frame losses during audio playback of the discontinuity-impaired stimulus (F). Lines of dots above time axis indicate, at which epoch time points each waveform significantly deviated from baseline. Error bands represent 95 % confidence intervals.

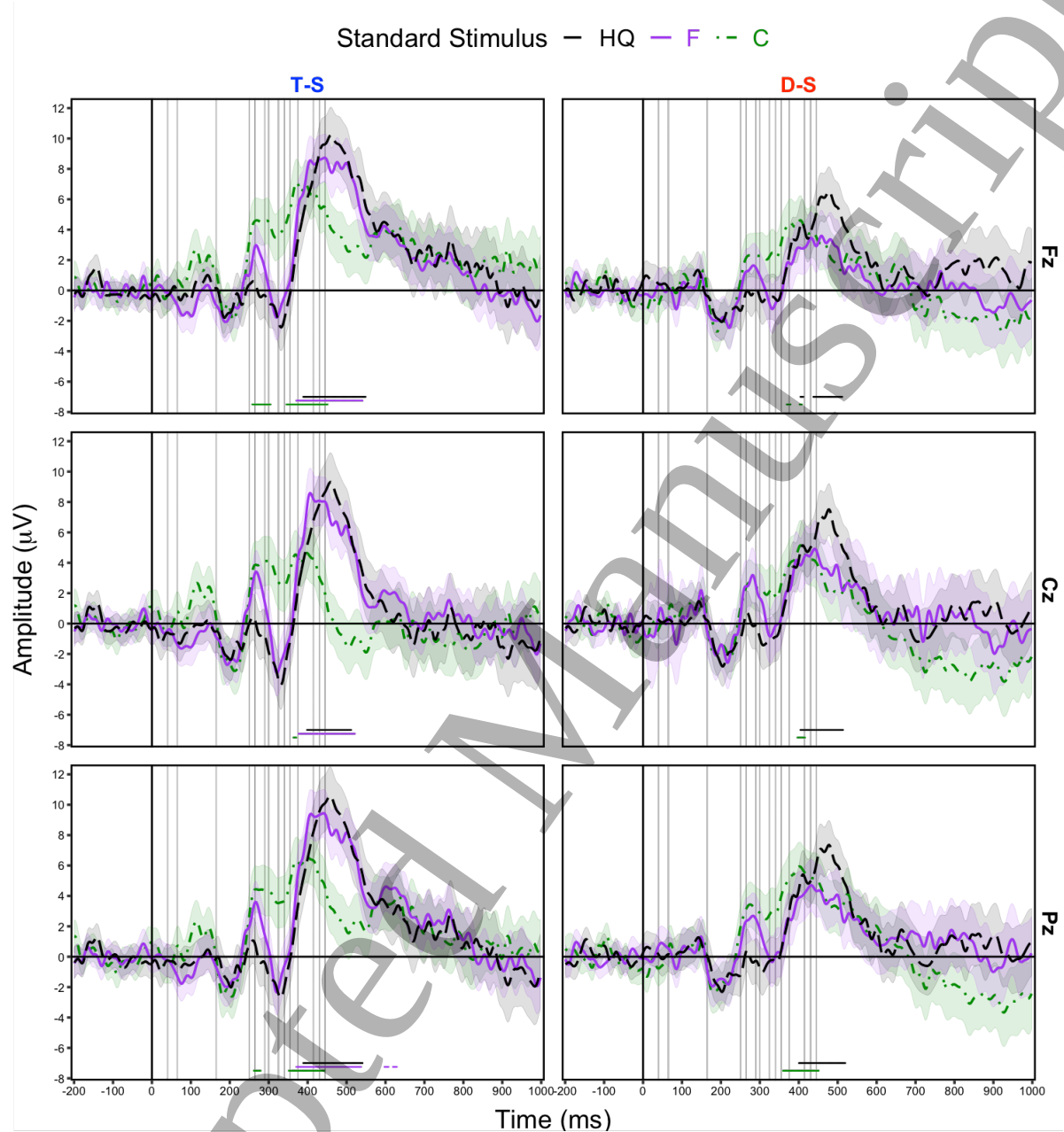


Figure 8: Grand average difference waveforms for each oddball type (T-S = left column, D-S = right column) and channel (Fz = top row, Cz = medium row, Pz = bottom row), showing variation in neural responses to *standard stimulus* (HQ, F, C) with N serving as constant oddball stimulus ($N = 28$). Vertical gray lines mark onsets of discrete frame losses during audio playback of the discontinuity-impaired stimulus (F). Lines of dots above time axis indicate, at which epoch time points each waveform significantly deviated from baseline. Error bands represent 95 % confidence intervals.

for either behavioral response measure. In general, average hit response times remained largely constant over the time course of the oddball task. The first block, however, was characterized by a visibly lower and more variable hit rate, which increased over the following blocks, eventually reaching a saturation point at about 97.5 % (see Figure 5). This trend seemed to reflect an initial habituation to general task demands, instead of more specific demands posed by individual stimulus sets.

Furthermore, task difficulty as determined by the discriminability between T and S presumably manifested in both behavioral response measures depending on *stimulus set*: High T-S discriminability should result in faster hit response times and increased hit rates, and vice versa. Pairwise comparisons between the twelve stimulus sets revealed that hit response times were highest for all stimulus sets in which F served as T (see Figure 4 and Table 4). With regard to hit rate, no significant post-hoc differences were found. Since F resulted from the multiplication of the source audio file with a RFE pattern, thus causing a number of discrete frame losses instead of a continuous degradation of the speech signal, it might have been more difficult for some participants to discriminate both stimulus types. On the other hand, hit response times were lowest when C was assigned to T, implying that participants were able to detect task-relevant variation in the frequency composition (C) of the speech signal much faster than variation in noise level (N). A possible explanation relies on the fact that the initial voiceless glottal fricative /h/ in the spoken word /Haus/ possesses only low acoustic energy. As change in signal-correlated noise depends on the speech signal amplitude, its sensory detectability in /h/ would have been reduced, thus entailing slower behavioral responses. By comparison, a change in spectral composition was better detectable in /h/ and therefore caused faster behavioral responses.

4.2.3. Electrophysiological results. Visual inspection of Figure 6, depicting grand average ERP waveforms (top row) and grand average difference waveforms (bottom row), allowed to identify distinctive P300 deflections over the epoch time course. It was confirmed that stimulus presentations during the oddball task had successfully evoked P3a and P3b components with typical temporal and morphological characteristics (see red and blue arrows in Figure 6). Despite that, grand average T-S and D-S difference waveforms drastically shifted in P300 latency depending on whether F or a continuously degraded stimulus (N, C) was assigned to either oddball type, which was attributed to a delayed detection of frame losses in the speech signal part of F (see vertical gray lines, which mark the onsets of discrete frame losses within the stimulus in Figure 6). For stimulus set HQ-F-N (see right column in Figure 6), temporal shifting of the T/T-S waveforms also comprised a reduction in amplitude. Presumably, attentional resources were mobilized after both stimulus onset and degradation onset. For continuous quality impairments (N/C) both onsets coincided, causing a summation of allocated attentional resources. However, for discontinuous quality impairments (F) both onsets were temporally separated, therefore this summation was less pronounced. Error bands indicate considerable variation in neural responses, which was partially

caused by constraints in the experimental design: Due to the necessity of testing each participant in twelve successive blocks as well as natural limitations in testing time and participants' ability to sustain attention, trial numbers per block were set to either 100 or 200—thus leaving a total trial ratio of either 70-15-15 or 140-30-30 for stimulus types S-T-D per block. As a consequence of those comparatively small trial numbers, signal-to-noise ratios of averaged neural responses for single participants were relatively low, which in turn increased variability both within and between participants.

Generally, larger event-related neural responses were observed after presentation of T versus D (see effects # 1, 3, 4, 6, 8, 9, 11, 12 listed in Table 5 and effects # 1, 3, 4, 8 listed in Table 6), which is in line with previous applications of the three-stimulus oddball paradigm in the auditory modality [29]. Thus, a manipulation check was provided that participants had indeed followed task instructions. Effects on P300 parameters also consistently declined in strength across electrode positions Fz, Cz and Pz (see effects # 5, 7, 10 listed in Table 5 and effect # 6 listed in Table 6), which matches the typical fronto-central distribution of the P3a subcomponent and emphasizes the functional importance of attentional orienting during the present oddball task [4,24].

In subsection 2.2.10, three internal causes were postulated to underlie effects of *oddball stimulus* and *standard stimulus* on P300 peak parameters (see Figure 2):

- *Temporal degradation pattern*, i.e., delayed sensory processing, as indicated by increased P3a peak latencies for F versus N/C serving as either oddball type. See again hypotheses (i) and (iii).
- *Sensory oddball-standard deviation*, i.e., enhanced attentional resource mobilization for high versus low oddball-standard deviations, as indicated by increased P3a peak amplitudes for HQ versus F/N/C serving as S. See again hypothesis (ii).
- *Perceptual target-standard discriminability*, i.e., increased cognitive load and processing time for low versus high T-S discriminability, as reflected by increased P3b peak latencies and reduced P3b peak amplitudes for F/N/C versus HQ serving as S. See again hypotheses (iv) and (v).

Looking at Figure 7, a temporal separation of grand average difference waveforms became apparent for both oddball types, indicating an effect of *oddball stimulus* when HQ served as constant standard stimulus: Fastest neural responses were observed for C, followed by N and F serving as oddball—thus confirming hypotheses (i). Interestingly, also a visible difference in peak latencies between N and C was observed, suggesting temporal variation in sensory/perceptual processing of the two continuous degradations. Moreover, apparent differences in peak amplitudes might point towards variation in the amount of allocated attentional resources between the degraded stimuli, despite their prior equalization in terms of perceived degradation intensity.

ERP parameter analysis demonstrated effects of *oddball stimulus* primarily on P300 peak latency: Elicited neural responses were shown to be delayed for F serving as oddball stimulus (see effects # 9 and 12, listed in Table 5 with corresponding post-hoc comparisons listed in Table A1)—hence confirming hypothesis (i). In addition,

also differences in the timing and size of neural responses between the two continuous degradations were observed (# 2, 7, 9). Some of these neural modulations only manifested for a particular *channel* or *oddball type*: First, when F served as standard stimulus (# 7), neural responses were fastest at Fz compared to Cz and Pz, again highlighting the role of attentional resource allocation for detecting temporal quality impairments in oddball stimuli. Second, when N served as standard stimulus (# 12), neural responses to F were prolonged only if the oddball stimulus had been categorized as task-relevant (T-S). This suggested that temporal quality features of oddball stimuli were only taken into account during task-related cognitive processing. Third, an effect on amplitude (# 2) occurred only if N/C had been categorized as task-irrelevant (D-S). Apparently, more attentional resources were allocated towards sensory and perceptual processing of N versus C. Since the perceived degradation intensity of both stimuli had been equalized beforehand, this finding might otherwise be explained by a compensatory effort from the attentional system, which automatically reacts to increased noise levels in the sensory signal, but not as much to changes in its spectral composition [58, 59].

Grand average difference waveforms depicted in Figure 8 show effects of *standard stimulus* in case N served as constant oddball stimulus, whose detection was delayed after repeated presentation of F—hence confirming hypothesis (iii).

A corresponding result pattern emerged from ERP parameter analysis (see Table 5): P300 peak latencies following presentations of C (see effect # 5, listed in Table 6 with the corresponding post-hoc comparison listed in Table A2) or N (# 10) as constant oddball stimulus were increased for F assigned to standard stimulus—again confirming hypothesis (iii). Besides, neural responses elicited by F occurring as oddball stimulus were faster when C versus N served as standard stimulus (# 7). It might again be assumed that after repeated presentations of N, a certain amount of attentional resources were still bound to automatically compensate the increased noise level, which would leave less attentional resources available for immediate processing of F. Effects of *standard stimulus* were also shown to be dependent on oddball type (# 10): Neural responses to N as oddball stimulus were slowed down after repeated presentation of F serving as standard stimulus, yet only if N had been categorized as task-irrelevant (D-S). Presumably, in this case no task-related mobilization of attentional resources ensued, leaving only the slowed neural response associated with delayed sensory processing. Contrary to hypothesis (ii), oddball-standard deviations were not notably larger when HQ versus quality-impaired stimuli served as standard stimulus. The same conclusion could be drawn with regard to T-S discriminability, since peak amplitude of HQ was not larger compared to the quality-impaired stimuli—thus rejecting hypotheses (iv). The only significant effect on P300 amplitude seemed to suggest a reduction in cognitive load if F versus N was presented as standard stimulus, from which C serving as oddball stimulus had to be actively distinguished (# 2, for T-S): Accordingly, perceptual discrimination between discontinuous (F) and continuous (C) quality impairments was easier than between the two continuous types of impairments (N vs. C), which could be attributed to differences in temporal degradation patterns. The absence of other

significant amplitude modulations provided support for a successful equalization of perceived degradation intensities on the electrophysiological level of analysis, especially considering that significant differences in peak amplitude between HQ and quality-impaired stimuli had originally been reported by the preliminary study [33]. Lastly, also no significant latency modulations following presentations of T were found (# 10, for T-S), which contradicted hypothesis (v).

To sum up, experimental evaluation of initial hypotheses derived in subsubsection 2.2.10 led to the following conclusions:

- (i) Effects of *oddball stimulus* on P3a peak latency ($F > N/C$) were confirmed by temporal shifts of ERP waveforms (see Figure 7) and corresponding P300 latency modulations (see effects # 9 and 12 listed in Table 5 and post-hoc comparisons listed in Table A1).
- (ii) Effects of *standard stimulus* on P3a peak amplitude ($HQ > F/N/C$) could not be confirmed by the present electrophysiological data.
- (iii) Effects of *standard stimulus* on P3a peak latency ($F > N/C$) were confirmed by temporal shifts of ERP waveforms (see Figure 8) and corresponding P300 latency modulations (see effects # 5 and 10 listed in Table 6 and post-hoc comparisons listed in Table A2).
- (iv) Effects of *standard stimulus* on P3b peak amplitude ($HQ > F/N/C$) could not be confirmed by the present electrophysiological data.
- (v) Effects of *standard stimulus* on P3b peak latency ($HQ < F/N/C$) could not be confirmed by the present electrophysiological data.

5. Conclusion and Future outlook

The present study consisted of the generation of degraded word stimuli with equal MOS values (pretest) to validly examine the impact of independent speech quality dimensions (“discontinuity”, “noisiness”, “coloration”) on P300 characteristics (main experiment). Hypotheses were deduced from a revised conceptual framework describing the functional roles of P3a and P3b subcomponents in quality perception (see Figure 2). Building upon the experimental design of a preliminary study, P300-related neural responses were triggered in the context of an active three-stimulus oddball task, which involved presentations of behaviorally relevant (T = target) and irrelevant (S = standard, D = distractor) stimulus types. The rationale of the main experiment was based on the assumption that if perceived degradation intensities were held constant, modulations in temporal and morphological P300 characteristics would only be attributable to changes in the particular nature of the induced degradations, i.e., the underlying quality dimensions (F = discontinuity-impaired, N = noisiness-impaired and C = coloration-impaired). Analysis of electrophysiological data revealed the following primary results:

First, the absence of effects of oddball-standard deviation and T-S discriminability on P300 amplitude provided further support that perceived degradation intensities of

the three quality-impaired stimuli had been successfully equalized. As a consequence, observed neural modulations could be more validly ascribed to changes in the characteristic nature of quality degradations.

Second, a basic distinction could be drawn between a “discontinuous” (F) and “continuous” (N, C) character of speech degradations, which was detectable either from the beginning of the speech signal or only from the first impaired sample onward. Corresponding effects were observed both on the behavioral (hit response times) and electrophysiological level of analysis (P300 peak parameters). In addition, average ERP waveforms and P300 peak parameters also revealed possible dissociations between the two continuous degradations (N vs. C) given the perceptual contexts of specific stimulus sets.

Third, the sensitivity of the present neurophysiological indicators to changes in perceived quality features often dependent on whether the oddball stimulus had been categorized as behaviorally relevant (T) or irrelevant (D). Furthermore, neural modulations often occurred over frontal (Fz) and central (Cz) electrode positions, most likely reflecting P3a-related activity. Such interaction effects highlight the importance of both behavioral task requirements (*task relevance*) and topographic location within the analyzed P300 scalp distribution (*channel*) as potential influencing factors, which should therefore be taken into consideration when conducting EEG-based quality assessments.

Several improvements and extensions with regard to the present experimental design are still left to be implemented by future studies:

First, instead of predicating the equalization of perceived degradation intensities on ACR ratings averaged across a large sample of participants (as realized in the pretest), appropriate test stimuli might be selected individually for each participant depending on his/her quality rating scores. Consequently, additional single-subject and/or single-trial analyses could be performed on the electrophysiological data by following the procedure of an earlier study [13].

Second, another more fundamental confounding factor has not been addressed so far: It could still be argued that the neural modulations observed in the present study (and even earlier studies) were in fact not related to changes in quality dimension, but resulting from physical changes entirely unrelated to quality, e.g., variation in stimulus content or even change in acoustic stimulation per se. To counter this alternative explanation within the present experimental design, portions reflecting mere physical changes would have to be subtracted from the evoked neural signal, leaving only portions that truly reflect quality changes. The former could be achieved, e.g., by presenting different high-quality (HQ) words as standard and oddball stimuli, thus causing a physical deviation unrelated to quality.

Third, to ensure equal loudness across participants, more detailed audiometric data could be collected and included in the analysis (e.g., individual hearing thresholds). Assessment of word intelligibility over the course of the experiment further permits the examination of another implicit assumption made by the present study, namely that the quality manipulations would induce only superficial impairments on the level of the

physical speech signal, but not on the level of semantic content.

Finally, future studies might prioritize the comparison of the two continuous quality dimensions, N and C, to avoid drastic temporal shifts of ERP waveforms and enable ERP parameter extraction with two separate time windows for P3a and P3b like in the preliminary study [33]. Also, HQ stimuli should be assigned to both oddball types (T, D) in order to achieve a full factorial design and allow for additional stimulus set comparisons.

Acknowledgments

We thank the anonymous reviewers for their feedback, which has helped to improve the quality of the present article. We thank Elvira Ibragimova for her help in collecting data during the pretest.

This work was supported by the strategic partnership program between Technische Universität Berlin, Germany, and the Norwegian University of Science and Technology, Trondheim, Norway.

Authors' contributions

SU, SM and JNVA conceptualized the study. SU and GM generated the stimuli. SU collected the data, ran the analyses, and prepared a first draft of the article with SM and JNVA. All authors read, revised and approved the article.

Appendix

Table A1: Post-hoc linear contrasts for effects of *oddball stimulus* on P300 peak parameters. Asterisks indicate statistical significance. Bold subtable indices point to correspondingly numbered effect indices (#) in Table 5.

(2)

	<i>Contrast</i>	<i>Estimate</i>	<i>SEM</i>	<i>t</i>	<i>p</i>
T-S	C - N	0.438	0.853	0.514	0.607
D-S	C - N	-3.227	0.814	-3.963	< 0.001*

(5)

	<i>Contrast</i>	<i>Estimate</i>	<i>SEM</i>	<i>t</i>	<i>p</i>
Fz	D-S - T-S	28.009	8.981	3.119	0.002*
Cz	D-S - T-S	20.083	8.981	2.236	0.025*
Pz	D-S - T-S	-14.594	8.981	-1.625	0.104

(7)

	<i>Contrast</i>	<i>Estimate</i>	<i>SEM</i>	<i>t</i>	<i>p</i>
Fz	C - N	22.925	8.743	2.622	0.009*
Cz	C - N	-16.855	8.743	-1.928	0.054
Pz	C - N	12.009	8.743	1.374	0.17

(9)

	<i>Contrast</i>	<i>Estimate</i>	<i>SEM</i>	<i>t</i>	<i>p</i>
	C - F	-27.133	5.019	-5.406	< 0.001*
	C - N	-12.887	4.960	-2.598	0.01*
	F - N	14.247	5.083	2.803	0.01*

(10)

	<i>Contrast</i>	<i>Estimate</i>	<i>SEM</i>	<i>t</i>	<i>p</i>
	Cz - Fz	19.904	4.989	3.990	< 0.001*
	Cz - Pz	0.036	4.989	0.007	0.994
	Fz - Pz	-19.869	4.989	-3.983	< 0.001*

(12)

	<i>Contrast</i>	<i>Estimate</i>	<i>SEM</i>	<i>t</i>	<i>p</i>
T-S	C - F	-33.61	7.27	-4.624	< 0.001*
D-S	C - F	-2.119	6.734	-0.315	0.753

Table A2: Post-hoc linear contrasts for effects of *standard stimulus* on P300 peak parameters. Asterisks indicate statistical significance. Bold subtable indices point to correspondingly numbered effect indices (#) in Table 6.

(2)						(5)					
	<i>Contrast</i>	<i>Estimate</i>	<i>SEM</i>	<i>t</i>	<i>p</i>		<i>Contrast</i>	<i>Estimate</i>	<i>SEM</i>	<i>t</i>	<i>p</i>
	T-S F - HQ	2.207	0.77	2.867	0.008*		F - HQ	17.123	5.045	3.394	0.002*
	T-S F - N	3.182	0.774	4.109	< 0.001*		F - N	3.387	5.068	0.668	0.5
	T-S HQ - N	0.975	0.778	1.253	0.21		HQ - N	-13.737	5.068	-2.71	0.013*
	D-S F - HQ	0.2	0.735	0.271	0.837						
	D-S F - N	-0.595	0.735	-0.809	0.837						
	D-S HQ - N	-0.794	0.733	-1.084	0.836						
(6)						(7)					
	<i>Contrast</i>	<i>Estimate</i>	<i>SEM</i>	<i>t</i>	<i>p</i>		<i>Contrast</i>	<i>Estimate</i>	<i>SEM</i>	<i>t</i>	<i>p</i>
	Cz - Fz	16.385	5.026	3.26	0.002*		C - HQ	-17.408	5.101	-3.413	0.001*
	Cz - Pz	-2.295	5.026	-0.457	0.648		C - N	-22.287	5.028	-4.433	< 0.001*
	Fz - Pz	-18.68	5.026	-3.717	< 0.001*		HQ - N	-4.879	5.061	-0.964	0.335
(10)											
	<i>Contrast</i>	<i>Estimate</i>	<i>SEM</i>	<i>t</i>	<i>p</i>						
	T-S C - F	-9.283	7.483	-1.241	0.644						
	T-S C - HQ	-5.179	7.486	-0.692	0.978						
	T-S F - HQ	4.105	7.376	0.557	0.978						
	D-S C - F	-29.284	7.017	-4.173	< 0.001*						
	D-S C - HQ	-47.022	6.989	-6.728	< 0.001*						
	D-S F - HQ	-17.737	6.992	-2.537	0.011*						

References

[1] S. Möller and A. Raake. Quality of Experience: Terminology, Methods and Applications. *PIK - Praxis der Informationsverarbeitung und Kommunikation*, 37(4), 2014.

[2] A. Raake and S. Egger. Quality and Quality of Experience. In S. Möller and A. Raake, editors, *Quality of Experience*, pages 11–33. Springer International Publishing, Cham, 2014.

[3] U. Jekosch. *Voice and Speech Quality Perception: Assessment and Evaluation*. Signals and Communication Technology. Springer, Berlin, Germany, 2005.

[4] J. Polich. Updating P300: An integrative theory of P3a and P3b. *Clinical Neurophysiology*, 118(10):2128–2148, 2007.

[5] ITU-T Recommendation P.800. *Methods for subjective determination of transmission quality*. International Telecommunication Union, Geneva, Switzerland, 1996.

[6] U. Reiter, K. Brunnström, K. De Moor, M.-C. Larabi, M. Pereira, A. Pinheiro, J. You, and A. Zgank. Factors Influencing Quality of Experience. In S. Möller and A. Raake, editors, *Quality of Experience*, pages 55–72. Springer International Publishing, Cham, 2014.

[7] U. Engelke, D. P. Darcy, G. H. Mulliken, S. Bosse, M. G. Martini, S. Arndt, J.-N. Antons, K. Y. Chan, N. Ramzan, and K. Brunnström. Psychophysiology-Based QoE Assessment: A Survey. *IEEE Journal of Selected Topics in Signal Processing*, 11(1):6–21, 2017.

- [8] D. A. Pizzagalli. Electroencephalography and High-Density Electrophysiological Source Localization. In J. T. Cacioppo, L. G. Tassinary, and G. Berntson, editors, *Handbook of Psychophysiology*, pages 56–84. Cambridge University Press, Cambridge, 3 edition, 2007.
- [9] C. C. Duncan, R. J. Barry, J. F. Connolly, C. Fischer, P. T. Michie, R. Näätänen, J. Polich, I. Reinvang, and C. V. Petten. Event-related potentials in clinical research: Guidelines for eliciting, recording, and quantifying mismatch negativity, P300, and N400. *Clinical Neurophysiology*, 120(11):1883–1908, 2009.
- [10] 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 *Proceeding of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 2690–2693. IEEE, 2010.
- [11] J.-N. Antons, A. K. Porbadnigk, R. Schleicher, B. Blankertz, S. Möller, and G. Curio. Subjective Listening Tests and Neural Correlates of Speech Degradation in Case of Signal-correlated Noise. In *Audio Engineering Society (AES) 129th Convention*, San Francisco, CA, USA, 2010. Curran Associates, Inc.
- [12] J.-N. Antons, R. Schleicher, S. Arndt, S. Möller, A. K. Porbadnigk, and G. Curio. Analyzing Speech Quality Perception Using Electroencephalography. *IEEE Journal of Selected Topics in Signal Processing*, 6(6):721–731, 2012.
- [13] A. K. Porbadnigk, M. S. Treder, B. Blankertz, J.-N. Antons, R. Schleicher, S. Möller, G. Curio, and K.-R. Müller. Single-trial analysis of the neural correlates of speech quality perception. *Journal of Neural Engineering*, 10(5):056003, 2013.
- [14] J.-N. Antons. *Neural Correlates of Quality Perception for Complex Speech Signals*. T-Labs Series in Telecommunication Services. Springer International Publishing, Cham, 2015.
- [15] 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. *Neuroscience Letters*, 500:e26–e27, 2011.
- [16] J.-N. Antons, K. u. R. Laghari, S. Arndt, R. Schleicher, S. Möller, D. O’Shaughnessy, and T. H. Falk. Cognitive, affective, and experience correlates of speech quality perception in complex listening conditions. pages 3672–3676. IEEE, 2013.
- [17] L. Lindemann and M. Magnor. Assessing the quality of compressed images using EEG. pages 3109–3112. IEEE, 2011.
- [18] S. Scholler, S. Bosse, M. S. Treder, B. Blankertz, G. Curio, K.-R. Müller, and T. Wiegand. Toward a Direct Measure of Video Quality Perception Using EEG. *IEEE Transactions on Image Processing*, 21(5):2619–2629, 2012.
- [19] S. Arndt, J.-N. Antons, R. Schleicher, S. Möller, and G. Curio. Using Electroencephalography to Measure Perceived Video Quality. *IEEE Journal of Selected Topics in Signal Processing*, 8(3):366–376, 2014.
- [20] L. Acqualagna, S. Bosse, A. K. Porbadnigk, G. Curio, K.-R. Müller, T. Wiegand, and B. Blankertz. EEG-based classification of video quality perception using steady state visual evoked potentials (SSVEPs). *Journal of Neural Engineering*, 12(2):026012, 2015.
- [21] S. Arndt. *Neural Correlates of Quality During Perception of Audiovisual Stimuli*. T-Labs Series in Telecommunication Services. Springer Singapore, Singapore, 2016.
- [22] S. Bosse, L. Acqualagna, W. Samek, A. K. Porbadnigk, G. Curio, B. Blankertz, K.-R. Müller, and T. Wiegand. Assessing Perceived Image Quality Using Steady-State Visual Evoked Potentials and Spatio-Spectral Decomposition. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2017.
- [23] J. Polich. *Neuropsychology of P300*. Oxford University Press, 2012.
- [24] D. Friedman, Y. M. Cycowicz, and H. Gaeta. The novelty P3: an event-related brain potential (ERP) sign of the brain’s evaluation of novelty. *Neuroscience & Biobehavioral Reviews*, 25(4):355–373, 2001.
- [25] E. N. Sokolov, editor. *The orienting response in information processing*. L. Erlbaum Associates,

- Mahwah, N.J., 2002.
- [26] H. Gaeta, D. Friedman, and G. Hunt. Stimulus characteristics and task category dissociate the anterior and posterior aspects of the novelty P3. *Psychophysiology*, 40(2):198–208, 2003.
 - [27] A. Kok. On the utility of P3 amplitude as a measure of processing capacity. *Psychophysiology*, 38(3):557–577, 2001.
 - [28] E. Courchesne, S. A. Hillyard, and R. Galambos. Stimulus novelty, task relevance and the visual evoked potential in man. *Electroencephalography and Clinical Neurophysiology*, 39(2):131–143, 1975.
 - [29] J. Katayama and J. Polich. Auditory and visual P300 topography from a 3 stimulus paradigm. *Clinical Neurophysiology*, 110(3):463–468, 1999.
 - [30] M. Wältermann, A. Raake, and S. Möller. Quality Dimensions of Narrowband and Wideband Speech Transmission. *Acta Acustica united with Acustica*, 96(6):1090–1103, 2010.
 - [31] M. Wältermann, A. Raake, and S. Möller. Direct Quantification of Latent Speech Quality Dimensions. *Journal of the Audio Engineering Society*, 60(4):246–254, 2012.
 - [32] M. Wältermann. *Dimension-based Quality Modeling of Transmitted Speech*. T-Labs Series in Telecommunication Services. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.
 - [33] S. Uhrig, S. Arndt, S. Möller, and J.-N. Voigt-Antons. Perceptual references for independent dimensions of speech quality as measured by electroencephalography. *Quality and User Experience*, 2(1), 2017.
 - [34] A. C. Papanicolaou, D. W. Loring, N. Raz, and H. M. Eisenberg. Relationship Between Stimulus Intensity and the P300. *Psychophysiology*, 22(3):326–329, 1985.
 - [35] M. Lindín, M. Zurrón, and F. Díaz. Stimulus intensity effects on P300 amplitude across repetitions of a standard auditory oddball task. *Biological Psychology*, 69(3):375–385, 2005.
 - [36] B. J. McDermott. Multidimensional Analyses of Circuit Quality Judgments. *The Journal of the Acoustical Society of America*, 45(3):774–781, 1969.
 - [37] J. Katayama and J. Polich. P300, probability, and the three-tone paradigm. *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section*, 100(6):555–562, 1996.
 - [38] J. Polich. Attention, probability, and task demands as determinants of P300 latency from auditory stimuli. *Electroencephalography and Clinical Neurophysiology*, 63(3):251–259, 1986.
 - [39] J. Polich. Task difficulty, probability, and inter-stimulus interval as determinants of P300 from auditory stimuli. *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section*, 68(4):311–320, 1987.
 - [40] R. Verleger, P. Jaśkowski, and E. Wascher. Evidence for an Integrative Role of P3b in Linking Reaction to Perception. *Journal of Psychophysiology*, 19(3):165–181, 2005.
 - [41] R. Verleger. P3b: Towards some decision about memory. *Clinical Neurophysiology*, 119(4):968–970, 2008.
 - [42] S. Uhrig, S. Möller, and J.-N. Voigt-Antons. Dissociating perceptual quality dimensions of transmitted speech using electroencephalography. *IEEE*, 2018.
 - [43] ITU-T Recommendation P.810. *Modulated noise reference unit (MNRU)*. International Telecommunication Union, Geneva, Switzerland, 1996.
 - [44] ITU-T Recommendation G.191. *Software tools for speech and audio coding standardization*. International Telecommunication Union, Geneva, Switzerland, 2010.
 - [45] ITU-T Recommendation P.56. *Objective measurement of active speech level*. International Telecommunication Union, Geneva, Switzerland, 2011.
 - [46] S. Olejnik and J. Algina. Generalized Eta and Omega Squared Statistics: Measures of Effect Size for Some Common Research Designs. *Psychological Methods*, 8(4):434–447, 2003.
 - [47] J.-N. Antons, S. Arndt, R. Schleicher, and S. Möller. Brain Activity Correlates of Quality of Experience. In S. Möller and A. Raake, editors, *Quality of Experience*, pages 109–119. Springer International Publishing, Cham, 2014.
 - [48] M. R. Nuwer, G. Comi, R. Emerson, A. Fuglsang-Frederiksen, J. M. Guerit, H. Hinrichs, A. Ikeda, F. J. Luccas, and P. Rappelsburger. IFCN standards for digital recording of clinical

- EEG. International Federation of Clinical Neurophysiology. *Electroencephalography and clinical neurophysiology*, 106(3):259–261, 1998.
- [49] R. Johnson. On the neural generators of the P300 component of the event-related potential. *Psychophysiology*, 30(1):90–97, 1993.
- [50] M. Fabiani, G. Gratton, and K. D. Federmeier. Event-Related Brain Potentials: Methods, Theory, and Applications. In J. T. Cacioppo, L. G. Tassinary, and G. Berntson, editors, *Handbook of Psychophysiology*, pages 85–119. Cambridge University Press, Cambridge, 3 edition, 2007.
- [51] M. Sabri, D. A. Kareken, M. Dzemidzic, M. J. Lowe, and R. D. Melara. Neural correlates of auditory sensory memory and automatic change detection. *NeuroImage*, 21(1):69–74, 2004.
- [52] 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: Auditory processing that leads to conscious perception. *Psychophysiology*, 48(1):4–22, 2011.
- [53] J. Katayama and J. Polich. Stimulus context determines P3a and P3b. *Psychophysiology*, 35(1):23–33, 1998.
- [54] R. Verleger, N. Baur, M. F. Metzner, and K. Śmigajewicz. The hard oddball: Effects of difficult response selection on stimulus-related P3 and on response-related negative potentials: Oddball-P3 and S-R mapping. *Psychophysiology*, 51(11):1089–1100, 2014.
- [55] M. D. Comerchero and J. Polich. P3a, perceptual distinctiveness, and stimulus modality. *Cognitive Brain Research*, 7(1):41–48, 1998.
- [56] M. D. Comerchero and J. Polich. P3a and P3b from typical auditory and visual stimuli. *Clinical Neurophysiology*, 110(1):24–30, 1999.
- [57] R. Tibon and D. A. Levy. Striking a balance: analyzing unbalanced event-related potential data. *Frontiers in Psychology*, 6, 2015.
- [58] M. Wöstmann. *Neural dynamics of selective attention to speech in noise*. PhD Thesis, Leipzig, Univ., 2015.
- [59] M. Wöstmann, E. Schröger, and J. Obleser. Acoustic Detail Guides Attention Allocation in a Selective Listening Task. *Journal of Cognitive Neuroscience*, 27(5):988–1000, 2015.