A comparison of the temporal weighting of annoyance and loudness

Kerstin Dittrich[a] and Daniel Oberfeld
Department of Psychology, Johannes Gutenberg-Universität Mainz, 55099 Mainz, Germany

(Received 20 August 2008; revised 16 July 2009; accepted 26 August 2009)

The influence of single temporal portions of a sound on global annoyance and loudness judgments was measured using perceptual weight analysis. The stimuli were 900-ms noise samples randomly changing in level every 100 ms. For loudness judgments, Pedersen and Ellermeier [J. Acoust. Soc. Am. **123**, 963–972 (2008)] found that listeners attach greater weight to the beginning and ending than to the middle of a stimulus. Qualitatively similar weights were expected for annoyance. Annoyance and loudness judgments were obtained from 12 listeners in a two-interval forced-choice task. The results demonstrated a primacy effect for the temporal weighting of both annoyance and loudness. However, a significant recency effect was observed only for annoyance. Potential explanations of these weighting patterns are discussed. Goodness-of-fit analysis showed that the prediction of annoyance and loudness can be improved by allowing a non-uniform weighting of single temporal portions of the signal, rather than assuming a uniform weighting as in measures like the energy-equivalent level (L_{eq}). A second experiment confirmed that the listeners were capable of separating annoyance and loudness of the stimuli. Noises with the same L_{eq} but different amplitude modulation depths were judged to differ in annoyance but not in loudness.

© *2009 Acoustical Society of America.* [DOI: 10.1121/1.3238233]

PACS number(s): 43.66.Mk, 43.66.Cb, 43.66.Ba, 43.50.Ba [BCM] Pages: 3168–3178

I. INTRODUCTION

A. Annoyance and noise exposure

The perceptual dimension annoyance has received considerable interest over the last few decades (for recent reviews see Kryter, 2007; Marquis-Favre *et al.*, 2005a, 2005b). Parameters influencing annoyance can be divided into acoustical parameters (cf. Zwicker, 1991), such as the presence of tonal components (e.g., Hellman, 1984, 1985) or frequency (e.g., Leventhall, 2004), and non-acoustical variables, such as individual noise-sensitivity (e.g., Zimmer and Ellermeier, 1996). The non-acoustical variables could explain why listeners' evaluations of the annoyance of sounds differ widely. Loud rock music, for example, can be a pleasant event as well as an annoying disturbance.

An important focus of research on noise is to examine the consequences of noise exposure, which is commonly associated with annoyance reactions. Noise can have negative effects on the auditory system, for example, inner ear damage. Non-auditory effects also occur, for example, sleep disturbance, impairment of work performance, or interference with daily activities (e.g., Michaud *et al.*, 2008; see Marquis-Favre *et al.*, 2005a for a recent review). Given the fact that noise exposure has a lot of negative consequences, engineering standards and laws have been developed to protect people against these negative consequences (for an international example see, e.g., Guidelines for Community Noise, WHO, 1999). These regulations mostly use technical measures like the energy-equivalent level (L_{eq} or L_{Aeq}) in order to assess the annoyance or loudness of noises.[1] The measures used to quantify noise can be divided into different categories (cf. Marquis-Favre *et al.*, 2005a), for example, those related to the sound pressure level (e.g., L_{A}), energy-based indices (e.g., L_{eq}), or statistical indices (e.g., N_5).[2] Given the fact that sound intensity is an important factor for both annoyance and loudness (e.g., Zwicker, 1991; Hellman, 1982), the same measures are frequently used for the two dimensions (cf. Marquis-Favre *et al.*, 2005a; Schomer *et al.*, 2001). Nevertheless, annoyance and loudness depend in a different manner on the characteristics of sounds. For example, Zwicker (1991) proposed that besides loudness, amplitude modulation depth and sharpness should be taken into account in annoyance calculations.

For assessing the annoyance of longer sounds that fluctuate in level, several alternative measures have been proposed (e.g., N_5, L_{eq}, or L_{A}; cf. Zwicker and Fastl, 1999). Most countries use some variant of the A-weighted energy-equivalent level (cf. Schomer *et al.*, 2001). The validity of L_{eq} and L_{Aeq} for estimating the annoyance of real-world noises was partially confirmed in some studies (e.g., Hiramatsu *et al.*, 1983; Kuwano and Namba, 2000). These measures take into account acoustical parameters such as sound pressure level and frequency spectrum. However, the correlations between these measures and annoyance judgments are frequently found to be rather weak (see Marquis-Favre *et al.*, 2005a, 2005b, for recent reviews). The weak correlation can be ascribed to at least two different causes. First, these measures do not take into account non-acoustical factors such as individual sound sensitivity. Second, relevant acoustical parameters might not be considered in these measures.

[a] Author to whom correspondence should be addressed. Present address: Department of Psychology, Albert-Ludwigs-Universität Freiburg, 79085 Freiburg, Germany. Electronic mail: dittrich@psychologie.uni-freiburg.de

B. Temporal aspects of annoyance

One acoustical parameter which has not received much consideration until now is the *temporal aspect* of annoyance (notable exceptions are Hiramatsu *et al.*, 1983; Dornic and Laaksonen, 1989; Namba and Kuwano, 1979, 1980). This study is concerned with the question of whether and how the influence of single temporal portions of a longer stimulus on annoyance varies as a function of the temporal position within the sound. In a two-interval forced-choice task, two noises consisting of nine contiguous 100-ms segments were presented. The task was to select the more annoying noise. On each trial, the sound pressure levels of the nine segments were drawn independently from a normal distribution for each of the two noises, with a 1 dB difference in mean level between the two intervals. In such a setting, the *perceptual weight* is defined as the relative influence that the level of a given temporal segment had on the decision of the listener. These weights can be estimated from the trial-by-trial data using *molecular* analyses (e.g., Ahumada and Lovell, 1971; Berg, 1989; Richards and Zhu, 1994).

If listeners are asked to judge the overall loudness of the described type of sounds, the initial and final portions of the stimulus receive greater weight than its temporal center (e.g., Ellermeier and Schrödl, 2000; Oberfeld, 2008a, 2008b; Pedersen and Ellermeier, 2008). In other words, primacy and recency effects are observed. Pedersen and Ellermeier (2008) suggested that an interaction of perceptual and cognitive processes leads to the observed primacy/recency weighting pattern. This assumption seems to be plausible given the fact that primacy and recency effects are not specific to this type of loudness judgment, but are ubiquitous in cognitive psychology. In studies of learning and memory of serially sorted information, the serial position curve frequently shows both a primacy and a recency effect (e.g., Postman and Phillips, 1965; Anderson *et al.*, 1998; Jones *et al.*, 2004). Similar memory effects have been found for the recall of nonverbal acoustical stimuli (McFarland and Cacace, 1992; Surprenant, 2001). In these studies, serial position effects were examined for tonal sequences with an overall duration up to 4 s. For loudness, one can assume that the levels of the single segments of a noise are processed as serially sorted information in a system exhibiting similar characteristics to short-term memory (Oberfeld, 2008b). The beginning and the ending can be assumed to be more distinct than the middle of the noise, and therefore have a stronger influence on a decision, as, for example, a loudness judgment (see Neath *et al.*, 2006 for a detailed discussion).

The present study compared the temporal weighting of loudness and annoyance. Primacy and recency effects were expected to show for both perceptual dimensions. One reason for this expectation was the close relation between loudness and annoyance (e.g., Zwicker, 1966; Hellman, 1984, 1985). Additionally, if the processing of the segments as serially sorted information caused the non-uniform temporal weighting, this effect should show for annoyance as well as for loudness.

Insight into the temporal weighting of annoyance is especially relevant for technical measures used in noise quantification. Conventional measures like L_{eq} assume that listeners weight the information provided by each temporal portion of a sound uniformly. The present study examined whether this approach is compatible with the perception of annoyance or whether temporal aspects should be considered in the estimation of annoyance.

In Experiment 1, listeners evaluated the relative annoyance and the relative loudness of two 900-ms samples of noise. The sound pressure level of the noise was changed randomly every 100 ms by drawing the level repeatedly and independently from a normal distribution. The influence of single temporal segments of this level-fluctuating noise on annoyance and loudness judgments was estimated using perceptual weight analysis (cf. Berg, 1989). Goodness-of-fit analysis was used to test whether the prediction of annoyance and loudness can be improved by allowing for a non-uniform weighting of single temporal portions of the signal.

A potential problem for the within-subjects comparison of temporal weights for loudness and annoyance is that listeners may not be capable of separating loudness and annoyance when repeatedly judging the same type of stimuli. If Experiment 1 actually showed the expected differences between the temporal weighting patterns for annoyance and loudness, then this would demonstrate that loudness and annoyance represented separate dimensions. If, on the other hand, an identical pattern of weights was found for annoyance and loudness, it might have been the case that subjects always evaluated the noises according to their loudness, even when they were asked to judge the stimuli according to their annoyance, or vice versa.

We used two methods to assess these possibilities. First, in Experiment 1, two groups of listeners were assigned to different task orders. One group made only annoyance judgments in the first part of the experiment, and only loudness judgments in the second part. For the other group, the order of the tasks was reversed. If the “true” weighting patterns for loudness and annoyance differed in any respect, then an effect of task order on the patterns of weights would indicate a failure of the listeners to switch between loudness and annoyance judgments. Second and more important, the listeners from Experiment 1 participated in an additional experiment (Experiment 2b) designed to more directly test whether they were capable of judging the level-fluctuating stimuli independently according to their loudness and their annoyance. To this end, we presented noises differing in modulation depth, that is, in the variability of the nine segment levels. Modulation depth has been reported to produce a dissociation between loudness and annoyance (Widmann, 1994). In Experiment 2, the two noises presented on each trial had the same L_{eq} while differing in modulation depth. We expected the listeners to perceive these sounds as similar in loudness, but to perceive the sound with the higher modulation depth as more annoying (Zwicker, 1991), in line with our assumption that listeners were able to separate the dimensions loudness and annoyance for our stimuli.

II. EXPERIMENT 1: COMPARISON OF THE TEMPORAL WEIGHTING OF ANNOYANCE AND LOUDNESS

A. Method

1. Listeners

Twelve listeners (8 women, 5 men, age 20–31 years) participated. Most were psychology students at the Johannes Gutenberg-Universität Mainz, participated for course credit, and had no experience in comparable psychoacoustic tasks. All listeners reported normal hearing. Detection thresholds in the right ear, as measured by a two-interval forced-choice, adaptive procedure with a three-down, one-up rule (Levitt, 1971), were better than 15 dB HL (hearing level; relative to the reference levels provided by Han and Poulsen, 1998) at all octave frequencies between 250 and 4000 Hz.

The individual noise-sensitivity of the listeners was assessed using the noise-sensitivity questionnaire of Zimmer and Ellermeier (1996). Noise-sensitivity is viewed as a trait reflecting individual differences in the tolerance of environmental noise. The questionnaire assesses the perceptual, cognitive, affective, and behavioral reactions to noise in different contexts. The mean noise-sensitivity was $M=85$, standard deviation $SD=8$, with a range from 71 to 100. Since the scale ranges from 0 to 186, none of the listeners could be considered as being extremely high or low in his or her noise-sensitivity. However, the listeners differed widely concerning their self-rated noise-sensitivity. On an 11 point rating scale ranging from 0 (not at all noise sensitive) to 10 (very noise sensitive), the range of self-rated noise-sensitivity was 2–9 ($M=5.7$, $SD=2.6$).

2. Apparatus

The stimuli were generated digitally, played back via two channels of an RME ADI/S digital-to-analog converter ($f_S=44.1$-kHz, 24-bit resolution), attenuated (two TDT PA5s), buffered (TDT HB7), and presented diotically via Sennheiser HDA 200 headphones calibrated according to IEC 318 (1970). The experiment was conducted in a single-walled sound-insulated chamber. Listeners were tested individually.

3. Stimuli and experimental procedure

The stimuli were presented in a two-interval procedure. Figure 1(a) shows a schematic depiction of a trial. On each trial, two level-fluctuating noises were presented. The stimuli were Gaussian wide-band noises consisting of nine contiguous temporal segments. The duration of each segment was 100 ms. On each trial and for each interval, the sound pressure levels of the nine temporal segments were drawn independently from a normal distribution. In the interval containing the less intense noise, the mean of the distribution was $\mu_L=64.5$-dB sound pressure level (SPL) and the SD was 2.5 dB. In the interval containing the more intense noise, the mean was $\mu_H=65.5$-dB SPL, also with $SD=2.5$ dB. Although the estimation of perceptual weights would be possible without a difference in mean level between the two intervals, we introduced this difference in level mainly to make the task easier for the subjects and also to be compatible with previous experiments (e.g., Berg, 1989; Ellermeier and Schrödl, 2000). The more intense noise was presented in interval 1 or interval 2 with identical *a priori* probability. To avoid overly loud sounds, the range of levels was restricted to $\mu\pm2.5$SD. Therefore, the maximal level difference between the most intense and the least intense segment within a given noise was 12.5 dB. The standard deviation of the nine segment levels (SD_{levels}) is a measure of the modulation depth. Across all trials, the mean modulation depth was 2.3 dB ($SD=0.55$ dB, range of 0.48–4.46 dB).

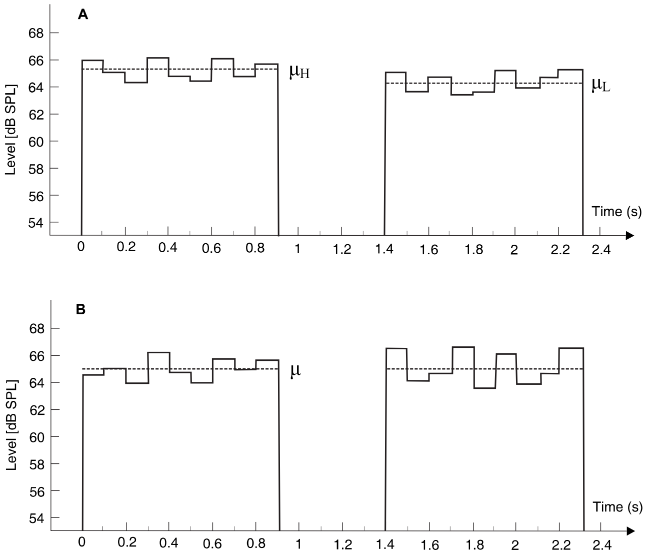

FIG. 1. Trial configurations used in Experiments 1 and 2. Two broad-band-noises consisting of nine contiguous segments were presented. On each trial, the level of each segment was drawn independently from one of two normal distributions differing in their means (Experiment 1) or standard deviations (Experiment 2). Task: First or second noise more annoying/louder? The dashed line represents the mean level. Panel (A): Experiment 1. Means of the normal distribution: $\mu_L=64.5$-dB SPL in one interval; $\mu_H=65.5$-dB SPL in the other interval. The SD was 2.5 dB in both intervals. Panel (B): Experiment 2. Small modulation depth (normal distribution with $SD=2$ dB) in one interval; large modulation depth (normal distribution with $SD=4$ dB) in the other interval. For the noise with the small modulation depth, the mean of the normal distribution was $\mu=65$ dB SPL. The two noises had the same L_{eq} or N_5.

The two noises were presented with a silent inter-stimulus interval of 500 ms. Depending on the task, the listeners selected the interval containing the more annoying or louder sound. No feedback was provided. The next trial followed the response after an inter-trial interval of 2 s.

In the last part of the experiment, magnitude estimates of the annoyance of the stimuli were obtained using a procedure without reference (e.g., Hellman and Zwislocki, 1961) and essentially the same instructions as in Hellman and Meiselman (1988). A single noise, fluctuating in level, was presented on each trial, with the segment levels drawn from a normal distribution with mean $\mu=65$ dB SPL and $SD=2$ dB.[3] Each listener judged 15 noises four times in randomized order. The listeners were asked to choose any positive number which seemed adequate to describe the annoyance of the presented noise. The geometric mean of the 60 numerical judgments was taken as the individual annoyance estimate. Across listeners, the mean magnitude estimate of annoyance was $M=0.75$ ($SD=0.13$, range of 0.55–1.00).

The listeners were randomly assigned to two experimental groups. Group 1 made only annoyance judgments in the first part of the experiment, and only loudness judgments in the second part. For Group 2, the order of tasks was reversed. The experiment was arranged in blocks of 50 trials. Each session comprised ten blocks and lasted approximately 60 min. Each listener completed six sessions. At the beginning of a session, the listeners received 50 practice trials. In the first session, the listeners from both groups completed the detection threshold measurements. In the second part of session 1, Group 1 received 300 trials of Experiment 2b (see description below) and judged the noises according to their annoyance. In sessions 2 and 3, the listeners of Group 1 received 1000 trials in the annoyance task of Experiment 1. In the second part of the experiment (sessions 4–6), Group 1 first made loudness judgments for 300 trials of Experiment 2b, and then completed 1000 trials in the loudness task of Experiment 1. For Group 2 the procedure was analogous.

At the end of session 6, the listeners provided magnitude estimates of the annoyance of the level-fluctuating noises (see above) and filled in the noise-sensitivity questionnaire (Zimmer and Ellermeier, 1996).

4. Estimation of temporal weights

Multiple binary logistic regression (PROC LOGISTIC, SAS 8.01) was used to estimate the weights from the trial-by-trial data.[4] For each trial and each segment ($i=1,\ldots,9$), the difference between the level of segment i in interval 2 and the level of segment i in interval 1 was computed. The binary responses served as the dependent variable and the nine within-trial segment level differences served as predictors. Due to the difference in mean level between the two intervals, the within-trial segment level differences were correlated. Therefore, separate logistic regression analyses were conducted for the trials in which the noise with the higher mean level (μ_H) occurred in interval 1, and for the trials in which the position of the noise with mean level μ_H was interval 2. Thus, a logistic regression was conducted for each combination of subject, task (annoyance/loudness), and position μ_H. Because modulation depth, that is, the variability of the levels within a sound, has an influence on annoyance (e.g., Widmann, 1994), the within-trial difference between the standard deviations of the nine segment levels (SD_{levels}) in interval 2 and the nine segment levels in interval 1 was included as a predictor. A comparison of the goodness-of-fit for models containing or not containing the within-trial difference in SD_{levels} as a predictor will be presented in Sec. II B.

The regression coefficients for the nine segment level differences were taken as weight estimates. The weights were normalized such that the sum of the absolute values was unity (see Kortekaas *et al.*, 2003), resulting in a set of relative temporal weights for each listener, task (annoyance/loudness), and position μ_H.

The unweighted residual sum-of-squares test (Copas, 1989) was used for assessing global goodness-of-fit. This test has been shown to perform favorably compared to some alternative tests (Hosmer *et al.*, 1997; Kuss, 2002). An SAS macro (GOFLOGIT; Kuss, 2001) was used to compute the test statistics. In global goodness-of-fit tests, the hypothesis is tested that the saturated (full) model containing as many parameters as observations does not provide a better description of the data than the fitted (restricted) model (cf. Agresti, 2002). Small p-values indicate lack-of-fit of the restricted model. It is usual to take p-values of less than 0.2 as an indication that the model did not fit adequately (cf. Agresti, 2002). For the 48 (Listener × Task × Position μ_H) fitted multiple logistic regression models the test produced a p-value below 0.2 in only five cases.

A summary measure of the predictive power of a logistic regression model is the area under the receiver operating characteristic (ROC) curve (cf. Agresti, 2002, Swets, 1986). This measure provides information about the degree to which the predicted probabilities are concordant with the observed outcome (see Hosmer and Lemeshow, 2000, for a critical discussion). The logistic regression model predicts the probability of a response as a function of the values of the predictors (e.g., the nine within-trial differences in segment level). For example, let $y=0$ denote the observed response "Louder noise in interval 1" and $y=1$ denote the response "Louder noise in interval 2." The logistic regression equation models the probability $\hat{\pi}$ of y being equal to 1. The predicted response is $\hat{y}=1$ when $\hat{\pi}>n_0$, and $\hat{y}=0$ when $\hat{\pi}\leq\pi_0$, for some cutoff π_0. The *sensitivity* of the model is $P(\hat{y}=1|y=1)$ and the *specificity* is $P(\hat{y}=0|y=0)$. In signal detection theory terms, the sensitivity corresponds to the proportion of hits, and the specificity corresponds to one minus the proportion of false alarms. The sensitivity and the specificity depend on the arbitrary cutoff π_0. For example, a value of π_0 close to 0 maximizes the sensitivity but minimizes the specificity. The area under the ROC curve, which is a plot of sensitivity as a function of (1—specificity), overcomes this limitation because it summarizes the predictive power for all possible cutoffs. In practice, the area under the ROC curve (AUC) is often computed via a Mann–Whitney U type of statistic for all pairs of $y=0$ and $y=1$ trials (Bamber, 1975). Areas of 0.5 and 1.0 correspond to chance performance and perfect performance, respectively. Across the 48 fitted logistic regression models, AUC ranged between 0.55 and 0.89 ($M=0.74$, $SD=0.097$), indicating reasonably good predictive power (Hosmer and Lemeshow, 2000). A repeated-measures analysis of variance (ANOVA) on AUC, with the within-subjects factors (annoyance and loudness) and position μ_H (first interval and second interval) and the between-subjects factor order of tasks (annoyance judgments first and loudness judgments first), showed no significant effects (all $p>0.1$).

B. Results and discussion

1. Temporal weights

The individual temporal weighting patterns are displayed in Fig. 2 separately for each task and each position of the noise with higher mean level. Clear primacy and recency effects for both tasks were observed for listeners 2, 4, 5, 8, and 9, while other listeners showed a more equal weighting pattern, e.g., 7 and 10. The position of the noise with higher mean level did not usually produce strong differences in the weighting patterns, an exception being for listener 1.

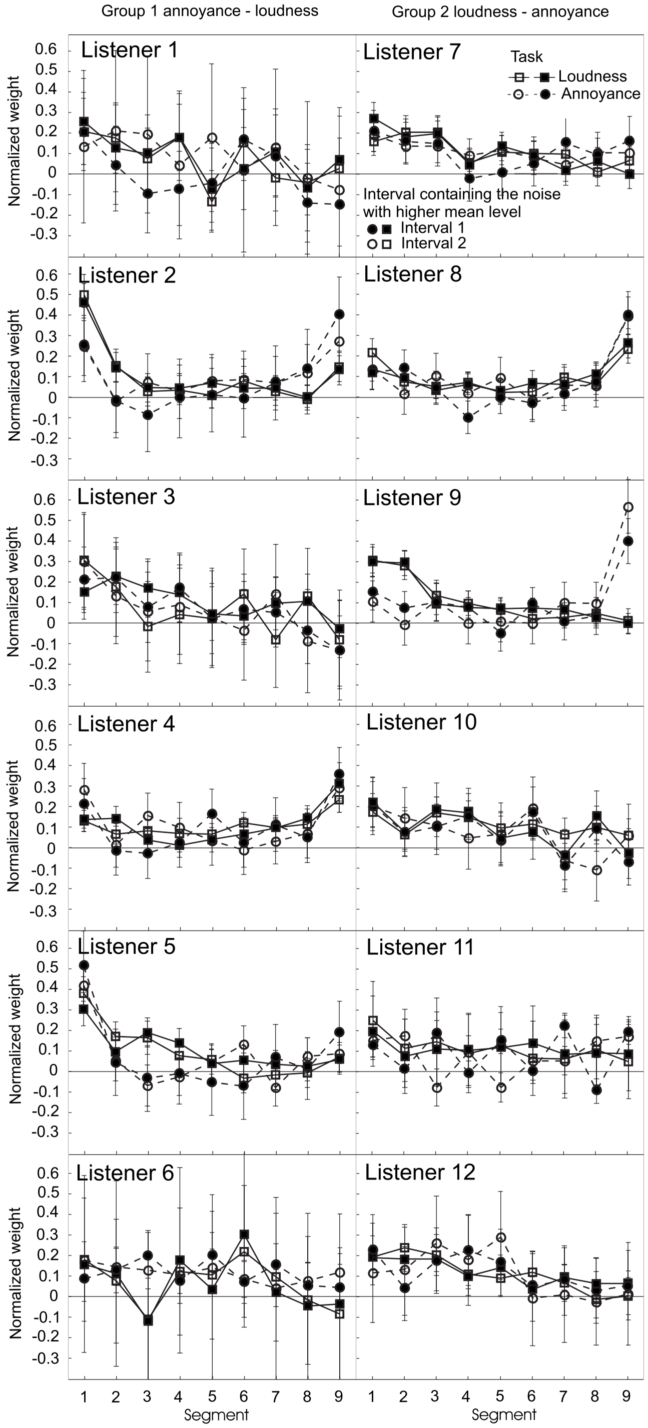

FIG. 2. Experiment 1. Individual relative normalized temporal weights are plotted as a function of segment number. Panels represent listeners. Left column: Group 1 (task order: annoyance-loudness). Right column: Group 2 (task order: loudness-annoyance). Squares and continuous lines: Loudness judgments. Circles and dashed lines: Annoyance judgments. Filled symbols: Interval 1 contained the noise with the higher mean level. Open symbols: Interval 2 contained the noise with the higher mean level. Error bars show the 95%-confidence intervals.

The normalized weights were analyzed via a repeated-measures ANOVA using a univariate approach. The Huynh–Feldt correction for the degrees-of-freedom was used where applicable (Huynh and Feldt, 1976), and the value of the *df* correction factor $\tilde{\varepsilon}$ is reported. The three within-subjects factors were segment (1–9), task (annoyance and loudness), and position μ_H (first interval and second interval). The order of tasks (annoyance judgments first and loudness judgments first) was included as a between-subjects factor. The results are displayed in Table I. There was a significant effect of segment. The Segment × Task interaction was significant, possibly because annoyance and loudness differed in their recency effects (see below). The effect of order of tasks was not significant. The Segment × Order of Tasks interaction and the Task × Order of Tasks interaction were also not significant. Thus, task order had no significant effect on the pattern of weights, indicating that the first task performed in the experiment did not strongly influence listeners' behavior in the second task. Due to the normalization of the weights, the main effect of task was also not significant. Because neither the main effect of the position of the noise with the higher mean level nor any interactions with this factor were significant, the weights were averaged across the two positions for further analysis. Figure 3 displays the mean temporal weights.

Primacy effects for the two conditions were compared in a three-factor repeated-measures ANOVA with the within-subjects factor section (weight assigned to Segment 1 versus average weight assigned to Segments 2–8) and task, and order of tasks as between-subjects factor. The analysis revealed a significant main effect of section $[F(1,10)=37.04, p<0.001]$, confirming that there was a primacy effect in both tasks. The effect of task was not significant ($p=0.09$). The Section × Task interaction was also not significant ($p>0.1$), failing to confirm our hypothesis that primacy effects are stronger for annoyance than for loudness. Neither the main effect of order of tasks nor the Section × Order of Tasks and Task × Order of Tasks interactions were significant (all $p>0.1$). Recency effects for the two conditions were compared in an ANOVA with the within-subjects factor section (weight assigned to Segment 9 versus average weight assigned to Segments 2–8) and task, and order of tasks as a between-subjects factor. The Section × Task interaction was significant $[F(1,10)=5.41, p=0.046]$, confirming the observation of a stronger recency effect for annoyance than for loudness. A *post-hoc* pairwise comparison between the weights assigned to Segment 9 showed a significant difference between the two tasks, $t(11)=3.71$, $p=0.003$ (two-tailed). In the ANOVA, no further main effects and interactions were significant (all $p>0.1$).

Additionally, it was tested whether the temporal weighting patterns for loudness and annoyance differed in uniformity. Therefore, we examined whether the variance of the nine temporal weights differed between loudness and annoyance. For this purpose, the coefficient of variation ($CV=SD/M$) of the nine temporal weights was calculated for each listener and each task. For annoyance, the mean CV was 1.21 ($SD=0.57$). For loudness, the mean CV was 0.97 ($SD=0.45$). A repeated-measures ANOVA with the within-subjects factor task and order of tasks as between-subjects factor showed no significant difference between the CVs of the weights in the two tasks $[F(1,10)=1.55, p=0.241]$. Neither the main effect order of task nor the Task × Order of Tasks interaction was significant (both $p>0.3$).

TABLE I. Results of the ANOVA conducted for the normalized weights from Experiment 1. Within-subjects factors: segment (S), task (T), and position μ_H (P). Between-subjects factor: order of tasks (O). Values in parentheses represent mean square errors. $\underset{\sim}{S}$=subjects. Partial η^2: Variance due to the effect of interest expressed as a proportion of the sum of the error variance and the effect-of-interest variance.

Source	df	F	p	Partial η^2	$\tilde{\epsilon}$
			Between subjects		
Order of tasks (O)	1	3.43	0.094	0.255	
$\underset{\sim}{S}$ within-group error	10	(0.007)			
			Within subjects		
Segment number (S)	8	7.759[a]	0.001	0.430	0.394
Task (T)	1	2.844	0.123	0.202	
Position (P)	1	0.264	0.618	0.011	
S × T	8	2.276[b]	0.045	0.181	0.788
S × P	8	0.450	0.887	0.052	1.0
S × O	8	1.095	0.357	0.107	
T × P	1	2.423	0.151	0.166	
T × O	1	0.122	0.734	0.013	
P × O	1	0.026	0.875	0.012	
S × T × P	8	0.505	0.771	0.043	0.626
S × T × O	8	0.715	0.572	0.064	
S × P × O	8	0.749	0.648	0.075	
T × P × O	1	2.362	0.155	0.191	
S × T × P × O	8	0.639	0.671	0.049	
S×$\underset{\sim}{S}$	80	(0.047)			
T×$\underset{\sim}{S}$	10	(0.005)			
P×$\underset{\sim}{S}$	10	(0.002)			
S×T×$\underset{\sim}{S}$	80	(0.013)			
S×P×$\underset{\sim}{S}$	80	(0.004)			
T×P×$\underset{\sim}{S}$	10	(0.002)			
S×T×P×$\underset{\sim}{S}$	80	(0.006)			

[a] $p < 0.01$.
[b] $p < 0.05$.

2. Model comparisons

The weights analyzed above were estimated via a multiple logistic regression model including the within-trial difference in modulation depth as a predictor. For comparison, a simpler regression model containing as predictors only the nine within-trial differences in segment level was fitted. The two models are nested so that a likelihood-ratio test (Agresti, 2002) can be used for model comparison. This test uses the statistic $-2(LL_{\text{restricted}} - LL_{\text{full}})$, where $LL_{\text{restricted}}$ and LL_{full} are the log likelihood of the model containing only the nine differences in segment level and the model additionally containing the difference in modulation depth, respectively. The test statistic is asymptotically distributed as χ^2_1 because the full model contains one additional free parameter. Of the 48

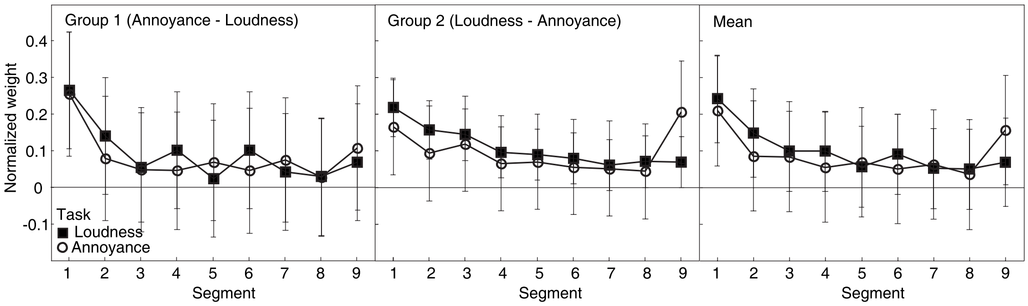

FIG. 3. Experiment 1. Mean relative normalized temporal weights plotted as a function of segment number. Squares: Loudness judgments. Circles: Annoyance judgments. Left panel: Group 1 (task order: annoyance-loudness). Middle panel: Group 2 (task order: loudness-annoyance). Right panel: All listeners (Groups 1 and 2 aggregated). Error bars show the 95%-confidence intervals.

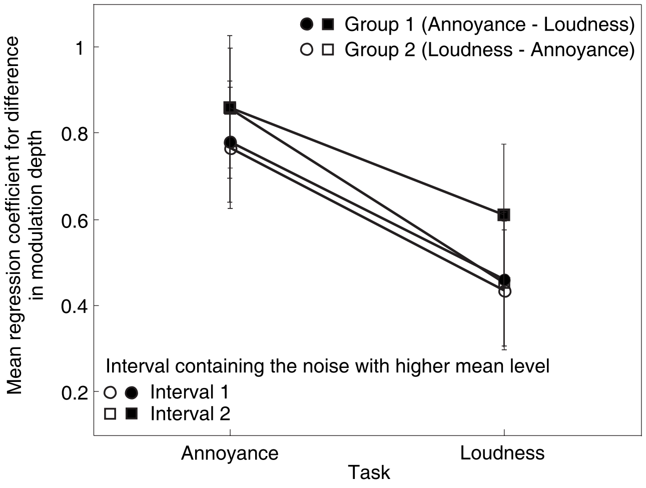

FIG. 4. Multiple logistic regression model for Experiment 1, containing as predictors the within-trial differences in segment level, and the difference in modulation depth. Shown are mean regression coefficients for the difference in modulation depth as a function of task and order of tasks. Circles: Interval 1 contained the noise with the higher mean level. Squares: Interval 2 contained the noise with the higher mean level. Filled symbols: Group 1 (task order: annoyance-loudness). Open symbols: Group 2 (task order: loudness-annoyance). Error bars show the 95%-confidence intervals.

(Listener × Task × Position μ_H) fitted models including both the differences in segment level and the difference in modulation depth, all but 14 fitted the data significantly better than the model not containing the difference in modulation depth ($p<0.05$). Notably, of these 14 models, only 2 were from the annoyance task, compatible with the expected stronger influence of modulation depth on annoyance than on loudness judgments. This pattern was also evident in the regression coefficients for the difference in modulation depth obtained for the full model, which are shown in Fig. 4. These regression coefficients were analyzed via a repeated-measures ANOVA with the within-subjects factors task and position μ_H and the between-subjects factor order of tasks. A significant effect of task confirmed the observation that the difference in modulation depth had a stronger influence for the annoyance than for the loudness task [$F(1,10)=7.40$, $p=0.022$]. The remaining effects were not significant ($p>0.15$).

As discussed in the Introduction, the traditional view is that the loudness and the annoyance of a sound can be predicted from L_{eq} and, at least in the case of annoyance, the modulation depth (e.g., Zwicker, 1991). To test whether a model allowing for a non-uniform temporal weighting of the nine segment levels provides a better account of loudness and annoyance, the fit of two different multiple logistic regression models was compared. The restricted model was compatible with the traditional approach and contained as predictors the difference between L_{eq} in interval 2 and L_{eq} in interval 1 (ΔL_{eq}), and the difference between the modulation depths (i.e., the standard deviation of the nine segment levels) in interval 2 and in interval 1 ($\Delta SD_{\mathrm{levels}}$). The full model also contained the predictors ΔL_{eq} and $\Delta SD_{\mathrm{levels}}$, but additionally the nine within-trial differences in segment level.

For the restricted model, the regression coefficient for the within-trial difference in L_{eq} was generally positive. It was not significantly different from 0 ($p>0.05$, two-tailed) in only five cases. Figure 5 shows the mean data. Across the 48 fitted models, the regression coefficient for the within-trial difference in modulation depth was mostly greater than 0, and in only one case significantly smaller than 0. In 26 of the 48 fitted models, it was significantly different from 0 ($p<0.05$, two-tailed). Figure 5 shows the mean data. As expected, the influence of modulation depth was stronger for annoyance than for loudness.

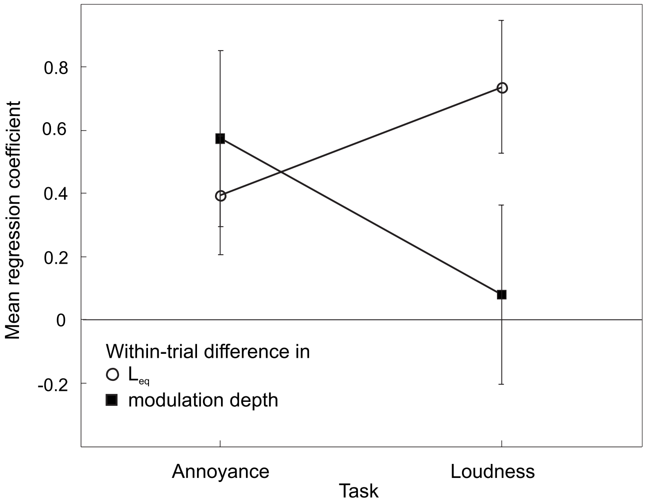

FIG. 5. Mean regression coefficients for a multiple logistic regression model for Experiment 1, containing as predictors the within-trial difference in L_{eq} (circles), and the within-trial difference in modulation depth (squares). Error bars show the 95%-confidence intervals.

The full model containing as predictors the nine segment levels as well as L_{eq} and modulation depth, and the restricted model containing only L_{eq} and modulation depth are nested. The full model has nine additional free parameters so that the test statistic is distributed as χ^2_9. In 30 of the 48 cases, the fit of the full model was significantly better than the fit of the restricted model ($p<0.05$), compatible with the hypothesis that the prediction of loudness and annoyance can be improved by allowing for a non-uniform weighting of the sound pressure level of single temporal portions of the signal. This notion was corroborated by an analysis of the predictive power of the two alternative models in terms of the area under the ROC curve. Figure 6 shows the mean outcome. A repeated-measures ANOVA with the within-subjects factors model (full versus restricted), task, and position μ_H, and the between-subjects factor order of tasks showed a significant effect of model [$F(1,10)=45.92$, $p<0.001$], confirming the higher predictive power of the model with a non-uniform weighting of the segment levels. The remaining effects were not significant ($p>0.1$). Two *post-hoc* ANOVAs were conducted to examine the AUC differences for the loudness task and the annoyance task separately. There was a significant effect of model for both tasks [$F(1,10)=30.99$, $p<0.001$ and $F(1,10)=31.71$, $p<0.001$, respectively], demonstrating that including temporal weights can improve the prediction of both perceptual dimensions.

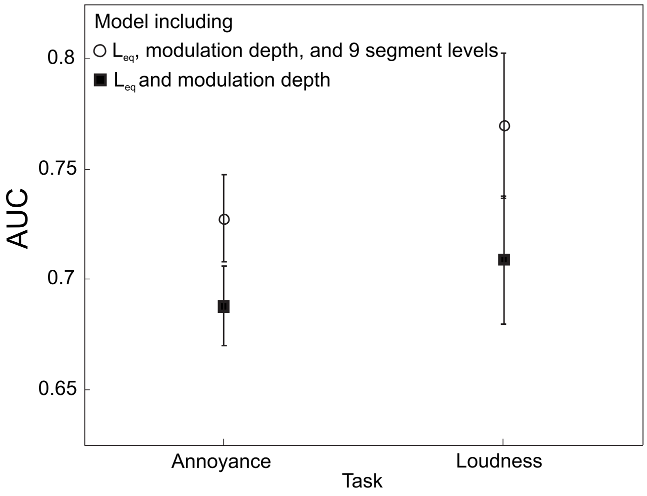

FIG. 6. Mean predictive power of two models for the results of Experiment 1, in terms of the AUC, as a function of task. Squares: Model containing as predictors the within-trial differences in L_{eq} and in modulation depth. Circles: Model containing as predictors the within-trial differences in L_{eq} and in modulation depth, and the nine within-trial differences in segment level. Error bars show the 95%-confidence intervals.

III. EXPERIMENT 2: CAN THE STIMULI USED IN EXPERIMENT 1 BE JUDGED SEPARATELY ACCORDING TO THEIR ANNOYANCE AND LOUDNESS?

In the Introduction, we discussed the potential problem of listeners failing to separate loudness and annoyance. The significant differences between the patterns of temporal weights for loudness and annoyance found in Experiment 1 and the only very weak effect of task order indicated that this problem was negligible in the present study. In Experiment 2, we used a different and more direct approach to examine whether the listeners were capable of independently judging the type of stimuli used in Experiment 1 according to their loudness or their annoyance. The rationale of this experiment was to introduce a parameter previously reported to produce a dissociation between loudness and annoyance. To this end, we presented noises differing in amplitude modulation depth. Level-fluctuating sounds have been suggested to be similar in loudness to steady sounds with the same L_{eq} or N_5 (Berglund *et al.*, 1976; Zwicker and Fastl, 1999, Chap. 16), although some studies reported amplitude-modulated sounds to be slightly louder than steady sounds with the same root-mean-square level (e.g., Zhang and Zeng, 1997; Grimm *et al.*, 2002), while Moore *et al.* (1999) found a small effect in the opposite direction for sinusoidal rather than noise carriers. For annoyance, on the other hand, it has been proposed that sounds differing in modulation depth may differ considerably in their annoyance even if the L_{eq} or the N_5 is constant (Zwicker, 1991; Widmann, 1994; but see Hiramatsu *et al.*, 1983). The method for Experiment 2b was similar to Experiment 1, but we constructed the two noises presented within a trial so that they were similar in loudness while differing in modulation depth. We expected the listeners to perceive these sounds as similar in loudness, but to perceive the sound with the higher modulation depth as more annoying, in line with our assumption that listeners were able to separate the dimensions loudness and annoyance. Additionally, we varied the task order (loudness judgments followed by annoyance judgments or vice versa) just as in Experiment 1 so that a tendency of the listeners to adhere to the type of judgment they had given during the first several hundreds of trials could be detected.

A. Experiment 2a: Comparison of L_{eq} and N_5

We are not aware of studies reporting psychophysical loudness measurements for exactly the same type of stimuli used in our experiments. Therefore, we conducted Experiment 2a as a pretest in order to decide whether N_5 or L_{eq} is more suitable for constructing nine segment noises differing in modulation depth but similar in loudness.

On each trial, two noises fluctuating in level were presented in a two-interval forced-choice task. One of the intervals contained a noise with higher modulation depth. The listeners decided which of two noises presented in a given trial was louder. Within a trial, the stimuli were constructed so that they had either the same L_{eq} or the same N_5.

1. Method

Thirteen listeners (8 women, 5 men, age 22–51 years) participated. None of them had taken part in Experiment 1. Most were psychology students and had no experience in comparable psychoacoustic tasks. All listeners reported normal hearing.

The same apparatus as in Experiment 1 was used. On each trial, two level-fluctuating noises with different modulation depth (small or large) but the same L_{eq} or N_5 were presented. The stimuli were identical to Experiment 1 except for the following differences. On each trial, the level-fluctuating noise with small modulation depth was generated by independently drawing the sound pressure levels of the nine temporal segments from a normal distribution with mean $\mu=65$ dB SPL and $SD=2$ dB. The noise with large modulation depth was generated by independently drawing the nine segment levels from a normal distribution with $\mu=65$ dB SPL and $SD=4$ dB. Subsequently, the mean level of the noise with large modulation depth was adjusted so that either L_{eq} or N_5 (depending on the experimental condition) was identical to that of the noise with small modulation depth. For identical L_{eq}, the level of each segment of the noise with large modulation depth was adjusted by an identical amount (e.g., +1.1 dB). For identical N_5, the level of each segment of the noise with large modulation depth was adjusted by an identical amount so that the highest segment level was identical to the highest segment level for the noise with small modulation depth.[5]

The procedure was the same as before except for the following differences. Each participant received 300 trials with the same L_{eq} and 300 trials with the same N_5. Trials with identical L_{eq} and N_5 were randomly interleaved in each block. The noise with large modulation depth was presented in interval 1 or interval 2 with identical *a priori* probability. Each block consisted of 50 trials. The listeners' task was to decide whether the first or the second noise was louder. The experiment started with a practice block consisting of 40 trials. See Fig. 1(b) for a schematic depiction of a trial. The duration of the experiment was approximately 35 min.

2. Results and discussion

The proportion of trials in which the noise with the large standard deviation was chosen as louder was analyzed. A mean proportion of 0 would indicate that the noise with the large standard deviation was never chosen as louder; a mean proportion of 1 would indicate that the noise with the large standard deviation was always chosen as louder. For noises identical in L_{eq}, the mean proportion was $M=0.50$, $SD=0.09$, indicating that noises with large and small modulation depths were perceived as equally loud. A one-sample t-test showed that this mean proportion was not significantly different from 0.5 [$t(12)=0.13$, $p=0.899$ (two-tailed)]. For noises identical in N_5, the mean proportion was $M=0.28$, $SD=0.08$, indicating that the noise with the smaller modulation depth was perceived as louder. The mean proportion differed significantly from 0.5 [$t(12)=7.64$, $p=0.001$ (two-tailed)]. A repeated-measures ANOVA with the within-subjects factor loudness measure (L_{eq},N_5) showed a significant effect [$F(1,12)=105.45$, $p<0.01$]. The results demonstrate a clear advantage for L_{eq} compared to N_5 for the purpose of constructing noises with different modulation depth but identical loudness. Therefore, L_{eq} was used in Experiment 2b. The results are consistent with the report of Moore *et al.* (1999) that the loudness of sounds with the same rms level is only weakly influenced by amplitude modulation, but see Zwicker and Fastl (1999) for a different claim.

B. Experiment 2b

1. Method

The same listeners as in Experiment 1 participated in this experiment. The division of the listeners into Groups 1 and 2 was the same as before. The same apparatus and essentially the same stimuli and experimental procedure as in Experiment 2a were used. In a two-interval task, listeners selected the louder or more annoying sound. On each trial, two noises with the same L_{eq} (and therefore approximately the same loudness) but with different modulation depths were presented. Noises with small and large modulation depths were generated by independently drawing the sound pressure levels of the nine temporal segments from a normal distribution with mean $\mu=65$ dB SPL and $SD=2$ or 4 dB. See Fig. 1(b) for a schematic depiction of a trial. The level of each segment of the noise with the large modulation depth was displaced by the same amount so that L_{eq} was equal to that of the noise with small modulation depth. Fifteen trials with the large SD in the first interval and the small SD in the second interval and 15 trials with the reverse order of modulation depths were generated and stored before the experiment started. Thus, the listeners evaluated exactly the same set of stimuli according to both their loudness and their annoyance. The same set of 30 trials was used for all listeners.

As already reported, Experiments 1 and 2b were interleaved. Experiment 2b consisted of practice blocks and experimental blocks. In session 1, Group 1 received each of the 30 stored trials ten times in random order and decided which interval contained the more annoying sound. In session 4, Group 1 again received 300 trials, but this time they made loudness judgments. The stimuli were presented in blocks of 50 trials. For Group 2, which started with the loudness judgments, the procedure was analogous.

2. Results and discussion

The proportion of trials in which the noise with the large standard deviation was chosen as louder or more annoying was analyzed. For loudness, the mean proportion was $M=0.52$ ($SD=0.09$), compatible with the hypothesis that noises with the same L_{eq} are perceived as equally loud. A one-sample t-test showed that this mean proportion was not significantly different from 0.5 [$t(11)=0.63$, $p=0.54$]. For annoyance, the mean proportion was $M=0.66$ ($SD=0.08$), indicating that, as expected, the noise with the larger modulation depth was perceived as more annoying. The mean proportion differed significantly from 0.5 [$t(11)=6.79$, $p<0.001$]. A repeated-measures ANOVA with the within-subjects factor task and the between-subjects factor order of tasks showed a significant effect of task [$F(1,10)=45.75$, $p<0.001$, $\tilde{\varepsilon}=1.0$]. Neither the effect of order of tasks [$F(1,10)=2.47$] nor the Task × Order of Tasks interaction [$F(1,10)=0.25$] was significant.

The results of Experiment 2b showed that noises with the same energy-equivalent level but different modulation depths were judged to differ in annoyance but not in loudness. These findings indicate that listeners were capable of separating loudness and their annoyance for the type of stimuli presented in Experiment 1. Thus, Experiment 2 provided further evidence for the appropriateness of the within-subjects design used in Experiment 1.

IV. GENERAL DISCUSSION

The present study examined the temporal weighting of annoyance for broadband noises fluctuating in level. The pattern of temporal weights was compared to the temporal weighting of loudness for the same stimuli and the same listeners. The results of Experiment 2 and the non-significant effect of task order in Experiment 1 showed that the listeners were able to separate loudness and annoyance. Additional evidence for the separability of the perceptual dimensions annoyance and loudness was provided by the goodness-of-fit analyses which showed that in Experiment 1 the variance of the intensity fluctuations (modulation depth) had a clear effect on annoyance but not on loudness.

Consistent with previous studies (e.g., Ellermeier and Schrödl, 2000; Plank, 2005; Oberfeld, 2008b; Pedersen and Ellermeier, 2008), a primacy effect was found for loudness and for annoyance. The listeners assigned higher weight to the level of the beginning of a sound than to its middle portion. The size of the primacy effect did not differ between annoyance and loudness. A significant recency effect (i.e., higher weight assigned to the end than to the temporal center of the sound) was observed for annoyance only. In the study of Pedersen and Ellermeier (2008) a recency effect for loudness was evident in the mean data. However, only three of the five listeners in their experiment showed a recency effect, while all listeners showed a primacy effect. Thus, just as in Experiment 1 of the present study, the primacy effect was stronger than the recency effect. Note that Plank (2005) also found no recency effect in a loudness judgment task where the noise segments were separated by pauses.

As discussed in the Introduction, the processing of the nine segments as serially sorted information represents a potential explanation of the primacy and the recency effects, for example, within the framework of the distinctiveness concept (e.g., Neath *et al.*, 2006). Supporters of this account hypothesize that the difficulty in correctly recalling an item depends on the degree to which it is distinct (or "stands out"). In serially sorted information sets, for example, word lists or the stimuli used in the present study, middle items have two neighboring items. However, end and beginning items only have one neighboring item and therefore are more distinct. It remains for future work to assess whether the temporal weights in loudness and annoyance can be explained in this way.

How can the stronger recency effect for the temporal weighting of annoyance be explained? One potential framework is the "peak-end rule" (Fredrickson and Kahneman, 1993) observed for retrospective evaluations of negative experiences such as painful medical treatments (Redelmeier and Kahneman, 1996) or exposure to aversive sounds (Schreiber and Kahneman, 2000).[6] Kahneman and coworkers found that such judgments are strongly influenced by the worst and the final part of the episode. Thus, a tentative explanation for our observation that the recency effect was significant for annoyance judgments only would be that judging annoyance implies negative emotions while judging the loudness of moderately loud sounds is a "neutral" task and therefore does not elicit a "peak-end" type of retrospective evaluation.

The level profile of stimuli presented in this study was flat because all segment levels within a noise were drawn from the same distribution. For loudness judgments, a gradual increase in level over the first few segments results in a delayed primacy effect: the weights assigned to the attenuated fade-in part are close to zero, and the maximum weight is assigned to the first segment presented at the full level (Oberfeld and Plank, 2005; Oberfeld, 2008a, 2008b). It remains to be shown whether this pattern is paralleled in annoyance judgments.

An important implication of the results of the present study is that technical measures of annoyance and loudness used in noise quantification should consider temporal aspects. Particularly the beginning and ending of a noise should be taken into account more strongly. The goodness-of-fit analyses conducted for Experiment 1 demonstrated that the prediction of both loudness and annoyance can be improved significantly by allowing for a non-uniform weighting of single temporal portions of the signal, rather than assuming that each temporal portion of a sound contributes equally to annoyance and loudness, which is the concept underlying measures like L_{eq} or N_5.

With respect to noise quantification, a limitation of the present study is that the stimuli were shorter than environmental noises, for example, aircraft noise. However, this does not preclude practical applications of our findings. For instance, car alarms frequently use short repeating patterns.[7] Our results indicate that sound designers trying to either increase or decrease the annoyance or loudness of such warning sounds (cf. Suied *et al.*, 2008) should focus on the beginning of the repetitive patterns. Nevertheless, additional research is necessary to clarify whether primacy and recency effects in the temporal weighting pattern of annoyance can be found for longer stimuli.

Finally, it should be noted that Experiment 2a demonstrated that L_{eq} is a better estimate of the loudness of the type of noises used in our experiments than N_5, which is favored by some authors (e.g., Zwicker and Fastl, 1999).

ACKNOWLEDGMENTS

We thank Brian Moore and two anonymous reviewers for helpful comments on earlier versions of this article.

[1] L_{eq}: Energy-equivalent sound pressure level. The sound pressure level of a steady sound that has the same total acoustic energy as a fluctuating sound with the same duration. The equivalent sound pressure level with an A-weighting is referred to as L_{Aeq}.

[2] L_{A}: A-weighted sound pressure level. This rating is based on the 40-phone equal loudness contour and is expressed in dB(A). N_5 is the maximal loudness (frequently estimated via a loudness model) that is reached or exceeded in 5% of the measurement time (i.e., the 95th percentile of the loudness distribution; Zwicker and Fastl, 1999).

[3] In the magnitude estimation procedure we presented noises with the same modulation depth as for Experiment 2b (levels drawn from a normal distribution with $SD=2$ dB). Therefore, the modulation depth was smaller than for the stimuli of the first part of Experiment 1, where the *SD* was 2.5 dB. Thus, due to the effect of modulation depth on annoyance, the stimuli in Experiment 1 were likely slightly more annoying than the stimuli for which the magnitude estimates were obtained.

[4] Besides multiple binary logistic regression, there exist other techniques for weight estimation (e.g., Ahumada and Lovell, 1971; Berg, 1989; Richards and Zhu, 1994), which are all based on a similar decision model (cf. Lutfi and Jesteadt, 2006) and produce similar estimates (Plank, 2005; Tang *et al.*, 2005). See Pedersen and Ellermeier (2008) for a discussion of the advantages of multiple logistic regression.

[5] The stimuli consisted of nine 100-ms segments. For these stimuli, N_5 corresponds to the loudness of the segment with the highest sound pressure level because this loudness is reached during one-ninth of the stimulus duration.

[6] We are grateful to Pieter Jan Stallen for suggesting this explanation.

[7] We thank an anonymous reviewer for suggesting the potential applications discussed in this paragraph.

Agresti, A. (**2002**). *Categorical Data Analysis* (Wiley, New York).

Ahumada, A., and Lovell, J. (**1971**). "Stimulus features in signal detection," J. Acoust. Soc. Am. **49**, 1751–1756.

Anderson, J. R., Bothell, D., Lebiere, C., and Matessa, M. (**1998**). "An integrated theory of list memory," J. Mem. Lang. **38**, 341–380.

Bamber, D. (**1975**). "The area above the ordinal dominance graph and the area below the receiver operating characteristic graph," J. Math. Psychol. **12**, 387–415.

Berg, B. G. (**1989**). "Analysis of weights in multiple observation tasks," J. Acoust. Soc. Am. **86**, 1743–1746.

Berglund, B., Berglund, U., and Lindvall, T. (**1976**). "Scaling loudness, noisiness, and annoyance of community noises," J. Acoust. Soc. Am. **60**, 1119–1125.

Copas, J. (**1989**). "Unweighted sum of squares test for proportions," Appl. Stat. **38**, 71–80.

Dornic, S., and Laaksonen, T. (**1989**). "Continuous noise, intermittent noise and annoyance," Percept. Mot. Skills **68**, 11–18.

Ellermeier, W., and Schrödl, S. (**2000**). "Temporal weights in loudness summation," in *Fechner Day 2000. Proceedings of the 16th Annual Meeting of* *the International Society for Psychophysics*, edited by C. Bonnet (Université Louis Pasteur, Strasbourg, France), pp. 169–173.

Fredrickson, B. L., and Kahneman, D. (**1993**). "Duration neglect in retrospective evaluations of affective episodes," J. Pers. Soc. Psychol. **65**, 45–55.

Grimm, G., Hohmann, V., and Verhey, J. L. (**2002**). "Loudness of fluctuating sounds," Acta. Acust. Acust. **88**, 359–368.

Han, L. A., and Poulsen, T. (**1998**). "Equivalent threshold sound pressure levels for Sennheiser HDA 200 earphone and Etymotic Research ER-2 insert earphone in the frequency range 125 Hz to 16 kHz," Scand. Audiol. **27**, 105–112.

Hellman, R. P. (**1982**). "Loudness annoyance, and noisiness produced by single-tone-noise complexes," J. Acoust. Soc. Am. **72**, 62–73.

Hellman, R. P. (**1984**). "Growth rate of loudness, annoyance, and noisiness as a function of tone location within the noise spectrum," J. Acoust. Soc. Am. **75**, 209–218.

Hellman, R. P. (**1985**). "Perceived magnitude of two-tone-noise complexes: Loudness, annoyance, and noisiness," J. Acoust. Soc. Am. **77**, 1497–1504.

Hellman, R. P., and Meiselman, C. H. (**1988**). "Prediction of individual loudness exponents from cross-modality matching," J. Speech Hear. Res. **31**, 605–615.

Hellman, R. P., and Zwislocki, J. (**1961**). "Some factors affecting the estimation of loudness," J. Acoust. Soc. Am. **33**, 687–694.

Hiramatsu, K., Takagi, K., and Yamamoto, T. (**1983**). "Experimental investigation on the effect of some temporal factors of nonsteady noise on annoyance," J. Acoust. Soc. Am. **74**, 1782–1793.

Hosmer, D. W., Hosmer, T., leCessie, S., and Lemeshow, S. (**1997**). "A comparison of goodness-of-fit tests for the logistic regression model," Stat. Med. **16**, 965–980.

Hosmer, D. W., and Lemeshow, S. (**2000**). *Applied Logistic Regression*, 2nd ed. (Wiley, New York).

Huynh, H., and Feldt, L. S. (**1976**). "Estimation of the Box correction for degrees of freedom from sample data in randomized block and split-plot designs," J. Educ. Stat. **1**, 69–82.

IEC 318 (**1970**). "An IEC artificial ear, of the wide band type, for the calibration of earphones used in audiometry" (International Electrotechnical Commission, Geneva).

Jones, D. M., Macken, W. J., and Nicholls, A. P. (**2004**). "The phonological store of working memory: Is it phonological and is it a store?," J. Exp. Psychol. Learn. Mem. Cogn. **30**, 656–674.

Kortekaas, R., Buus, S., and Florentine, M. (**2003**). "Perceptual weights in auditory level discrimination," J. Acoust. Soc. Am. **113**, 3306–3322.

Kryter, K. D. (**2007**). "Acoustical sensory, and psychological research data and procedures for their use in predicting effects of environmental noises," J. Acoust. Soc. Am. **122**, 2601–2614.

Kuss, O. (**2001**). "A SAS/IML Macro for goodness-of-fit testing in logistic regression models with sparse data," in Proceedings of the 26th Annual SAS Users Group International Conference, Paper No. 265–226.

Kuss, O. (**2002**). "Global goodness-of-fit tests in logistic regression with sparse data," Stat. Med. **21**, 3789–3801.

Kuwano, S., and Namba, S. (**2000**). "Psychological evaluation of temporally varying sounds with L_{Aeq} and noise criteria in Japan," J. Acoust. Soc. Jpn. (E) **21**, 319–322.

Leventhall, H. G. (**2004**). "Low frequency noise and annoyance," Noise Health **6**, 59–72.

Levitt, H. (**1971**). "Transformed up-down methods in psychoacoustics," J. Acoust. Soc. Am. **49**, 467–477.

Lutfi, R. A., and Jesteadt, W. (**2006**). "Molecular analysis of the effect of relative tone level on multitone pattern discrimination," J. Acoust. Soc. Am. **120**, 3853–3860.

Marquis-Favre, C., Premat, E., and Aubrée, D. (**2005a**). "Noise and its effects—A review on qualitative aspects of sound. Part II: Noise and annoyance," Acta. Acust. Acust. **91**, 626–642.

Marquis-Favre, C., Premat, E., Aubrée, D., and Vallet, M. (**2005b**). "Noise and its effects—A review on qualitative aspects of sound. Part I: Notions and acoustic rating," Acta. Acust. Acust. **91**, 613–625.

McFarland, D. J., and Cacace, A. T. (**1992**). "Aspects of short-term acoustic recognition memory: Modality and serial position effects," Audiology **31**, 342–352.

Michaud, D. S., Keith, S. E., and McMurchy, D. (**2008**). "Annoyance and disturbance of daily activities from road traffic noise in Canada," J. Acoust. Soc. Am. **123**, 784–792.

Moore, B. C. J., Vickers, D. A., Baer, T., and Launer, S. (**1999**). "Factors affecting the loudness of modulated sounds," J. Acoust. Soc. Am. **105**, 2757–2772.

Namba, S., and Kuwano, S. (**1979**). "An experimental study on the relation between long-term annoyance and instantaneous judgment of level-fluctuation sound," in *Proceedings Inter-noise '79*, edited by S. Czarnecki (Institute of Fundamental Technological Research Polish Academy of Sciences, Warsaw, Poland), Vol. **II**, pp. 837–842.

Namba, S., and Kuwano, S. (**1980**). "The relation between overall noisiness and instantaneous judgment of noise and the effect of background noise level on noisiness," J. Acoust. Soc. Jpn. (E) **1**, 99–106.

Neath, I., Brown, G. D. A., McCormack, T., Chater, N., and Freeman, R. (**2006**). "Distinctiveness models of memory and absolute identification: Evidence for local, not global, effects," Q. J. Exp. Psychol. **59**, 121–135.

Oberfeld, D. (**2008a**). "Does a rhythmic context have an effect on perceptual weights in auditory intensity processing?," Can. J. Exp. Psychol. **62**, 24–32.

Oberfeld, D. (**2008b**). "Temporal weighting in loudness judgments of time-varying sounds containing a gradual change in level," J. Acoust. Soc. Am. **123**, 3307.

Oberfeld, D., and Plank, T. (**2005**). "Temporal weighting of loudness: Effects of a fade in," in *Fortschritte der Akustik (Advances in Acoustics)—DAGA '05*, edited by Deutsche Gesellschaft für Akustik (DEGA, Berlin), pp. 227–228.

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

Plank, T. (**2005**). "Auditive Unterscheidung von zeitlichen Lautheitsprofilen (Auditory discrimination of temporal loudness profiles),"Ph.D. thesis, Universität Regensburg, Regensburg, Germany.

Postman, L., and Phillips, L. W. (**1965**). "Short-term temporal changes in free-recall," Q. J. Exp. Psychol. **17**, 132–138.

Redelmeier, D. A., and Kahneman, D. (**1996**). "Patients' memories of painful medical treatments: Real-time and retrospective evaluations of two minimally invasive procedures," Pain **66**, 3–8.

Richards, V. M., and Zhu, S. (**1994**). "Relative estimates of combination weights, decision criteria, and internal noise based on correlation coefficients," J. Acoust. Soc. Am. **95**, 423–434.

Schomer, P. D., Suzuki, Y., and Saito, F. (**2001**). "Evaluation of loudness-level weightings for assessing the annoyance of environmental noise," J. Acoust. Soc. Am. **110**, 2390–2397.

Schreiber, C. A., and Kahneman, D. (**2000**). "Determinants of the remembered utility of aversive sounds," J. Exp. Psychol. Gen. **129**, 27–42.

Suied, C., Susini, P., and McAdams, S. (**2008**). "Evaluating warning sound urgency with reaction time," J. Exp. Psychol., Appl. **14**, 201–212.

Surprenant, A. M. (**2001**). "Distinctiveness and serial position effects in tonal sequences," Percept. Psychophys. **63**, 737–745.

Swets, J. A. (**1986**). "Indices of discrimination or diagnostic accuracy: Their ROCs and implied models," Psychol. Bull. **99**, 100–117.

Tang, Z., Richards, V. M., and Shih, A. (**2005**). "Comparing linear regression models applied to psychophysical data," J. Acoust. Soc. Am. **117**, 2597.

WHO (**1999**). "Guidelines for community noise," edited by B. Berglund, T. Lindvall, and D. H. Schwela, World Health Organization, Geneva, http://www.who.int/docstore/peh/noise/giudelines2.html (Last viewed 8/20/08).

Widmann, U. (**1994**). "Zur Lästigkeit von amplitudenmoduliertem Breitbandrauschen (The annoyance of amplitude modulated broadband noises)," in *Fortschritte der Akustik (Advances in Acoustics)—DAGA '94, Aachen*, edited by Deutsche Gesellschaft für Akustik (DEGA, Berlin), pp. 1121–1124.

Zhang, C., and Zeng, F.-G. (**1997**). "Loudness of dynamic stimuli in acoustic and electric hearing," J. Acoust. Soc. Am. **102**, 2925–2934.

Zimmer, K., and Ellermeier, W. (**1996**). "Construction and evaluation of a noise-sensitivity questionnaire," in *Recent Trends in Hearing Research: Festschrift for Seiichiro Namba*, edited by H. Fastl (BIS, Bibliotheks- und Informationssystem der Universität Oldenburg), pp. 163–170.

Zwicker, E. (**1966**). "Ein Beitrag zur unterscheidung von Lautstärke und Lästigkeit (A contribution to the distinction between loudness and annoyance)," Acustica **17**, 22–25.

Zwicker, E. (**1991**). "Ein Vorschlag zur Definition und zur Berechnung der unbeeinflussten Lästigkeit (A proposal for defining and calculating the unbiased annoyance)," Zeitschrift für Lärmbekämpfung **38**, 91–97.

Zwicker, E., and Fastl, H. (**1999**). *Psychoacoustics—Facts and Models* (Springer, Berlin).