Bibliographical reference

Contribution of Peak Events to Overall Loudness, André Fiebig and Roland Sottek, Acta Acustica **vol. 101** (Number 6), 2015, pp. 1116-1129

DOI

https://doi.org/10.3813/AAA.918905

Acta Acustica united with Acustica
Vol. 101 (2015) 1116 – 1129

DOI 10.3813/AAA.918905

Contribution of Peak Events to Overall Loudness

André Fiebig, Roland Sottek
HEAD acoustics GmbH, Eberstr. 30a, Herzogenrath, 52134, Germany. andre.fiebig@head-acoustics.de

Summary
The general topic of this work was to investigate whether humans apply unifying principles to form retrospective overall loudness assessments of noise episodes. In three within-subjects factorial design of experiments, the contribution of systematically varied peak and background magnitudes to the assessment of overall loudness was studied. It was observed that the loudness level of peak and background contributed significantly to the assessments of overall loudness. The complete absence of any interaction between peak and background level in the experiments suggests additivity. The experimental findings underline the relevance of the average of momentary perceptual levels to the overall assessment of the whole episode. This means that if participants are requested to judge the overall loudness of noises, then participants did not deliberately ignore certain parts of the presented noise episodes. This applies for the considered bounded episodes of duration of 10 s having a recognizable start and end. Factorial design is frequently criticized, because it draws participants' attention to the manipulated variables and provides strong clues to the participants about the experimenter's hypothesis probably resulting in demand characteristics. Therefore, further experiments were performed, where the noise stimuli from the factorial design of experiment were judged in different stimuli contexts. First, filler material was added to the stimuli set to obscure the aim of the study and to distract attention from the systematically manipulated features. In a further experiment additional overall sound assessments were requested besides the assessment of overall loudness. It was observed that the importance of peak and background magnitudes was similar over all experimental conditions. As a tendency, it can be stated that the introduced measures to obfuscate the study aim distracting from the manipulated features lead to a greater cognitive averaging of the momentary experiences and to a less importance of distinct peak events for the overall assessment of the whole sound episode.

PACS no. 43.66.Cb, 43.66.Lj

1. Introduction

Human beings experience their surrounding world unfolding over time through a stream of transient states that often vary from moment to moment in their intensity [1]. Those experiences are frequently the basis for decisions, choices and predictions about the future. To make a meaningful decision, it is inevitable to retrospectively summarize "infinite" or bounded episodes of experiences. Thus, the stream of momentary experiences must be converted or combined to a summary assessment of the entire experience profile. But how do human beings combine streams of momentary experiences into an overall impression or an overall assessment, respectively? Do humans base their decisions on some unifying principles or do they possess a variety of cognitive schemas, each of which can be evoked or suppressed by subtle contextual features? [2] In general, it seems that there is a broad consensus that a global judgment is determined primarily by the relevant properties of an episode [3]. But, is the process of deriving overall evaluations of experienced episodes based on a kind of temporal integration of momentary affect or is the process more linked to a weighted averaging of selected moments? [4] According to Schäfer *et al.* it is still hard to tell from recent studies if the remembrance and overall evaluation of a past experience rely on integrated moments, such as the average, or on distinct moments, such as peak or end [5]. Moreover, it appears likely that different mechanism and cognitive effects are triggered in dependence of stimulus duration and time of assessment. For example, for overall loudness assessments of short-term noises primacy effects are frequently observed [6], which mean that listeners assign higher weight to the level of the beginning of a sound than to its middle portion. Susini *et al.* observed a perceptual asymmetry between increasing and decreasing ramps for global loudness judgments obtained at the end of longer noise stimuli; the asymmetry was in favor of increasing ramps referring to a recency effect [7]. Comparably, Algom *et al.* observed memory and percept invariance in the visual domain possibly due to the flow of information from short-term to long-term memory [8]. Beyond doubt it is evident that top-down processes are involved in perception processes and affect how information is gathered, processed, and stored [9].

Received 05 May 2015,
accepted 15 October 2015.

In general, if perception of sound is investigated, frequently retrospective overall assessments of time-variant noises are requested. An overall assessment of a sound episode is a retrospective act, where a magnitude of sensation, perception or affective appraisal is (consciously) assigned to the past period of experienced sound. This is the standard case in listening experiments independent of method or object of investigation, and only sometimes instantaneous evaluations of sounds are requested [10].

In psychoacoustic experiments, where retrospective overall assessments of time-variant noises are collected, it is frequently observed that prominent portions of sound contribute to the overall impression stronger than less prominent portions [11], since it seems unlikely that a memory representation of every instance of longer episodes is fully available to form the overall judgment [12]. Accordingly, Kuwano *et al.* observed that the judgment of overall loudness is not the same as the simple average of instantaneous loudness judgments [11]. Schreiber and Kahneman presented evidence as well that people tend to use selected moments of extended experiences to form overall unpleasantness assessments of synthetic sound stimuli [13]. This notion goes beyond a simple parametric representation of a sound in terms of psychoacoustic variables. It seems conceivable that due to human cognitive processing specific events are of higher importance than others leading to models claiming that snapshot measures are more efficient indexes of momentary affect and its meaning to overall evaluations than any average of the time-series curve of instantaneous perception or real-time continuous judgments respectively [4].

The perception-adequate interpretation of loudness of time-variant noises for deriving an overall loudness value is already subject to investigation over decades. Natural sounds are usually varying over time [14] and listeners, if requested, have to combine the varying loudness levels into an overall loudness assessment. It is likely that the neglect of duration, which is frequently observed in other empirical contexts such as assessments of pain [15] or pleasantness of movies [4], applies for the assessment of overall loudness as well and that duration effects tend to be small [13]. Paulsen has shown that for steady sounds, such as highway or random broadband noises, the duration does not play a significant role with respect to the assessment of overall loudness, annoyance or unpleasantness [16]. According to Paulsen, this is true for durations long enough to build up an impression which is representative for the used original sounds, which means durations much longer than 1 s. In case of longer sequences (20 s, 50 s and 80 s) the sound assessments were not influenced by the duration of the noise stimuli possessing the same L_{Aeq} [16]. In contrast to it, Schreiber and Kahneman found a small effect of duration of noise stimuli on overall annoyance ratings [13]. In general, the concept of duration neglect – the relative insensitivity of retrospective overall evaluations to the duration of an episode – is broadly accepted for several contexts [17]; however, as pointed out by Ariely and Loewenstein in some cases duration can matter for judgment and decision making [18]. For short sounds due to temporal integration of loudness the duration of stimulus plays a significant role [19]. Moreover, as shown by Olsen, for specific signals, like synthetic up-ramps, global loudness ratings appear to change as a function of duration, even when the range of physical intensity change remains constant and end-level recency is controlled [20, 21].

Moreover, it was frequently observed that assessed overall loudness appears to be strongly influenced by loud single events. For example, Fastl has shown that judged overall loudness is considerably higher than the arithmetic mean of the loudness run over time due to an "overemphasis" of single loud events, which was consistently observed over different methods, such as category scaling, line scaling, and magnitude estimation [22]. In order to take into account the relevance of loud events on the perceived overall loudness Fastl proposed the indicator 4th percentile loudness N_4. In general, a percentile value on a scale of 100 indicates the percent of a distribution that is equal to or above it.[1] In case of the N_4 indicator, the loudness value of the loudness over time function of an episode is meant, which is equal or exceeded only 4% of the measurement time indicating more or less the level of loud events. This means that 96% of all loudness values of the considered episode are smaller than the N_4 loudness value. Comparably, Stemplinger found N_4 and N_5 to be valid indicators for perceived overall loudness of industrial and traffic noises [23, 24]. Sottek has recommended the use of N_{10} for statistically fluctuating loudness in order to predict overall loudness perception [25]. The German loudness standard for the calculation of loudness of time-variant sound, the DIN 45631/A1, recommends the use of N_5 as a representative single value for the perceived overall loudness [26]. Accordingly, the ISO 532-1 CD, as largely based on DIN 45631/A1, states that the statistical mean of time varying loudnesses leads, in general, to results that are too low in comparison to evaluated loudness, and the percentile loudness N_5 shall be given when stating the overall loudness perceived [27]. Moreover, in the ISO 532-1 standard it is mentioned that the N_5 indicator is not suited for impulses because the percentile loudness N_5 strongly depends on the measurement time. In contrast to it, Namba *et al.* proposed a loudness measure calculated by the mean energy level of third octave bands instead of percentile values of time-variant loudness [28]. Schlittenlacher *et al.* introduced a loudness indicator $LL(P)$, which is calculated from the loudness levels over time [29]. Thus, the $LL(P)$ loudness indicator differs from the loudness indicator proposed by Namba *et al.*, since they computed their energy-based loudness metric on the basis of sound pressure levels over time. Schlittenlacher *et al.* showed that for the presented 10 s long sound stimuli the $LL(P)$ was superior to the N_5 indicator and predicted better the

[1] Unfortunately, the interpretation of percentile values in psychoacoustics is inconsistent with the common definition and use of percentile values in statistics. However, since the use of percentile values in the described way is very common and established in psychoacoustics, the authors apply the psychoacoustics notion of percentile values in this paper.

observed overall loudness assessments. Friebe found that the arithmetic average of momentary loudness evaluations corresponds well with judged overall loudness in the context of different environmental noise sequences [30]. Ferguson *et al.* observed that the median and standard deviation of time-varying loudness values of musical pieces are related to emotional responses to a certain extent [31].

Rennies *et al.* observed that the accuracy of different loudness predictions varied in the context of speech and speech-like signals [32]. They found that the mean of the long-term loudness function calculated by the time varying loudness model according to Glasberg and Moore [33] was a better predictor of the loudness than the maximum of the short-term loudness function also determined by the time varying loudness model [33]. The short-term loudness is calculated by integrating the instantaneous loudness using a kind of automatic gain control with a short attack time constant (22 ms) and a longer release time constant (50 ms); the long-term loudness according to Glasberg and Moore is the result of a second temporal smoothing stage with an attack time constant of 99 ms and a release time constant of 2 s, applied to the short-term loudness. The long-term loudness tries to model memory effects [33].

Moreover, several cognitive effects, such as primacy or recency effects, were frequently observed in the context of overall loudness assessments. In general, a recency effect refers to a cognitive bias, where humans put greater weight on late information than on early information of an experienced episode. The primacy effect represents the opposite to the recency effect: the early information is of higher importance than information presented later on. For example, Höger *et al.* revealed the relevance of the recency effect for loudness assessments [34]. Susini *et al.* observed that the highest levels, their temporal positions and their duration of emergence are relevant for overall loudness of time-variant sounds using 1 kHz pure tones with varying levels [35]. Steffens and Guastavino observed that besides average of momentary judgments, the linear trend of the temporal experience plays a role for overall pleasantness ratings of environmental noises [36]. In the field of molecular psychophysics investigating temporal perceptual weights of stimulus components, a bowl-shaped pattern of the temporal perceptual weights is frequently observed (e.g. [37, 38]). Pedersen and Ellermeier observed that ten temporal segments of a stimulus sequence were not uniformly weighted and they found evidence for perceptual emphasis on onsets and offsets, whereas the recency effect was of weaker extent [39]. Moreover, Oberfeld and Plank observed a delayed primacy effect in case of stimuli, which were faded in [40]. Ponsot *et al.* investigated the loudness perception of 1 kHz pure tone stimuli consisting of 16 consecutive 125 ms stationary segments with levels drawn independently from a normal-truncated distribution and found a recency effect and a primacy effect for flat level profile sounds, which were both statistically significant [41]. Moreover, in case of increasing and decreasing level profile sounds Ponsot *et al.* observed a level dominance effect; the segments with the highest levels received greater attention independent from the ramp type [41].

All these studies provided evidence for complex rules of loudness perception and integration, which is not adequately described as a simple summation process [42]. Frequently, experimental results are incompatible with the notion of an automatic, accumulative integration process as hypothesized by most of the current loudness models [39]. Moreover, several studies reported on the influence of non-acoustical aspects on loudness judgments. For example, Laumann *et al.* showed that the perceived loudness of music depends on preference and they observed that in accordance with the respective preference level of the music stimuli, different loudness assessment strategies were applied [43]. They observed that along with decreasing preference of the presented music the difference between the means of overall judgments and averaged instantaneous judgments becomes larger. Menzel *et al.* showed that differences in estimates of loudness can occur depending on visual stimulation by images of different colored sports cars [44].

All in all, there is a substantial body of research on how humans combine components of bounded episodes with start and end to summarize their experiences, but as of yet studies failed in providing broadly acknowledged universal models and rules that govern retrospective overall (loudness) assessments of time-variant noise episodes [45]. The aim of the present study is to deepen the knowledge of cognitive stimulus integration in the context of overall loudness assessment.

2. Overview of experiments and aim of the study

The general aim of the study was to investigate which unifying principles are used to form retrospective overall loudness assessments of noise episodes consisting of a peak event and steady noise with lower magnitude before and after the peak event. In particular, the importance of distinct peak components and steady noise parts for overall loudness assessments were systematically investigated. The null hypotheses were that the assessments of overall loudness cannot be predicted by means of acoustic indicators related to i) the magnitude of peaks and ii) the magnitude of the background. These hypotheses were tested by means of different experiments. To study the susceptibility of retrospective overall loudness assessments based on potential unifying principles to experimental circumstances, three different experiments were performed in this study. All experiments used a within-subjects factorial design of experiment. The results provide insights regarding the components of a noise episode contributing to the overall loudness assessment and how people encode past experiences of loudness perception. Are specific segments of a sound ignored, when they were below a certain level [42], is only the prominent and salient part of an episode relevant to the overall loudness assessment [46] or does the

Table I. Overview of experiments and basic properties.

	Experiment 1	Experiment 2	Experiment 3
Sounds of interest (three peak levels crossed with three background levels)	3 × 3	3 × 3	3 × 3
Additional sounds presented for the purpose of distraction from the manipulated features	no	yes	yes
Type of evaluation	single attribute	single attribute	multiple attributes

overall loudness result from a kind of averaged loudness over time [29]?

In three within-subjects factorial design of experiments, the contribution of systematically varied peak and background magnitudes within a noise episode to the assessment of overall loudness was studied. In principle three white noises with low loudness levels (background factor) were factorial crossed with three white noises with higher loudness levels (peak factor). The resulting sounds possessed a 1 second long peak event in a 10 s long episode. The peak event was principally louder than the steady noise before and after the peak event. The factor *peak* and the factor *background* had both three levels resulting in nine sounds. The peak event was always in the middle of the sequence to control for primacy and recency effects. The duration of all stimuli was kept constant to control for potential duration effects. The null hypotheses were that first, the peak level has not any significant influence on overall loudness assessment and second, the background level does not play a significant role for overall loudness assessment.

The relative increase of the factor levels was deliberately designed to be equal according to the DIN 45631/A1. Since the German standard DIN 45631/A1 is the only current standard, which allows for the computation of loudness of arbitrary non-stationary, time-varying sounds, this computation method was used for the definition of stimuli magnitudes [26]. As pointed out by Genuit *et al.*, the use of time-variant loudness computation models is even for apparently steady noises required, such as white noise due to its stochastic variation [14]. In the following, all loudness calculations are based on the DIN 45631/A1. The loudness values of the peak and background signals were configured in a way that they increase stepwise by approximately 30% according to the DIN 45631/A1. By using other methods for the calculation of time-variant loudness, such as the Time Varying Loudness (TVL) model proposed by Glasberg and Moore [33] or the Chalupper and Fastl Dynamic Loudness Model (DLM) [47] very similar relative loudness differences are obtained.[2]

The nine systematically manipulated sounds were always assessed on the same unipolar 11-point category scale with respect to overall loudness, but in different experimental circumstances. In the first experiment only the nine systematically manipulated sounds were presented and assessed.

Since the principle of the systematically varied levels of the factors might be obvious to the participants in the first within-subjects design of experiment probably supporting demand effects, additional experiments were performed. According to Zizzo, filler questions or filler behavioral tasks can be employed, while not deceiving subjects as such, this may help in obfuscation [48]. Accordingly, Ariely *et al.* claimed as well that a factorial design of experiment is known to create some effects that they are intended to measure by reminding the participants of a consideration that they might otherwise have neglected [49]. Thus, by means of further experiments it was studied to what extent the apparentness of stimuli manipulation triggers the observed variances. Filler material in terms of further sound stimuli (called filler sounds) with different properties were included in the stimuli set and additional responses were elicited to obfuscate the experimental objectives and to distract participants from the manipulated features. By the presentation of twenty three additional natural and synthetic sounds, it was intended to make it difficult to recognize the way of the systematic manipulation of the stimuli and to identify cues providing information about what constitutes appropriate rating behavior. To sum up, in a second experiment all stimuli – sounds of interest and additional sounds – were assessed on the 11-point category scale applied also in the first experiment by a different group of participants.

In order to further change the potential impact of demand characteristics, a third experiment was performed. As stated by Ariely *et al.* within-subjects factorial design has powerful demand characteristics, which can be reduced by having participants' rate sequences that differed in several features. Thus, attention is not directed to only one of the features of the investigated sequences [49]. Therefore, in a third experiment the 32 sounds from the second experiment including the nine factorial manipulated samples of interest were also assessed, apart from overall loudness, with respect to overall unpleasantness, overall sharpness, and overall tonality. Due to different response elicitations, it was intended to obfuscate further any inference on which ones the experimenter is actually interested in (cf. [49]). Additionally, to reduce any special focus on the attribute of interest, the assessment of overall loudness was requested on the bottom of the presented four category scales. Potentially, requesting several overall assessments at the same time after a sound episode termi-

[2] In case of the TVL model a relative increase of the short-term loudness and long-term loudness of 31% is observed. The DLM predicts a 31% loudness increase of the factor levels as well.

Table II. Loudness according to DIN 45631/A1 of peak and background signals (P_i and B_i) estimated by means of loudness percentile values (N_5 for peak and N_{50} for background).

P_1[3]	P_2	P_3	B_1	B_2	B_3
9.7	12.7	17.0	5.4	7.2	9.6

nated can avoid to a certain extent a listening to the sound episodes in an analytical way; participants were probably forced to listen more holistically to the sound episode to be judged. Since the stimulus set in this experiment was identical with the second experiment, context effects due to range-frequency effects were expected to be small [50]. Table I displays the performed experiments and summarizes their differences: Three within-subjects factorial design of experiments were performed, where experimental conditions were changed to reduce potential demand characteristics.

3. Overall loudness assessment of synthetic sounds in factorial design of experiments with and without filler sounds

3.1. Method

3.1.1. Participants

Experiment 1: Sixteen participants (13 male, 3 female) took part in the first experiment. The age ranged from 22 to 45 years; the mean age was 33.1 years (standard deviation of 7.1 years). All test participants reported normal hearing.

Experiment 2: Nineteen participants (11 male, 8 female) participated in the second experiment. The age ranged from 24 to 47 years and the mean age was 31.8 years with a standard deviation of 9.2 years. All test participants reported normal hearing.

Experiment 3: Eighteen participants (11 male, 7 female) took part in the third experiment. The age ranged from 23 to 54 years and the mean age was 38.5 years with a standard deviation of 10.6 years. All test participants reported normal hearing.

No participant took part in more than one experiment. The collected data about age, self-reported experience in listening test participations, self-reported acoustic background knowledge and the judged level of difficulty of the evaluation task were collected in all experiments and analyzed. No link between personal data and loudness judgments was found.

3.1.2. Stimuli

The set of factorial designed noise samples judged with respect to overall loudness is displayed in Figure 1. All sounds were equally configured: The peak event lasts 1 s, i.e., 10% of the total duration of 10 s. The loudness of the noise before and after the peak event possesses the same loudness magnitude. Table II shows the loudness of the peak and background signals in terms of percentile values computed by the DIN 45631/A1 [26], which allows for calculating the loudness and loudness level of arbitrary non-stationary, time-varying sounds.

The loudness values of peak and background signals were configured in a way that they increase stepwise by approximately 30% according to the DIN 45631/A1. This means that a similar increase of magnitude of peak and background signals occurs. The noise samples were composed of white noises with varying bandwidth and center frequency (using band-pass filters with 'infinitely' steep spectral slopes) and varying sound pressure levels. The background noise, occurring before and after the peak event, was a random broadband noise with a constant power spectral density and a bandwidth of four critical bands ranging from 510 to 1080 Hz. The peak component was based on a narrow-band noise covering only one critical band ranging from 1080 to 1270 Hz. The different center frequency and bandwidth of the peak component should additionally make the peak event more distinct besides its higher loudness.[4] All sounds were presented diotically.

In the first experiment only the 3 × 3 factorial designed sounds were presented.

In the experiments 2 and 3, in addition to the nine factorial designed sounds, 23 additional synthetic and natural sounds were presented. All filler sounds lasted 10 s. The additional synthetic noises were generated according to the configuration principle of the investigated samples. A peak event occurred in the middle of the sound stimulus. However, the magnitudes of the background before and after the peak event differed from the factorial varied stimuli to distract the participants from the manipulated features. Moreover, the bandwidth, center frequency differed or an audible tone was added. The natural sounds were typical environmental noises, such as fountain, road traffic or marketplace noise. The natural sounds were included in the experiment 2 and 3 to further reduce the attention to the manipulated features. In total, 32 sounds were presented in experiment 2 and 3. An overview of the additional presented sounds and their loudness according to the DIN 45631/A1 is shown in Table III. The order of sounds was randomized in all experiments.

[3] Peak magnitudes were described in terms of the respective N_5 value of loudness according to DIN 45631/A1. For the lowest peak level and highest background level a slightly higher N_5 value is determined. This is due to the fact that the loudness of the background noise is close to loudness of the peak increasing slightly the N_5 value (see Figure 1). However, the magnitude of the peak component is not changed at all.

[4] Sottek observed that random broadband noises, which are steeply filtered at the edges, evoke a perception of strong tonal character for almost any bandwidth probably influencing the perception of loudness [59]. However, the test subjects did not mention after performing the overall loudness experiments in the interviews any disturbing tonality. However, a confounding influence of this effect on the experimental results cannot be ruled out.

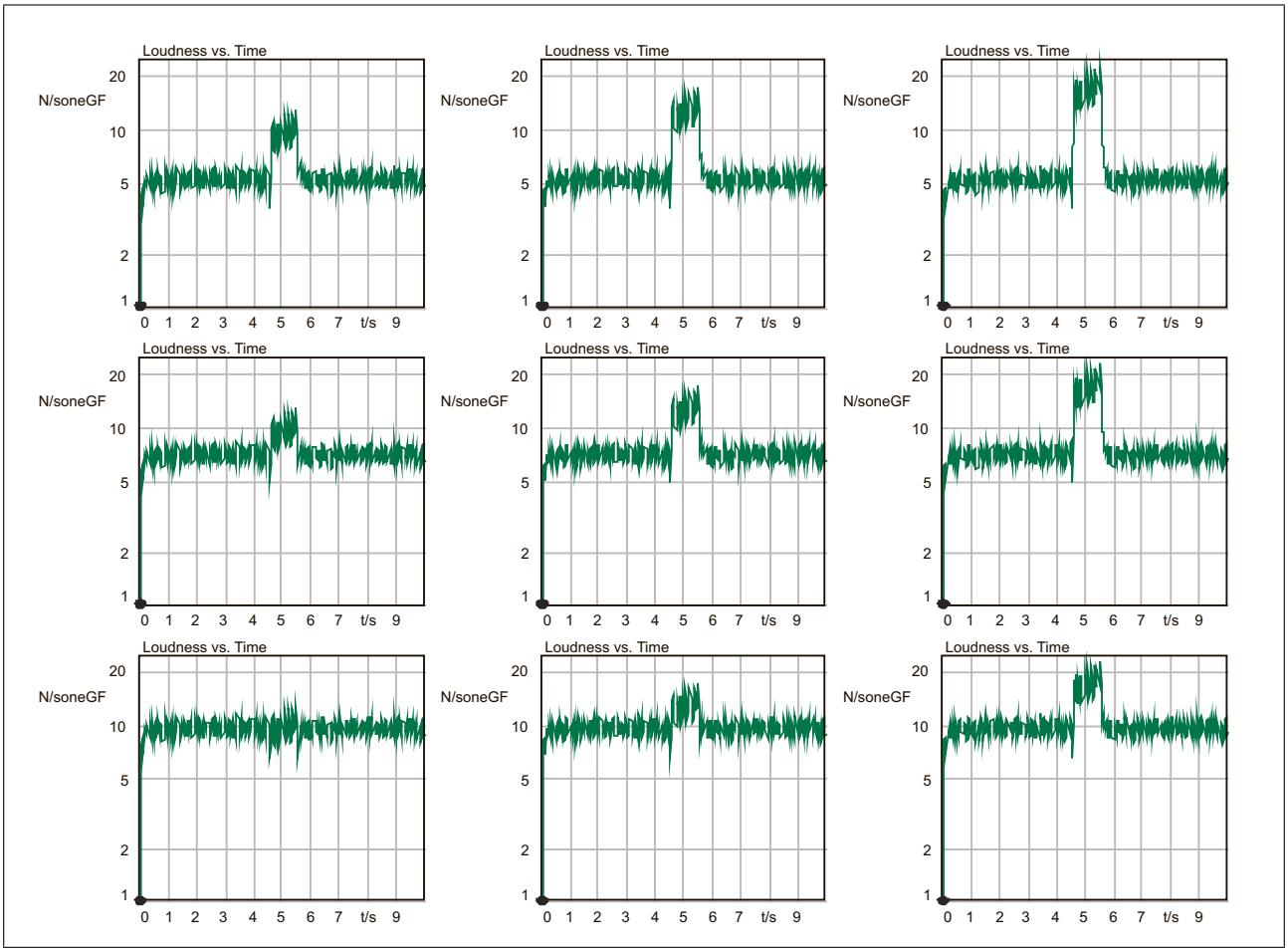

Figure 1. Loudness of noise samples over time according to DIN 45631/A1. Rows display sounds of constant background noise but changing magnitude of the middle sequence (peak). Columns show sounds of constant peak magnitude but changing background level.

3.1.3. Apparatus

The sounds were presented via digital equalizers (PEQ V, HEAD acoustics) and Sennheiser headphones HD 650 in an acoustically treated listening studio at the premises of the HEAD acoustics GmbH in Herzogenrath, Germany. The playback system was calibrated and equalized [51]. The instruction texts and interview questions were presented on computer screens. All sounds were played back with free-field equalization. The same apparatus was applied for all experiments.

3.1.4. Procedure

Experiment 1: The participants were not informed about the aim of the study and its hypotheses under test; it was explained that the experiment deals with the perception of noise in general. The overall loudness was measured by requesting category scale assessments. The experiment started with an instruction: *In the following you will listen to sounds with a duration of 10 s. Please judge the overall loudness after listening to the entire sound on an 11-point category scale, which ranges from "not at all loud" to "very loud"*. All categories were numbered. After the instruction, the participants listened to four sample stimuli to be familiar with the type of sounds to be presented in the experiment. The stimuli were randomly chosen from the stimuli set of the experiment. The demonstration of sounds, before the assessment of stimuli has started, was assumed to reduce scaling effects, such as floor and ceiling effects. The participants were requested to listen to the entire sound stimulus. During the playback of a sound, the input was blocked in order to force the participant to listen to the full sound stimulus before providing an assessment. Immediately after the completion of the sound stimulus, the participants could provide their rating of overall loudness via a touchscreen. It was not possible to repeat the playback of a sound stimulus. The participants could take as much time as needed to rate the stimulus; no time limit was given. After providing an assessment, the next sound stimulus was automatically played back. The duration of the first experiment with the nine systematically varied sounds was around 10 minutes. After the experiment, a short interview took place.

Experiment 2: The procedure was the same as in experiment 1, with one exception. In contrast to the first experiment, two randomly chosen synthetic and two randomly chosen natural sounds were played back before all sounds were presented for judgment. The duration of this experiment was due to the presentation of 32 sounds around 20 minutes.

Table III. Basic properties of sounds presented for distraction in addition to the nine sounds of interest. N_i [soneGF] according to DIN 45631/A1.

Sound description	Duration [s]	N_5	N_{50}	N_{90}
Small brook	10	9.2	7.4	5.7
River	10	12.2	10.9	10.0
Small fountain	10	8.4	6.2	5.1
Large fountain	10	11.6	11.2	10.7
Small waves	10	6.4	4.7	3.0
Ocean waves	10	12.8	10.4	8.8
Urban park	10	3.8	3.6	3.2
Road traffic noise	10	10.1	9.1	8.4
Marketplace	10	9.0	5.8	5.2
Car passing-by	10	12.4	2.5	1.4
Motorbike passing-by	10	24.6	4.2	2.2
Steam engine – Version 1	10	13.7	11.1	10.3
Steam engine – Version 2	10	9.1	7.2	6.7
White noise, where background loudness differs before and after peak – Version 1	10	12.7	6.6	5.1
White noise, where background loudness differs before and after peak – Version 2	10	17.0	8.9	5.1
White noise, where background loudness differs before and after peak – Version 3	10	12.7	6.6	5.1
White noise, where background loudness differs before and after peak – Version 4	10	17.0	8.9	5.1
Bandpass-filtered white noise with time-varying filter parameters (center frequency and bandwidth) – Version 1	10	11.0	10.0	9.2
Bandpass-filtered white noise with time-varying filter parameters (center frequency and bandwidth) – Version 2	10	11.0	10.0	9.3
Bandpass-filtered white noise with time-varying filter parameters (center frequency and bandwidth) – Version 3	10	11.0	10.1	9.5
Sum of white noise and a tone with varying level and frequency – Version 1	10	8.7	7.2	6.7
Sum of white noise and a tone with varying level and frequency – Version 2	10	9.9	9.4	9.0
Sum of white noise and a tone with varying level and frequency – Version 3	10	11.3	9.5	9.0

Experiment 3: The procedure was the same as in experiment 1 and 2, with one exception. All 32 sounds from the second experiment were judged by a kind of semantic differential technique requesting evaluations of multiple attributes. Besides the overall loudness, the participants were requested to assess overall sharpness, overall annoyance and overall tonality on unipolar 11-point category scales all ranging from not at all to very.[5] The order of judging the different attributes was not defined or controlled. It was free to the participant to choose the order of evaluating the sound on the four category scales. The duration of this experiment was around 25 minutes.

After the experiments, the participants were invited to express feelings and thoughts with respect to the procedure, the different scaling tasks and the degree of complexity of evaluation tasks. No participant reported any feelings of fatigue, exhaustion or overload.

3.2. Results

3.2.1. Experiment 1 - Overall loudness assessment in single category scaling experiment without filler sounds

By means of an ANOVA for two-factor repeated measures design with repeated measures on both factors it was found that both factors contribute statistically significant to the assessment of overall loudness for the considered synthetic noise stimuli. The background loudness ($F_{\text{background}}(2, 30) = 31.76$, $p < 0.01^{**}$) plays obviously a relevant role in constructing an overall loudness assessment and this factor was found to be highly statistically significant. Since the Mauchly's sphericity test indicated unequal group variances related to the peak factor, the Greenhouse-Geisser correction was applied and the degrees of freedom adjusted. The peak factor was highly statistically significant as well ($F_{\text{peak}}(1.19, 17.9) = 46.3$, $p < 0.01^{**}$, $\varepsilon_{GG} = 0.59$). It was verified on the basis of the analysis of skewness and kurtosis that the data was normally distributed.[6] The highly significant result means that at least two means of the three factor levels show a highly statistically significant difference, which mean that the null hypothesis must be rejected. An interaction effect was not observed.

Figure 2 shows the results as treatment mean plots of the two-factor experiment with repeated measures on both factors regarding the assessment of overall loudness. It can be clearly seen that both factors – level of peak as well as

[5] The applied method using 11-point unipolar category scales does not correspond to a classical semantic differential method. Usually semantic differential procedures use numerous bipolar category scales, where direction and intensity are indicated by means of verbal and numerical labels [60].

[6] Measures of skewness and kurtosis are frequently used to measure the extent of non-normality, where skewness measures the departure from symmetry and kurtosis the thickness of the tails of the distribution [61]. In all experiments, the assumption of normally distributed data was not seriously violated.

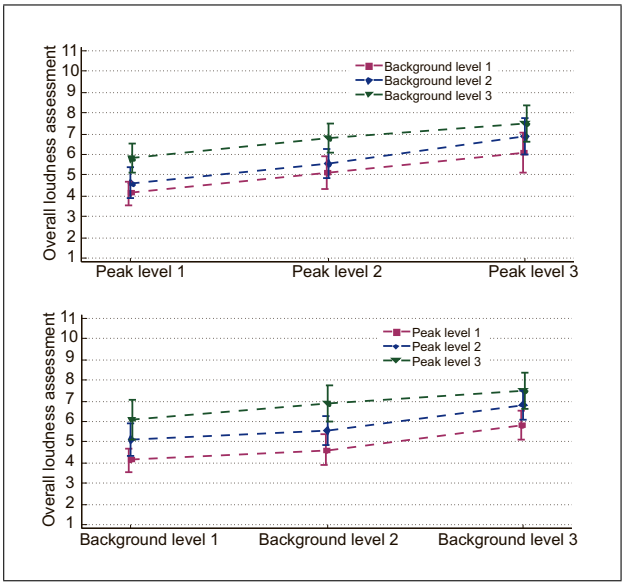

Figure 2. Treatment mean plots of Experiment 1: Relationship between overall loudness assessments of synthetic sounds and peak and background levels in a single category scaling experiment. The category scale ranges from not at all (1) to very (11). The arithmetic means and 95% confidence intervals are displayed. The corresponding loudness values of all peak and background levels are shown in Table II.

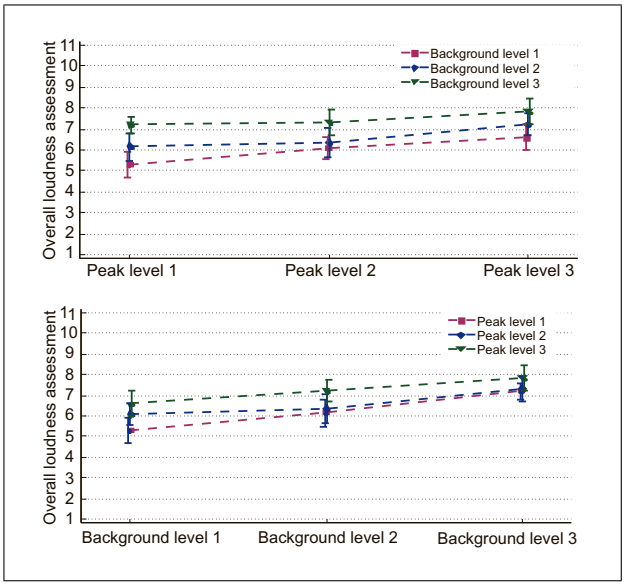

Figure 3. Treatment mean plots of Experiment 2: Relationship between overall loudness assessments and peak and background levels in a single category scaling experiment with filler sounds included. The category scale ranges from not at all (1) to very (11). The arithmetic means and 95% confidence intervals are displayed. The corresponding loudness values of all peak and background levels are shown in Table II.

level of background – are of importance for overall loudness assessments.

The effect sizes are rather large for the peak and background factor. The partial and generalized eta squared are $\eta_p^2 = 0.76$ and $\eta_g^2 = 0.25$ for the peak factor and amount $\eta_p^2 = 0.68$ and $\eta_g^2 = 0.18$ for the background factor.[7]

3.2.2. Experiment 2 - Overall loudness assessment in a single category scaling experiment with filler sounds included

A two-way ANOVA test for repeated measures design on both factors shows that both factors contribute significantly to the assessment of overall loudness for the 10 s stimuli in a single category scaling experiment with additional filler sounds presented. The Mauchly's sphericity test indicated a violation of group variance equality assumption for the peak factor. A highly statistically significant result for the peak factor was found using the Greenhouse-Geisser correction ($F_{\text{peak}}(1.49, 26.8) = 29.7$, $p < 0.01^{**}$, $\varepsilon_{GG} = 0.74$). The influence of background loudness on overall loudness was also highly statistically significant ($F_{\text{background}}(2, 36) = 25.6$, $p < 0.01^{**}$). An interaction effect between both factors was not identified. The results are shown in Figure 3.

The effect sizes are medium to large for the factors peak and background. The effect size of the peak factor amounts $\eta_p^2 = 0.62$ and $\eta_g^2 = 0.11$, whereas the effect size of the background factor regarding overall loudness assessments is $\eta_p^2 = 0.59$ and $\eta_g^2 = 0.20$. The greater meaning of the background magnitude for overall loudness assessments in this experiment can be recognized in Figure 3. In general, the effect sizes expressed in generalized eta squared values are smaller than observed in the first experiment. This indicates more error variance in the data probably resulting from distracting filler material.

3.2.3. Experiment 3 - Overall loudness assessment in a multiple category scaling experiment with filler sounds included

The analysis of the assessments of the manipulated noise samples showed similar experimental results as observed in the first two experiments requesting overall assessments on only one category scale (see Figure 4). Again both factors, peak loudness ($F_{\text{peak}}(2, 34) = 15.4$, $p < 0.01^{**}$) as well as the background loudness ($F_{\text{background}}(2, 34) = 14.9$, $p < 0.01^{**}$), contribute significantly to the assessment of overall loudness for the investigated synthetic noises. An interaction effect was not observed.

The effect sizes are small for the peak, and larger for the background factor. The effect size of the peak factor is $\eta_p^2 = 0.48$ and $\eta_g^2 = 0.10$. The effect size η_g^2 of the background factor is slightly larger with $\eta_p^2 = 0.47$ and $\eta_g^2 = 0.14$.

[7] In general, there is a broad discussion about adequate measures of strength of association for interpreting empirical findings. Partial eta squared η_p^2 is frequently proposed for repeated measures ANOVA to provide a measure of strength of association between an independent variable and a dependent variable that excludes variance produced by other factors and to offer a measure which allows for comparing the strength of association between the same independent variable and dependent variable across studies that have different factorial designs [62]. Bakeman recommends the use of the generalized eta squared η_g^2, which provides a value comparable across studies regardless of whether the factor is considered in between-subjects or within-subjects design of experiment [52].

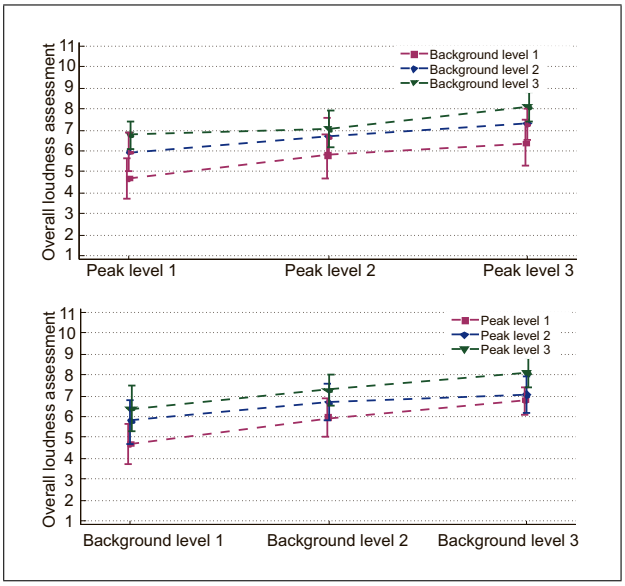

Figure 4. Treatment mean plots of Experiment 3: Relationship between overall loudness assessments and peak and background levels in a multiple category scales experiment (semantic differential technique). The category scale ranges from not at all (1) to very (11). The arithmetic means and 95% confidence intervals are displayed. The corresponding loudness values of all peak and background levels are shown in Table II.

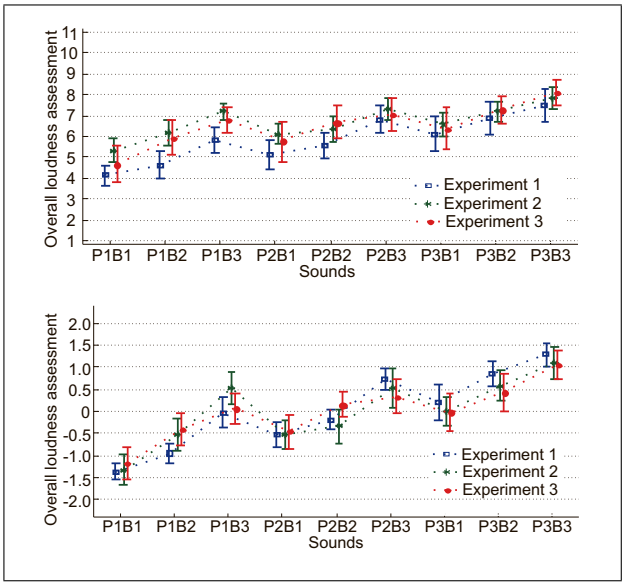

Figure 5. Top: Overall loudness assessments (arithmetic mean values and standard deviation) of sounds varying in the magnitude of peak (P) and background (B) judged under different conditions. Bottom: Overall loudness assessments as standard scores of sounds varying in the magnitude of peak (P) and background (B) judged under different conditions.

3.2.4. Comparison of experimental results and their relation to loudness indicators

In all experiments the null hypothesis regarding the importance of background loudness and peak loudness was rejected. The relevance of both factors, peak and background within a noise sequence, turned out to be statistically significant for the assessment of overall loudness. The effect sizes of peak and background for overall loudness are comparable over the experiments to a certain extent. Peak and background magnitude significantly influence assessed overall loudness, meaning that an increase in the magnitude of one factor by keeping the magnitude of the other factor constant leads to a statistically significant difference in the resulting overall loudness assessment.

Comparing the influence of the experimental circumstances on the overall loudness assessments, a 3-way ANOVA with unequal sample sizes showed a statistically significant main effect due to the experimental condition ($F(2, 450) = 13.6$, $p < 0.01^{**}$). Post-hoc tests on the basis of Scheffe's test showed that the first experiment differed statistically significant to experiment 2 and 3. This means that the experimental results in terms of the absolute category assessments of the noise stimuli differ statistically significant over the experiments mainly due to the introduction of filler sounds. The further distraction by requesting assessments regarding multiple attributes possibly changing the attention process and the way of listening did not significantly alter the overall loudness assessments. In general, due to frequency-range effects, it was highly expected that assessments of specific sounds on a particular 11-point category scale differ in dependence of their respective stimulus context [50]. The effect size of the experimental condition factor is relatively small ($\eta_p^2 = 0.06$).

The other main effects, the peak and background factor, were found to be highly statistically significant as well considering all experiments ($F_{\text{peak}}(2, 450) = 38.3$, $p < 0.01^{**}$, $F_{\text{background}}(2, 450) = 42.5$, $p < 0.01^{**}$). The effect size of the peak factor in terms of the partial eta squared is $\eta_p^2 = 0.15$. Comparably, the effect size η_p^2 of the background factor is $\eta_p^2 = 0.16$ over the three different experiments. This result underlines that the magnitude of the peak and background loudness play a meaningful role regarding overall loudness assessments independent from the experimental circumstances considered in this study. The assessments of overall loudness of the systematically manipulated synthetic noise stimuli correlate highly statistically significant over all experiments (product-moment correlation: $r > 0.92^{**}$).

Figure 5 illustrates the comparability of experimental results with respect to the nine factorial designed noise samples. In particular, if the standard scores are analyzed, the different experimental outcomes are very similar.

Moreover, it was observed that in the first experiment the magnitude of peak event resulted in larger effects than the background loudness. But, the effect size changes with the introduction of filler sounds and further by the introduction of potentially distracting additional evaluation scales; the effect sizes of the background factor were higher than the peak factor for the experiment 2 and 3 (see Table IV). However, post-hoc tests using Scheffe's method showed that in all cases, for the background and peak factor, a change from lowest to the highest level led always to a statistically significant difference in overall loudness.

Table IV. Effect sizes in generalized eta-squared η_g^2 for the peak and background factors over the conducted experiments [52]. A: Single category scaling experiment without filler sounds; B: Single category scaling experiment with filler sounds; C: Multiple category scaling experiment with filler sounds.

	A	B	C
Peak factor	0.25	0.11	0.10
Background factor	0.18	0.20	0.13

In the first experiment even all factor levels result in statistically significant overall loudness changes illustrating the potential impact of experimenter demand effects. In experiment 2 and 3 a loudness increase of approximately 30% of a factor level did not always lead to a statistically significant change in overall loudness. This means that the addition of filler material distracting participants from the manipulated features of the stimuli of interest did not considerably influence the experimental outcomes. Moreover, the change of the evaluation task from a single category scaling task to requesting assessments on multiple category scales possibly changing the attention process and the way of listening did not alter significantly the experimental result either.

Due to collinearity of potential loudness indicators and the low number of noise samples, it is not possible to reliably identify the most powerful loudness metric. In general, a significant relationship between mean indicators (like median loudness) and assessed overall loudness was observed. The same applies for the importance of peak events on overall loudness assessments; peak related indicators like N_5 correlate with the overall loudness assessments as well. Varying experimental conditions, such as adding filler material to reduce the attention to the manipulated features leading to a potentially less analytical listening, did not generally change the link between factors and assessed overall loudness. Since both peak and background magnitude play a statistically significant role in the assessment of overall loudness, power mean indicators based on psychoacoustic loudness values over time perform better than single loudness percentile values, like N_5 proposed in DIN 45631/A1 or N_{50}. In contrast to single percentile loudness indicators, power mean metrics incorporate background and peak magnitudes. The link between computed single loudness indicators representing the overall loudness of the sound and the respective assessments of overall loudness collected in the three different experiments is shown in Figure 6.

It can be seen that the root mean squared (quadratic mean) or root mean cubed (cubic mean) loudness metric correlate better with the overall loudness assessments collected under different experimental conditions than the N_5 indicator or the median loudness N_{50}. The loudness level mean value proposed by Schlittenlacher *et al.* [29], which is comparable to the cubic mean loudness metric due to the relationship between loudness level *LL* in phons and loudness N in sones, performs reasonably as well regarding the prediction of the overall loudness assessments of the presented noise samples. The cubic mean loudness metric and the loudness level *LL* in phons showed over all experiments the smallest root mean squared errors (0.29 and 0.31), whereas the root mean squared errors of the arithmetic and geometric mean of the loudness values over time were clearly larger (0.49 and 0.54). The loudness percentile metrics N_5 and N_{50} revealed even larger root mean squared errors (0.65 and 0.6). The mean squared error was determined assuming a linear relation between loudness metrics and the overall loudness assessments.

3.3. Discussion

In all experiments both factors, peak and background, were related to the assessment of overall loudness and their variation led to statistically significant differences in the means of overall loudness assessments. Statistically significant interaction effects were not found. This result demonstrates the insufficiency of single loudness percentile values, which cannot cover peak and background magnitudes at the same time. Any loudness indicator must encompass the magnitude of salient loud events as well as of steady parts within an episode.

However, in Figure 3 it can be seen that the stimuli with peak level 1 and peak level 2 were judged similar. Only the highest peak level leads to considerably higher overall loudness assessments compared to the lower peak magnitudes. This indicates that the peak event does only contribute to the overall loudness assessment, if a certain affective threshold is exceeded. Below this threshold under the experimental circumstances the overall loudness assessment is more or less only influenced by the long-lasting "background" magnitude. Therefore, the required weighting of the peak and background components is not conclusive. If beta weights of respective multiple linear regression models considering peak and background loudness are determined, the magnitudes of weights vary over the experiments. In two experiments the peak loudness has a slightly higher beta weight, whereas in experiment 2 the background loudness achieved a higher beta weight. For the considered stimuli and experimental conditions, the use of power mean indicators with exponents larger than 1 (e.g. quadratic and cubic mean) appear to predict the overall loudness assessments well. With increasing exponent a larger emphasis is put on loud events within an episode.

However, the level of generality of the results might be limited due to the limited set of stimuli and experimental circumstances. First of all, the investigation of the importance of peak or background levels is based on psychoacoustic calculations of loudness magnitudes as a proxy of the momentary intensity of loudness. It is implicitly assumed that listeners base their decisions on some "internal" computation of instantaneous loudness to derive overall loudness. However, it might be favorable to rely on real-time continuous judgments of loudness which might vary from each other instead of using psychoacoustic functions over time. However, the influence of online measurements of perception on overall assessments is unclear.

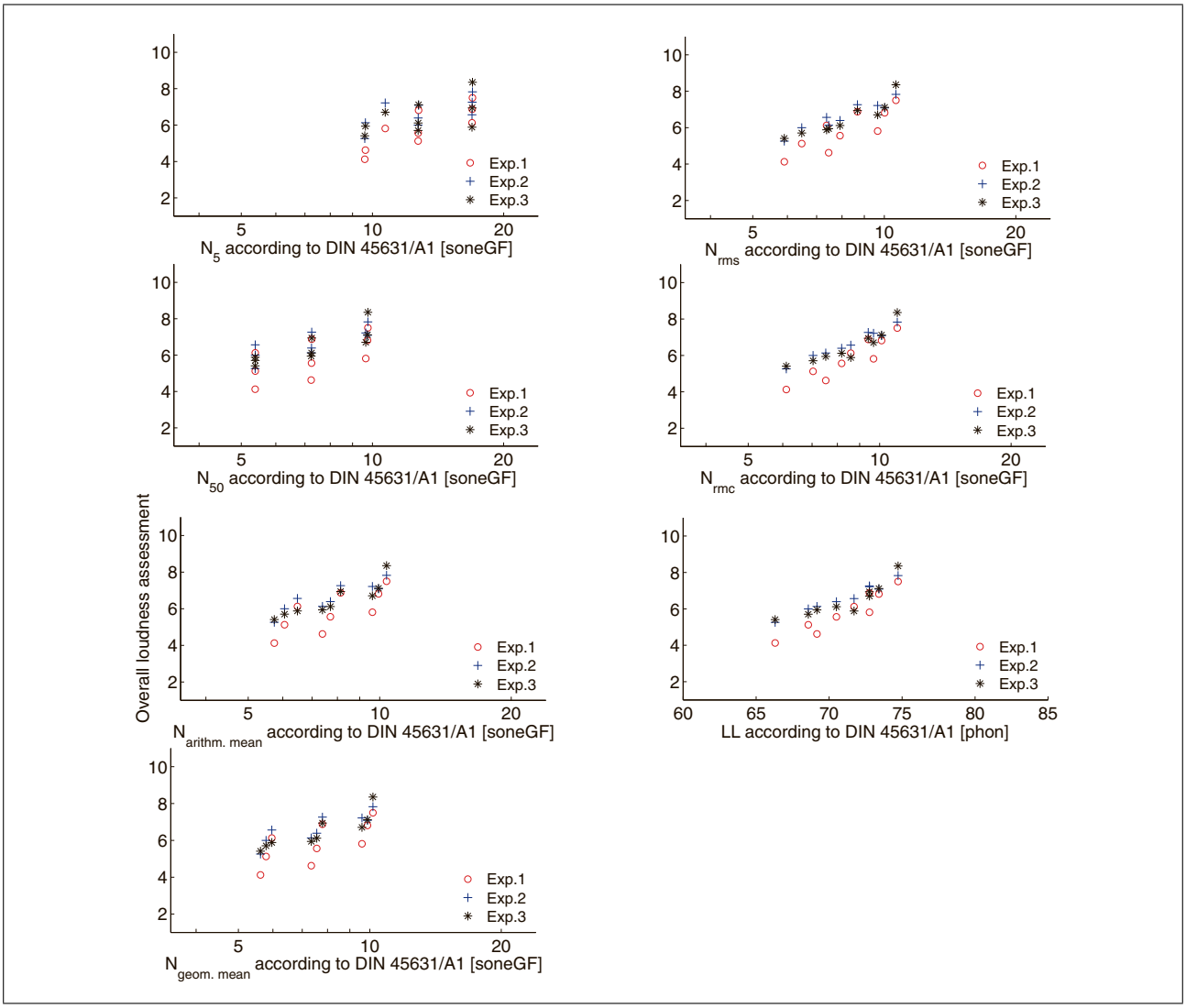

Figure 6. Relationship between different loudness metrics and overall loudness of nine noise samples assessed in three experiments. From top left to bottom right: Overall loudness assessments over N_5[8], overall loudness assessments over median loudness (N_{50}), overall loudness assessments over arithmetic mean of N ($N_{\text{arith.mean}}$), overall loudness assessments over geometric mean of N ($N_{\text{geom.mean}}$), overall loudness assessments over root mean square of N (N_{rms}), overall loudness assessments over root mean cubed of N (N_{rmc}), overall loudness assessments over loudness level mean *LL*. All single values are based on the loudness calculation according to DIN 45631/A1.

Some studies provided evidence regarding the impact of online evaluations on the relationship between hedonic patterns and overall assessments. According to Ariely and Zauberman, online measurements of momentary experiences caused retrospective evaluations to be closer to the mean of momentary experiences, which is interpreted as an increased reliance on the responses participants produced online [53]. In contrast to it, Steffens and Guastavino did not observe significant effects due to momentary judgments sustaining attentive listening on retrospective judgments [36]. However, Steffens and Guastavino noticed that the mean judgments of the control group, which provided overall retrospective judgments only, were closer to the center of the scale than the mean judgments of the experimental group, which provided overall as well as momentary judgments. Susini *et al.* observed that global judgments were slightly higher to those global judgments with prior provided continuous judgments [7]. Independent from discussions about "possible response contamination from the online evaluations to the retrospective evaluations" [54], it seems uncommon to permanently report instantaneous perception regarding everyday experiences and accordingly, Fredrickson concluded that simply experiencing an episode without requesting momentary ratings can be considered to be a more typical circumstance [17].

Moreover, the duration of a stimulus and the time making an overall assessment might play a particular role for

[8] For the lowest peak level and highest background level a slightly higher N_5 value is computed due to the similar loudness of background and peak, however the magnitude of the peak component is not changed at all.

summarized evaluations. The 10 s duration of presented stimuli might be too long to base an assessment on sensory memory as an out of cognitive control and automatic response act [55]. The use of the short-term memory might trigger different cognitive processes compared to sensory memory or long-term memory. In particular, the duration between the termination of a stimulus and the time at which an overall loudness assessment was provided varied between the experiments. In the third experiment the overall loudness was presumably judged several seconds after the end of the noise episode. Although the duration between stimulus end and overall loudness assessment might have been larger in the third experiment due to the presentation of additional category scales relative to the other two experiments, both main effects (peak and background) remain statistically significant. However, it must be assumed that by extending the time until an assessment is requested, forcing participants to apply increasingly the long-term memory, might lead to different results. According to Winkler and Cowan it is still unclear if specific acoustic information disappears as a function of absolute time or disappears as a function of the shifting context [56]. The observed relationships between factor levels and overall loudness assessments appear only to be valid for short-term memory assessments, where the working memory is applied.

4. Conclusions

By means of analyses of variances for two-factor repeated measures design with repeated measures on both factors it was confirmed in different experiments that both peak and background magnitudes are relevant for overall loudness assessments. The null hypotheses can be rejected and the alternative hypotheses can be considered to be true; loud events as well as longer quieter parts of a sound episode influence the assessment of overall loudness. The absence of any interaction effect between peak and background magnitudes suggests additivity. This can be interpreted that both variables can be "added" together in a certain way to estimate the dependent variable speaking for an "adding-type of integration" of both variables [57]. The factorial design experiments with varying test conditions have repeatedly confirmed that the participants did not deliberately neglect or even completely ignore larger parts of episodes and their respective intensity when they are requested to provide an overall sound assessment.

In general, the experimental results indicate that it is not sufficient to rely on either the N_5 loudness indicator or median loudness as a single predictor of overall loudness of time-variant noise episodes. Any loudness predictor must encompass the magnitude of salient sensational (perceptual) events as well as "background" magnitude(s) within a sound episode. Based on power mean values of the loudness over time, the level of prominent events and background can be considered at the same time; whereas the weighting of these aspects depends on the applied exponent. For the assessment of overall loudness the gained experimental results suggest an exponent to be larger than 1. Further experiments indicate different exponents of power mean values for other sound perception dimensions [45].

However, it is evident that everyday life episodes are mostly longer than a few seconds. If an experience lasts hours strongly varying in intensity, then a detailed representation of the perceptual profile appears ineffective, because it would produce an unnecessarily high cognitive load. Thus, potential unifying principles observed in laboratory experiments regarding the assessment of overall loudness of short-term noise episodes might be valid only to a limited degree to prolonged noise assessments in daily life and must be scrutinized with respect to their ecological validity.

Finally, it must be noted that it is not clear that a phenomenological representation of overall loudness in fact exists. According to Schreiber and Kahneman there is possibly no representation of total brightness, total loudness or total utility, since the intensity of pleasure, pain or loudness is only an attribute of a moment [13]. Thus, any psychological scale is only a concept, which the experimenter uses because it provides meaning and generality [58], but does not necessarily represent the way of human perception itself. Thus, the question seems justified, when participants are requested to provide an assessment of overall loudness of time-variant noises: Do humans perceive or do they construct an overall loudness? The elucidation of this general question will probably influence the concept of (overall) loudness perception.

Acknowledgement

The authors would like to thank two anonymous reviewers and the editor for their valuable comments, which improved the manuscript considerably and Klaus Genuit and Brigitte Schulte-Fortkamp for fruitful discussions.

References

[1] D. Ariely, Z. Carmon: Summary assessment of experiences: The whole is different from the sum of its parts. – In: Time and decision. Economic and Psychological Perspectives on Intertemporal Choice. G. Loewenstein, D. Read, F. Baumeister (eds.). Russel Sage Foundation, New York, USA, 2003.

[2] S. Frederick, G. Loewenstein: Conflicting motives in evaluations of sequences. J. Risk Uncertain, Springer, DOI 10.10070/s11166-008-9051-z, 2008.

[3] D. Kahneman: Evaluation by moments: Past and future. – In: Choices, values and frames. D. Kahneman, A. Tversky (eds.). Cambridge University Press, New York, USA, 1999.

[4] B. L. Fredrickson, D. Kahneman: Duration neglect in retrospective evaluations of affective episodes. Journal of Personality and Social Psychology **65** (1993) 45–55.

[5] T. Schäfer, D. Zimmermann, P. Sedlmeier: How we remember the emotional intensity of past musical experiences. Front. Psychol., August 2014, doi: 10.3389/fpsyg.2014.00911, 2014.

[6] K. Dittrich, D. Oberfeld: A comparison of the temporal weighting of annoyance and loudness. J. Acoust. Soc. Am. **126** (2009) 3168–3178.

[7] P. Susini, S. McAdams, B. K. Smith: Loudness asymmetries for tones with increasing and decreasing levels using continuous and global ratings. Acta Acustica united with Acustica **93** (2007) 623–631.

[8] A. Algom, Y. Wolf, B. Bergman: Integration of stimulus dimension in perception and memory: Composition rules and psychophysical relations. Jornal of Experimental Psychology: General **114** (1985) 451–471.

[9] B. A. Schneider, S. Parker: The evolution of psychophysics: From sensation to cognition and back. Proceedings of Fechner day 2010, 2010, Vol. 26.

[10] S. Großmann, H. Fastl: Echtzeitbeurteilung instationärer Signale in hörversuchen. Proceedings of DAGA 2010, Berlin, Germany, 2010.

[11] S. Kuwano, S. Namba, T. Kato, J. Hellbrück: Memory of the loudness and its relation to overall impression. Forum Acusticum, Sevilla, Spain, 2002.

[12] D. Västfjäll: The "end effect" in retrospective sound quality evaluation. Acoust. Sci. & Tech., Acoustical Letter **25** (2004) 170–172.

[13] C. A. Schreiber, D. Kahneman: Determinants of the remembered utility of aversive sounds. Journal of Experimental Psychology, General **129** (2000) 27–42.

[14] K. Genuit, R. Sottek, A. Fiebig: Comparison of loudness calculation procedures in the context of different practical applications. Internoise 2009, Ottawa, Canada, 2009.

[15] D. A. Redelmeier, J. Katz, D. Kahneman: Memories of colonoscopy: a randomized trial. Pain 104, Elsevier, 2003, 187–194.

[16] R. Paulsen: On the influence of the stimulus duration on psychophysical judgment of environmental noises taken in the laboratory. Internoise 1997, Budapest, Hungary, 1997.

[17] B. L. Fredrickson: Extracting meaning from past experiences: The importance of peaks, ends, and specific emotions. Cognition and Emotion **14** (2000) 577–606.

[18] D. Ariely, G. Loewenstein: When does duration matter in judgment and decision making? Journal of Experimental Psychology: General **129** (2000) 508–523.

[19] T. Poulsen: Loudness of tone pulses in a free field. J. Acoust. Soc. Am. **69** (1981) 1786–1790.

[20] K. N. Olsen: Intensity dynamics and loudness change: A review of methods and perceptual processes. Acoustics Australia **42** (2014) 159–165.

[21] K. N. Olsen: Perceptual bias and loudness change: An investigation of memory, masking, and psychophysiology. Doctoral thesis, University of Western Sydney, Sydney, Australia, 2011.

[22] H. Fastl: Beurteilung und Messung der äquivalenten Dauerlautheit. Z. f. Lärmbekämpf. **38** (1991) 98–103.

[23] I. Stemplinger: Globale Lautheit von gleichförmigen Industriegeräuschen. DAGA 1996, Bonn, Germany, 1996.

[24] I. Stemplinger: Beurteilung der globalen Lautheit bei Kombination von Verkehrsgeräuschen mit simulierten Industriegeräuschen. DAGA 1997, Kiel, Germany, 1997.

[25] R. Sottek: Modelle zur Signalverarbeitung im menschlichen Gehör. Doctoral thesis, RWTH Aachen University, Aachen, Germany, 1993.

[26] DIN 45631/A1: Calculation of loudness level and loudness from the sound spectrum - Zwicker method - Amendment 1: Calculation of the loudness of time-variant sound. Deutsches Institut für Normung, Beuth Verlag, Berlin, Germany, 2010.

[27] ISO 532-1 CD: Methods for calculating loudness. Part 1: Zwicker method. International Standardization Organization, Geneva, Switzerland, 2015.

[28] S. Namba, T. Kato, S. Kuwano: Evaluation of loudness level of time-varying sounds. Internoise, Osaka, Japan, 2011.

[29] J. Schlittenlacher, T. Hashimoto, S. Kuwano, S. Namba: Overall loudness of short time-varying sounds. Internoise 2014, Melbourne, Australia, 2014.

[30] S. Friebe: Äquivalente Dauerlautheit von Kindergeräuschen im Vergleich zu konventionellen Bewertungsverfahren. Diploma thesis, Technical University Ilmenau, Germany, 2011.

[31] S. Ferguson, D. Cabrera, E. Schubert: Comparing continuous subjective loudness responses and computational models of loudness for temporally varying sounds. 129th Audio Engineering Society Convention, San Francisco, CA, USA, 2010.

[32] J. Rennies, I. Holube, J. L. Verhey: Loudness of speech and speech-like signals. Acta Acustica united with Acustica **99** (2013) 268–282.

[33] B. R. Glasberg, M. B. C. J.: A model of loudness applicable to time-varying sounds. Journal of the Audio Engineering Society **50** (2002) 331–341.

[34] R. Höger, E. Matthies, E. Letzing: Physikalische versus psychologische Reizintegration: Der Mittelungspegel aus wahrnehmungspsychologischer Sicht. Z. f. Lärmbekämpf. **35** (1988) 163–167.

[35] P. Susini, S. McAdams, B. K. Smith: Global and continuous loudness estimation of time-varying levels. Acta Acustica united with Acustica **88** (2002) 536–548.

[36] J. Steffens, C. Guastavino: Trend effects in momentary and retrospective soundscape judgments. Acta Acustica united with Acustica **101** (2015) 713–722.

[37] W. Ellermeier, S. Schrödl: Zeitliche Gewichtung bei der Lautheitsintegration. DAGA 2000, Oldenburg, Germany, 2000.

[38] D. Oberfeld: The temporal weighting of the loudness of time-varying sounds reflects both sensory and cognitive processes. International Society for Psychophysics **25** (2009) 75–78.

[39] B. Pedersen, W. Ellermeier: Temporal weights in the level discrimination of time-varying sounds. J. Acoust. Soc. Am. **123** (2008) 963–972.

[40] D. Oberfeld, T. Plank: Temporal weighting of loudness: Effects of a fade in. DAGA 2005, Munich, Germany, 2005.

[41] E. Ponsot, P. Susini, G. Saint Pierre, S. Meunier: Temporal loudness weights for sound with increasing and decreasing intensity profiles. J. Acoust. Soc. Am. **134** (2013) EL321.

[42] B. Pedersen: Auditory temporal resolution and integration. Stages of analyzing time-varying sounds. Doctoral thesis, Aalborg University, Alborg, Denmark, 2006.

[43] K. Laumann, H. Fastl, S. Kuwano, S. Namba: Overall loudness versus average of instantaneous loudness for excerpts of music: Effects of musical style. DAGA 2007, Stuttgart, Germany, 2007.

[44] D. Menzel, H. Fastl, R. Graf, J. Hellbrück: Influence of vehicle color on loudness judgments. J. Acoust. Soc. Am. **123** (2008) 2477–2479.

[45] A. Fiebig: Cognitive stimulus integration in the context of auditory sensations and sound perceptions. Doctoral thesis (in press), Technical University Berlin, Berlin, Germany, 2015.

[46] H. Fastl, E. Zwicker: Psychoacoustics. Facts and models. Springer Verlag, Heidelberg, New York, Berlin, 2007.

[47] J. Chalupper, H. Fastl: Dynamic loudness model (DLM) for normal and hearing-impaired listeners. Acta Acustica united with Acustica **88** (2002) 378–386.

[48] D. J. Zizzo: Experimenter demand effects in economic experiments. Social science research network. 2008.

[49] D. Ariely, D. Kahneman, G. Loewenstein: Joint comment on "when does duration matter in judgment and decision making?" (ariely and loewenstein, 2000). Journal of Experimental Psychology: General **129** (2000) 524–529.

[50] A. Parducci: Perceptual and judgmental relativity. – In: Perspectives in psychological experimentation: Toward the year 2000. V. Sarris, A. Parducci (eds.). Lawrence Erlbaum Associates, Hillsdale, New Jersey, USA, 1984.

[51] K. Genuit, W. R. Bray: The AACHEN HEAD system - Binaural recording for headphones and speakers. U.S. Audio, December 1989, 58–66, 1989.

[52] R. Bakeman: Recommended effect size statistics for repeated measures designs. Behavior Research Methods **37** (2005) 379–384.

[53] D. Ariely, G. Zauberman: On the making of an experience: The effects of breaking and combining experiences on their overall evaluation. J. of Behav. Dec. Making **13** (2000) 219–232.

[54] D. Ariely: Combining experiences over time: The effect of duration, intensity changes and on-line measurements on retrospective pain evaluations. Journal of Behavioral Decision Making **11** (1998) 19–45.

[55] Z.-L. Lu, G. Sperling: Measuring sensory memory: Magnetoencephalography habituation and psychophysics. – In: Magnetic source imaging of the human brain. Z.-L. Lu, K. L. (eds.). Lawrence Erlbaum Associates, NJ, USA, 2003, 319–342.

[56] I. Winkler, C. Nelson: From sensory to long-term memory evidence from auditory memory reactivation studies. Experimental Psychology **52** (2004) 1–17.

[57] D. Algom: Differences between 'early' and 'late' processing of stimulus dimensions in perception? The role of context invariance. – In: Fechner Day 2008. B. A. Schneider, B. Ben-David, S. Parker, W. Wong (eds.). International Society for Psychophysics, Toronto, Canada, 2008, 59–64.

[58] H. Helson: Perception. – In: Contemporary approaches to psychology. H. Helson, W. Bevan (eds.). van Nostrand, Princeton, New Jersey, Toronto, London, 1967.

[59] R. Sottek: Progresses in calculating tonality of technical sounds. Internoise 2014, Melbourne, Australia, 2014.

[60] A. Zeitler: Auditory pleasantness. Methodological considerations in the applications of psychophysical scaling methods for sound quality evaluation. Doctoral Thesis, Logos Verlag Berlin, Berlin, Germany, 2002.

[61] A. Khan, G. D. Rayner: Robustness to non-normality of common tests for the many-sample location problem. Journal of applied mathematics and decision sciences **7** (2003).

[62] C. A. Pierce, R. A. Block, H. Aguinis: Cautionary note on reporting eta-squared values from multifactor ANOVA designs. Educational and Psychological Measurement **64** (2004) 916–924.