Temporal weighting of loudness: Comparison between two different psychophysical tasks

Emmanuel Ponsot, Patrick Susini, Daniel Oberfeld

Temporal weighting of loudness: Comparison between two different psychophysical tasks

Emmanuel Ponsot[a] and Patrick Susini
STMS Laboratory (IRCAM, CNRS, UPMC), 1 place Igor Stravinsky, 75004 Paris, France

Daniel Oberfeld
Experimental Psychology, Johannes Gutenberg-Universität Mainz, 55099 Mainz, Germany

(Received 30 July 2015; revised 4 January 2016; accepted 4 January 2016; published online 21 January 2016)

Psychophysical studies on loudness have so far examined the temporal weighting of loudness solely in level-discrimination tasks. Typically, listeners were asked to discriminate hundreds of level-fluctuating sounds regarding their *global loudness*. Temporal weights, i.e., the importance of each temporal portion of the stimuli for the loudness judgment, were then estimated from listeners' responses. Consistent non-uniform "u-shaped" temporal weighting patterns were observed, with greater weights assigned to the first and the last temporal portions of the stimuli, revealing significant primacy and recency effects, respectively. In this study, the question was addressed whether the same weighting pattern could be found in a traditional loudness estimation task. Temporal loudness weights were compared between a level-discrimination (LD) task and an absolute magnitude estimation (AME) task. Stimuli were 3-s broadband noises consisting of 250-ms segments randomly varying in level. Listeners were asked to evaluate the global loudness of the stimuli by classifying them as "loud" or "soft" (LD), or by assigning a number representing their loudness (AME). Results showed non-uniform temporal weighting in both tasks, but also significant differences between the two tasks. An explanation based on the difference in complexity between the evaluation processes underlying each task is proposed.
 [http://dx.doi.org/10.1121/1.4939959]

[EAS] Pages: 406–417

I. INTRODUCTION

Psychophysical reverse-correlation (Ahumada and Lovell, 1971; Beard and Ahumada, 1998; Ahumada, 2002), also termed perceptual weight analysis (Berg, 1989), has been shown to provide a unique framework for identifying decision strategies underlying various perceptual evaluations (see Dai and Micheyl, 2010; Murray, 2011). In auditory research, this method has been successfully employed to examine the spectral weighting of individual components of complex sounds (e.g., Leibold *et al.*, 2007; Leibold *et al.*, 2009; Jesteadt *et al.*, 2014), the temporal weighting of loudness for time-varying sounds (e.g., Pedersen and Ellermeier, 2008; Oberfeld and Plank, 2011; Ponsot *et al.*, 2013), or spectro-temporal weights (Oberfeld *et al.*, 2012). For the study of temporal loudness weighting, time-varying sounds composed of several temporal portions varying in level randomly and independently from trial-to-trial are typically judged in terms of *global loudness* (i.e., the loudness of the sound as a whole) over several hundred trials. Assuming that listeners' overall loudness judgments are based on a linear combination of each segment level, the relative weight of each portion can then be estimated, for example, using multiple logistic regression (Oberfeld, 2008). The weights thus obtained indicate how strongly the global loudness is impacted when the level of a temporal portion of the sound is changed. For stimuli with a flat level profile (i.e., all temporal portions of the sound are drawn from distributions having the same mean level), these studies consistently showed that the first 100–300 ms receive a higher weight than later portions of the stimulus (i.e., a primacy effect) (Ellermeier and Schrödl, 2000; Oberfeld, 2008; Pedersen and Ellermeier, 2008; Dittrich and Oberfeld, 2009; Rennies and Verhey, 2009; Oberfeld and Plank, 2011; Oberfeld *et al.*, 2012). This means that, for example, a 1-dB increase in the level of the first 100 ms of the sound causes a stronger increase in global loudness than a 1-dB increase in the level of the final 100 ms. Some studies also found a small recency effect (Pedersen and Ellermeier, 2008; Ponsot *et al.*, 2013), but this effects appeared to be weaker than the primacy effect (Dittrich and Oberfeld, 2009; Oberfeld and Plank, 2011). The observed non-uniform temporal weights are an important outcome for research in loudness, because primacy and recency effects are incompatible with the uniform temporal weighting assumed in current indicators of loudness such as L_{Aeq} (i.e., the A-weighted equivalent sound pressure level; ANSI S1.4, 1983) or N_5 (i.e., the 95th percentile of the loudness distribution; Zwicker and Fastl, 1999). For example, a 1-dB increase in the level of the first 100 ms of a sound would result in the same increase in L_{Aeq} as a 1-dB increase in any other temporal portion of the signal if the level and spectral content is equal for the two temporal portions. In this context, Oberfeld and Plank (2011) demonstrated that adding temporal weights to current loudness indicators provides significantly better predictions of the psychophysical data.

Previous studies on the temporal weighting of loudness used sample (or level) discrimination tasks (Berg and Robinson, 1987). In these tasks, the levels of the different temporal portions of the sound (temporal segments) are

[a]Electronic mail: ponsot@ircam.fr

drawn from random distributions. In the case where a one-interval task is used, there are two level distributions, one with a higher mean level [e.g., 61 dB sound pressure level (SPL)], and one with a lower mean level (e.g., 59 dB SPL) (e.g., Dittrich and Oberfeld, 2009). On each trial, the segment levels of the stimulus are all commonly drawn either from the "high" or from the "low" distribution (selected randomly). The listener's task is to decide whether the presented sound was rather "loud" (i.e., originated from the higher level distribution) or rather "soft" (i.e., originated from the lower level distribution). Thus, it corresponds to a one-interval, two-alternative forced-choice (1I, 2AFC) level-discrimination task (i.e., an absolute identification task; Braida and Durlach, 1972). In such a task, responses can be classified as being correct or incorrect so that the accuracy or the sensitivity can be computed. If each temporal segment provides the same amount of information concerning the correct response (which was typically the case in previous experiments), then the "ideal" strategy (maximizing the accuracy) in this task would be to apply identical weights to all temporal portions of a sound (Berg, 1989). However, as discussed above, listeners' weighting strategies are found to be significantly different from this flat "ideal weighting strategy." The use of sample discrimination tasks with two-interval paradigms is also very common (e.g., Rennies and Verhey, 2009; Oberfeld and Plank, 2011), and results in very similar temporal weights to those described for one-interval tasks (Oberfeld and Plank, 2011).

These findings raise the questions: How can the performance in the level-discrimination task be reconciled with loudness judgments that are typically obtained using more "traditional" loudness judgment tasks, such as magnitude estimation (e.g., Stevens, 1956)? That is, to what extent do the weighting patterns observed in level-discrimination tasks reflect the temporal weighting underlying a loudness evaluation in, for example, a magnitude estimation task? The aim of the present study was to address these issues by investigating whether the pattern of temporal weights estimated in a level discrimination (LD) task could also be found in an absolute magnitude estimation (AME) task (Hellman and Zwislocki, 1963).

Our hypothesis regarding this research question was that the temporal weighting pattern observed in the AME task should be similar to the pattern observed in the LD task. Thus, we expected a pronounced primacy effect and probably a weaker recency effect. Indeed, the internal decision variable used by observers in LD tasks as in AME tasks should be based on *global loudness* in both cases. As discussed by Oberfeld and Plank (2011), in the sample discrimination tasks, the listeners are typically instructed to classify each sound as being either "soft" or "loud," that is, to evaluate the *global loudness* of each sound with respect to the loudness of the previous sounds presented in a given block. Thus, it seems reasonable to assume that the subjective quality or sensory continuum on which listeners base their decisions in sample discrimination tasks is *loudness* (cf., Green and Swets, 1966; Durlach and Braida, 1969).

In the present study, two experiments were conducted. In within-subjects designs (to ensure good statistical power in the presence of interindividual differences), temporal loudness weights were measured and compared between AME and LD tasks. A number of experimental constraints were considered to accurately compare the weighting patterns between the two tasks, which deserve to be mentioned here. A first concern was related to the choice of the experimental procedures and the presentation paradigms. Since the AME task is a procedure where the stimuli are presented and evaluated one by one, we opted for an LD task based on a one-interval paradigm. A second experimental concern was that the stimuli had to induce sufficient variability in the magnitude estimates to estimate the weights. For this reason, the segment levels were drawn from normal distributions with comparably large standard deviations ($SD = 5$ dB), while previous studies used *SD*s between 2 and 3 dB. Finally, to minimize undesired psychoacoustical interaction and transient effects between consecutive segments caused by these large level modulations (such as forward masking caused by a loud segment preceding a much softer segment), the segments were rather long (250-ms), and $\cos^2$ functions were used to smooth intersegment amplitude variations (for further details, see the procedure section below). Thus, the stimuli presented in this study were comparably long noises (3 s) fluctuating randomly and slowly (4 Hz) in level. In comparison, previous studies presented shorter sounds (<1.5 s total duration) with shorter temporal segments (typically 100 ms, corresponding to 10-Hz random level modulations) and smaller level fluctuations (e.g., *SD*s of 2 dB).

In the first experiment, the stimuli presented in the two tasks varied around different mean levels. In the LD task, the segment levels were drawn from random distributions with means at 63.5 dB SPL ("low" distribution) and 66.5 dB SPL ("high" distribution) while in the AME task, random distributions with means at 54, 61, 68, and 75 dB SPL were used. We introduced this large range of levels (21 dB) in the AME task to ensure that subjects could easily judge the loudness of the stimuli by using different numbers. The second experiment was conducted to explore whether the results obtained in the first experiment were related, at least in part, to the larger range of stimuli levels used in the AME task compared to the LD task. To answer this question, in the second experiment, the stimuli presented in the AME task varied around the same mean levels (63.5 and 66.5 dB SPL) as in the LD task.

II. EXPERIMENT 1

A. Materials and method

1. Participants

Seven subjects (4 women, 3 men; age 23–31 years) participated voluntarily. All reported normal hearing. They gave their informed written consent according to the Declaration of Helsinki prior to the experiment and were paid for their participation. The participants were naive with respect to the hypotheses under test.

2. Stimuli

The stimuli were white (broadband) noises lasting 3 s. Stationary (constant-intensity) noises were used in the first

session of the AME task, presented diotically with levels of 54, 61, 68, or 75 dB SPL. Otherwise, all the stimuli were level-fluctuating noises made of 12 consecutive 250-ms stationary noise segments with levels drawn independently from normal-truncated distributions ($SD = 5$ dB, restricted to Mean $\pm$ 2.5 *SD*). As mentioned above, the variability in level of these noises was chosen to be sufficiently large to produce variability in participants' judgments, so that the temporal weights could be estimated. While 50-ms linear ramps were imposed on the amplitude envelopes at the onset and the offset of the stimuli, inter-segment level variations were smoothed using 100-ms temporal windows (half-periods of $\cos^2$ functions), to avoid unwanted abrupt changes of sound intensity or temporal loudness masking effects. On each trial, the levels of all segments were randomly and independently drawn from the same normal distribution. In the AME task, the mean of the distribution was 54, 61, 68, or 75 dB SPL, presented with equal probability. In the LD task, the distribution means were 63.5 or 66.5 dB SPL, selected with equal probability.

3. Apparatus

The stimuli were generated at a sampling rate of 44.1 kHz with 16-bit resolution using MATLAB. Sounds were converted using an RME Fireface 800 soundcard and presented diotically through headphones (Sennheiser HD 250 Linear II). Sound level was calibrated using a Brüel & Kjær artificial ear (type 4153, IEC318). Participants were tested individually in a double-walled IAC sound-insulated booth at IRCAM.

4. Procedure

The experiment was divided into two parts, which consisted of two different psychophysical tasks, as described below. The participants performed the two parts one after the other; their order of presentation was counterbalanced between participants. Each part involved several sessions scheduled on different days; part LD refers to the level discrimination (LD) task that consisted of three 1-h sessions, part AME refers to the AME task, which was divided into five 1-h sessions.

In part LD, a 1I, 2AFC procedure was employed. On each trial, a sound was presented with the segment levels drawn either from the "high" or the "low" distribution (see above), randomly chosen with *a priori* equal probability. The participant decided whether the stimulus type was "loud" or "soft." Listeners were explicitly asked to consider the *global loudness* of the stimuli when making their judgment, corresponding to the judgment of the loudness over the entire duration of the sound (Pedersen and Ellermeier, 2008; Ponsot *et al.*, 2013). Each session comprised five blocks of 90 trials each. The answers collected during the first session of this part, which served as a training session, were removed from the analysis. Thus, a total of 900 trials were collected per listener in this task. The participants did not receive any trial-by-trial feedback, but the percentage of correct identifications of the "low"/"high" distributions (defined as corresponding to "soft"/"loud" responses, respectively) were displayed on the computer screen at the end of each block to ensure sustained attention on the task (Ponsot *et al.*, 2013).

In part AME, an AME procedure was used. No standard/modulus corresponding to a certain number was provided. The task was simply to give a number best representing the *global loudness* of each sound, regardless of the numbers assigned to previous stimuli (Hellman, 1976). There was no training session in this part. In the first session of the AME part, the stationary broadband noises were presented to the participants during 160 trials (40 repetitions of each stimulus), in random order. The participants produced magnitude estimates of loudness using their own scale. Level-fluctuating noises were then presented in the four sessions that followed. In these sessions, each trial consisted of a level-fluctuating noise drawn from one of the four defined mean levels (see above). Each "level-fluctuating" session comprised 250 trials, divided into three blocks (90-80-80 trials; the 10 first estimates of the first block served as a training and were removed from the analysis). Thus, a total of 960 "level-fluctuating" trials were collected per listener in the AME task (240 trials per mean level and listener).

5. Fitting loudness functions

The magnitude estimates provided by each participant during the AME task were fitted with simple loudness functions. First, the levels (in dB SPL) of the stimuli (both stationary and level-fluctuating noises) presented in this part were converted into equivalent pressure units (in Pascals) using the following formula:

$$p_{eq} = p_0 10^{L_{eq}/20}, \tag{1}$$

where $p_0 = 20\,\mu$Pa and L_{eq} denotes the energy-equivalent sound pressure level of the entire stimulus (in dB). The magnitude estimates of the listeners were then fitted individually using a power function, which represents one of the most simple loudness functions (cf., Suzuki and Takeshima, 2004; Oberfeld *et al.*, 2012)

$$E = k p_{eq}^{\alpha}, \tag{2}$$

where E is the magnitude estimate (i.e., number) produced by the participant, and p_{eq} is the equivalent pressure of the stimulus. The constants k and α were estimated by non-linear regressions for each listener, for stationary and level-fluctuating noises separately. Thus, a total of 14 regressions were conducted. Individual loudness exponents α estimated by the regressions conducted on level-fluctuating noises estimates were used afterwards in the decision models to estimate temporal weights in both the AME and the LD tasks (see below).

6. Decision models

Multiple regression analyses were used to estimate the temporal weights in the two tasks. In previous studies, the predictors used in the models were simply based on the sound pressure level of the temporal segment levels (e.g.,

TABLE I. Parameters of the individual loudness functions (α, k) estimated by fitting non-linear power functions to the numbers assigned to both constant and level-fluctuating noises in the AME task of experiment 1. Columns represent listeners (S1–S7). For each fitted model, ordinary-R^2 obtained by the regression is indicated as a measure of goodness-of-fit.

Experiment 1										
		S1	S2	S3	S4	S5	S6	S7	Average	SD
Constant noises	k	31.73	21.48	206.27	16.64	2.48	24.39	106.29	**58.47**	73.33
	α	0.53	0.45	0.57	0.34	0.28	0.72	0.62	**0.50**	0.16
	R^2	0.78	0.75	0.90	0.80	0.87	0.87	0.88	**0.83**	0.06
Fluctuating noises	k	19.2	10.16	122.71	14.81	9.89	10.34	97.55	**40.67**	48.12
	α	0.36	0.29	0.45	0.33	0.2	0.56	0.61	**0.39**	0.15
	R^2	0.66	0.78	0.86	0.64	0.61	0.77	0.90	**0.74**	0.11

Oberfeld and Plank, 2011). However, because of the large variations imposed on the segment levels of the stimuli in the present study, we decided to use predictors based on *loudness* in order to match more accurately the human perceptual intensity scale. The loudness of each segment N_i was estimated using the individual loudness exponents α obtained from the regressions conducted on the estimates attributed to level-fluctuating noises in the AME task (the individual values of α were taken from the fitted level-fluctuating noises loudness functions are reported in Table I)

$$N_i = p_i^{\alpha}, \tag{3}$$

where p_i is the equivalent pressure (in Pa) of the ith segment, computed according to Eq. (1). Thus, 12 predictors, corresponding to the loudness values of the 12 segments, were used in the regression models to estimate the weights in the two tasks. In the AME task, the dependent variable (DV) was the magnitude estimate (number) given by the participants to evaluate global loudness. In the LD task, the DV was the binary response ("soft" or "loud") entered by the participants to evaluate global loudness.

In the AME task, estimates were assumed to be a linear combination of the loudness values of the 12 segments. Segment levels L_i (in dB) were first converted into pressure units p_i (in Pa) using Eq. (1). Second, using Eq. (3), we estimated their equivalent loudness N_i. Therefore, the estimate (E) could be expressed as the linear combination of each segment loudness N_i

$$E(\mathbf{N}) = \sum_{i=1}^{12} w_i N_i + c, \tag{4}$$

where $\mathbf{N}$ is the vector of segment loudness values, N_i is the loudness of segment i, w_i is the perceptual weight assigned to segment i, and c is a constant.

The decision model chosen to account for the LD task was similar to previous studies on temporal weighting of loudness (e.g., Oberfeld and Plank, 2011), except that the loudness values rather than the sound pressure levels of the segments were used as predictors. The loudness of each segment (N_i) was estimated using Eq. (3), i.e., using the exponents inferred from the estimates given to fluctuating-noises in the AME task. The decision variable D was also assumed to be a linear combination of the loudness of the 12 segments,

$$D(\mathbf{N}) = \sum_{i=1}^{12} w_i N_i + c, \tag{5}$$

where $\mathbf{N}$ is the vector of segment loudness values, N_i corresponds to the loudness of the ith segment, the w_i are the perceptual weights, and c is a constant. The model assumes that a listener responds that the noise presented on a given trial was loud rather than soft if $D(\mathbf{N}) > 0$, and that

$$p(\text{"}loud\text{"}) = \frac{e^{D(\mathbf{N})}}{1 + e^{D(\mathbf{N})}}, \tag{6}$$

which corresponds to a logistic regression model (McCullagh and Nelder, 1989).

These two decision models are strictly identical except that the dependent variable is continuous in the first model (AME task) and binary in the second model (LD task). Multiple logistic regression was thus used to estimate the temporal weights in the LD task (Pedersen and Ellermeier, 2008; Oberfeld and Plank, 2011; Ponsot *et al.*, 2013), and multiple linear regression was used in the AME task. Regressions were conducted separately for each listener and task. For each task, the twelve regression coefficients w_i were taken as the twelve temporal weights. The weights were normalized individually so the sum of their absolute values was 1 (Kortekaas *et al.*, 2003). Statistical analyses were conducted with R (R Core Team, 2015)

B. Results

1. Loudness functions

Magnitude estimates were obtained for both stationary and level-fluctuating sounds in the AME task of experiment 1 for each listener. As an example, the results obtained for one subject (S7) are reported in Fig. 1. Overall, subjects were well able to produce numbers not only reflecting the mean level of the stimuli but also reflecting level-modulations that were introduced: four overlapping scatterplots could be observed for each subject. Power functions were used to fit the magnitude estimates of the stationary noises and the level-fluctuating noises separately (see above). These fits correspond to the black line and the grey line as plotted in Fig. 1 for S7, respectively. The parameters of each individual power function, estimated using non-

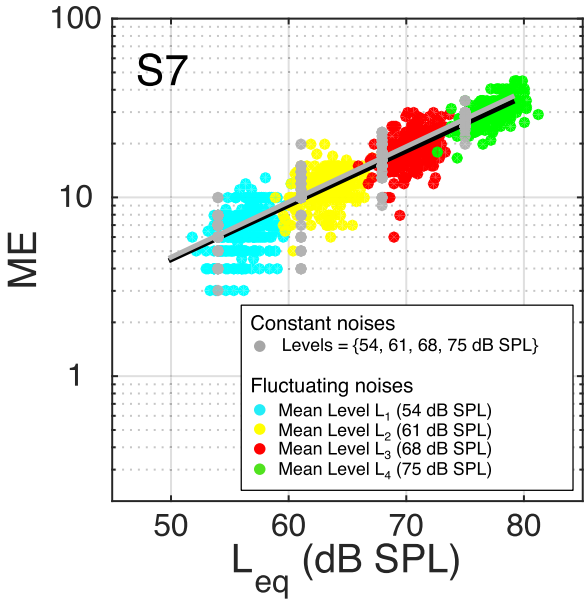

FIG. 1. (Color online) Raw magnitude estimates produced by one subject (S7) in the AME task of experiment 1. Subjects evaluated the global loudness of stationary noises and level-fluctuating noises at different mean levels. Estimates are plotted on a log-scaled y-axis as a function of the L_{eq} of the stimuli (asymmetric distributions of cloud of dots around the levels of stationary noises can thus be observed since the L_{eq} is dominated by segments with higher sound pressure level). Separate power functions were fitted to the MEs for constant noises estimates (grey line) and for level-varying noises estimates (black line). Parameters of the loudness functions are reported in Table I for the different subjects.

linear least square fits, are reported in Table I. Except for S5, who deliberately and notably changed his scale between the first (stationary noises) and the following sessions in the AME task (level-fluctuating noises) (indeed, the subject told the experimenter he wanted to use another scale with greater numbers—as it can be seen from the change of intercept reflected by the parameter k in Table I), similar loudness function parameters (k and α) were obtained for stationary and level-fluctuating noises. The goodness of fit of these non-linear regressions, evaluated in terms of the proportion of variance accounted for (R^2), was reasonably high. It was not significantly higher for stationary than for level-fluctuating noises [$t(6) = 2.286$, $p = 0.062$]. Loudness exponents α (see Table I) were in line with the values reported in the literature for white noise stimuli (e.g., Canévet *et al.*, 2003; Teghtsoonian *et al.*, 2005). These exponents were smaller for level-fluctuating noises than for stationary noises for every subject, leading to a significant difference [$t(6) = 3.911$, $p = 0.008$]. Because the range of level was slightly larger for fluctuating noises than for stationary noises due to the random distributions (as can be seen in Fig. 1), this result is consistent with the reported outcome that the exponent of the loudness function becomes smaller when the range of stimuli level variation is increased (for a review, see Arieh and Marks, 2011). Finally, because the subjects were free to use their own response scale in the AME procedure, comparably large inter-individual differences were found regarding the multiplicative constant k (see the values reported in Table I), compatible with the literature (Hellman and Meiselman, 1988). Values of k were smaller for fluctuating noises than for stationary noises, except for S5 who deliberately changed his scale.[1]

2. Temporal weighting patterns

As explained above, the exponent α of the loudness functions of fluctuating noise obtained for each subject was fed into the regression models to estimate the temporal weights in each task. The temporal weights were then obtained for the seven subjects in the two tasks. The averaged temporal weights obtained in the two tasks are presented in Fig. 2. As noted above, an "ideal observer" would apply uniform weights to the 12 segments (at least in the LD task), because each temporal segment element provides the same amount of information concerning the "correct" response (Berg, 1989). The data presented in Fig. 2 revealed significant deviations from this uniform weighting pattern, as indicated by the confidence intervals of the weights that do not contain the horizontal black line corresponding to uniform weighting. Moreover, the weighting patterns obtained in the two tasks were rather different. In the AME task, a clear primacy effect but no recency effect was observed whereas in the LD task a small recency and a weaker primacy effect could be noticed.

The normalized weights were analyzed with a repeated-measures analysis of variance (rmANOVA) using a univariate approach, with the correction for the degrees of freedom of Huynh and Feldt (1976) where applicable. The Huynh-Feldt correction factors ($\tilde{\varepsilon}$) are reported. Unless otherwise specified, all the tests were two-tailed and used a probability level of 0.05 to test for significance. Effect sizes are reported using partial eta-squared, η_p^2. The within-subjects factors were segment and task. The effect of segment was significant [$F(11, 66) = 3.509$, $p = 0.032$,

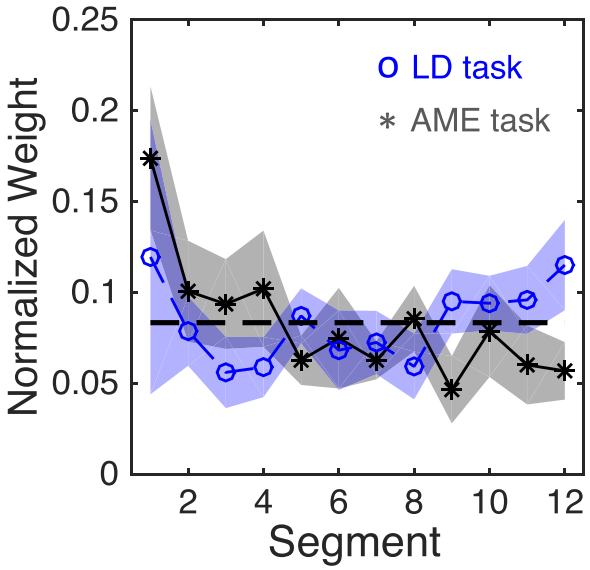

FIG. 2. (Color online) Mean normalized temporal weights for both tasks of experiment 1, presented as a function of the segment position (1–12) are used for the level discrimination (LD) task and asterisks are used for the absolute magnitude estimation (AME) task. Shaded areas correspond to 95% confidence intervals. The horizontal dotted line represents the uniform temporal weighting pattern of an "ideal observer" in the LD task.

TABLE II. Goodness of fit of full models (containing as predictors the L_{eq} and the 12 temporal weights) and restricted models (containing only the L_{eq} as predictor) used to predict individual loudness judgments obtained in each task of experiment 1. Different indexes were employed to compare the models in the two tasks: R^2 for the AME task, AUC for the LD task. Outputs of likelihood-ratio tests (AUC) and F-tests (R^2) are reported to indicate significant improvements of the full model containing temporal weights as compared to the restricted model in each case. Columns represent listeners (S1–S7).

Experiment 1										
Goodness-of-fit index (Task)	Model	S1	S2	S3	S4	S5	S6	S7	Average	SD
AUC (LD Task)	restricted	0.881	0.847	0.841	0.878	0.882	0.877	0.922	**0.876**	0.027
	full	0.892	0.869	0.903	0.884	0.890	0.889	0.937	**0.895**	0.022
	p-value	0.003	0.001	<0.001	0.335	0.071	0.001	0.001		
R^2 (AME Task)	restricted	0.652	0.777	0.851	0.638	0.608	0.761	0.895	**0.740**	0.111
	full	0.665	0.782	0.869	0.641	0.614	0.767	0.901	**0.749**	0.113
	p-value	0.001	0.063	<0.001	0.756	0.268	0.013	0.001		

$\eta_p^2 = 0.369$, $\tilde{\varepsilon} = 0.30$], showing that the segments did not receive a uniform weighting. A significant segment × task interaction was found [$F(11, 66) = 5.014$, $p < 0.001$, $\eta_p^2 = 0.455$, $\tilde{\varepsilon} = 1.0$], supporting the view that a different weighting was applied in the two tasks. As can be observed in Fig. 2, larger weights were applied to the first segments of the stimuli in the AME task, while moderately larger weights were applied to the first segment and to the final three segments in the LD task. Due to the weights normalization, there was of course no significant effect of task [$F(1, 6) = 1.000$, $p = 0.336$, $\eta_p^2 = 0.143$]. Additional ANOVAs with segment as a within-subjects factor were performed for each task separately. For the AME task, a significant effect of segment confirmed the presence of non-uniform weights [$F(11, 66) = 6.781$, $p < 0.001$, $\eta_p^2 = 0.531$, $\tilde{\varepsilon} = 0.83$]. For the LD task, the effect of segment was not significant [$F(11, 66) = 2.018$, $p = 0.149$, $\eta_p^2 = 0.252$, $\tilde{\varepsilon} = 0.27$]. One may notice the unexpected higher weight obtained for the fifth segment compared to other segments situated in the middle section of the stimulus. We have no explanation for this result.

Deviations from the flat weighting pattern were quantified on an individual basis by computing the coefficient of variation CV of the 12 weights (SD/M), for each listener and each task (see Oberfeld and Plank, 2011). The mean CV was lower in the LD task ($M = 0.45$, SD = 0.27) than in the AME task ($M = 0.54$, SD = 0.16), indicating slightly more uniform weights in the LD task, but the difference was not significant [$t(6) = 1.058$, $p = 0.331$]. One subject (S3) showed particularly strong primacy effects in the two tasks, and higher CVs than the remaining subjects.

3. Predictive power of the decision models

The goodness of fit of the decision models was evaluated using the area under the ROC curve (AUC) for the models of the LD task, which is an index of the predictive power of the logistic regression model (Dittrich and Oberfeld, 2009). The proportion of variance (R^2) accounted for was used in the AME task. Fair to good model predictions were found for the two tasks: for the LD task, AUC ranged from 0.81 to 0.93 ($M = 0.87$, SD = 0.036). For the AME task, R^2 ranged from 0.61 to 0.90 ($M = 0.74$, SD = 0.11).

4. Increased predictive power by including temporal weights

The benefit of using estimated temporal weights to predict loudness was evaluated in the two tasks. This was done by comparing different models to predict the present results: Restricted models containing only L_{eq} as predictor were compared with full models containing L_{eq} plus the twelve temporal weights as predictors (see Dittrich and Oberfeld, 2009; Oberfeld and Plank, 2011). The results are reported in Table II.

In the LD task, the AUC of the full model was significantly higher than the AUC of the restricted model, [$t(6) = 2.607$, $p = 0.040$]. Although this difference in AUC is small, Cohen's d_z (Cohen, 1988) indicates a large effect size ($d_z = 0.985$). Individual likelihood-ratio tests were conducted to compare the goodness-of-fit of the full and the restricted models (see Oberfeld and Plank, 2011). A significantly better goodness-of-fit ($p < 0.05$) was obtained with the full model for all subjects but S4 and S5, who correspond to those having the more flat weighting patterns (i.e., smaller CVs).

In the AME task, the R^2 of the full model was also significantly higher than the R^2 of the restricted model [$t(6) = 4.098$, $p = 0.006$]. The effect size was large, $d_z = 1.549$. F-tests on individual data showed that a significantly better goodness-of-fit was obtained with the full model for subjects S1, S3, S6, and S7 ($p < 0.05$).

Overall, these results indicate a significant benefit of using the temporal weights to predict judgments both in the LD task and in the AME task. However, the AUC and R^2 only increased by small factors (2% and 1%, respectively). The likely reason for the small improvement by including temporal weights in the AME task is the large variation in overall loudness due to the variation of mean level from 54 to 75 dB SPL. Indeed, this 21 dB variation in level, which is completely captured by the L_{eq}, accounts for the greatest part of the variance of the magnitude estimates. Even with the rather large SD of the level perturbations, the variation caused by the level fluctuations around the mean segment level is of course weaker than the effect of the mean level. To examine the "real" improvement of considering temporal weights in the models for the AME task (by setting aside this large variation in mean level), additional weight analyses were conducted separately per mean level in the AME task.

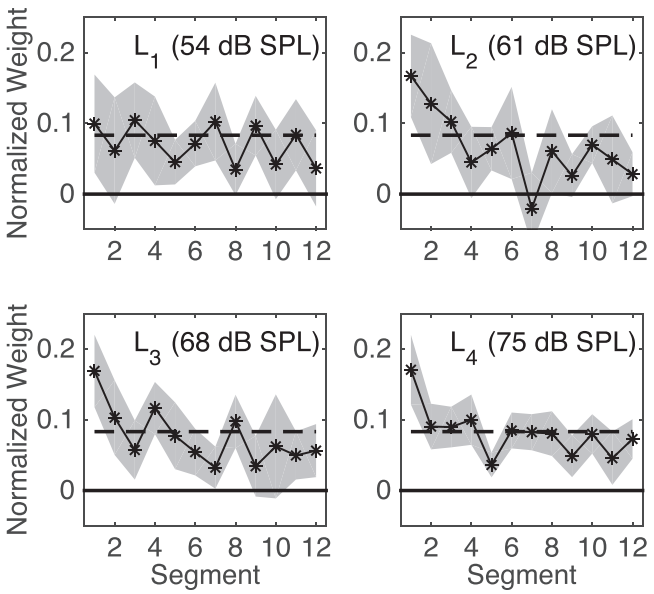

FIG. 3. Normalized averaged temporal weights for the AME task of experiment 1, from the additional analyses conducted at each mean level separately.

5. Additional analyses per mean level in the AME task

We conducted separate regressions for each mean level with the results obtained in the AME task. The mean temporal weighting patterns thus obtained for each level are presented in Fig. 3.

An rmANOVA indicated a significant effect of segment [$F(11,\ 66) = 4.900,\ p < 0.001,\ \eta_p{}^2 = 0.450,\ \tilde{\varepsilon} = 0.97$]. The level × segment interaction was not significant [$F(33,\ 198) = 1.522, p = 0.071, \eta_p{}^2 = 0.203, \tilde{\varepsilon} = 0.71$]. These results support what was observed in the overall analysis. The segments did not receive a uniform weighting, but rather a primacy effect was observed. Figure 3 shows that this weighting process depended descriptively (although not significantly) on the mean level: the weighting patterns estimated at 61, 68, and 75 dB SPL exhibit primacy effects, while a flat pattern was observed at 54 dB SPL, i.e., the lowest mean level employed in the AME task.

We then compared the goodness of fit of the full and restricted models based on these weights obtained for each mean level. The proportions of variance accounted for by the models are reported at each mean level in Table III. Of course, the mean R^2 values are this time much lower than in the overall analysis presented above, but the data show large increases in R^2 when adding the temporal weights to the decision models (107%, 196%, 57%, and 31% for each of the mean levels in ascending order).

C. Discussion

In experiment 1, the temporal weights of loudness were measured for the same participants in two psychophysical tasks, an LD task and an AME task. The loudness exponents estimated for each observer in the AME task were employed in the decision models to estimate the temporal loudness weights. Especially for the AME task, where the mean sound pressure level varied across a range of 21 dB, this procedure based on individual loudness values is in our view superior to the analyses based on sound pressure level used in most previous studies. Descriptively, additional separate analyses conducted in the AME task for the different mean levels showed slightly more uniform temporal weights at low levels. However, the effect of mean level on the weights was not significant and therefore, this aspect remains to be specifically addressed in future studies.

Overall, the data from experiment 1 show that listeners assigned significantly non-uniform temporal weighting patterns in both tasks. This is an interesting finding because the sound duration (3 s) was considerably longer than in previous studies. Contrary to our hypothesis, there was a significant difference between the weighting patterns in the LD and the AME task. On average, participants assigned higher weights to the first three segments compared to later segments in the AME task (i.e., a primacy effect). In the LD task, we observed both a primacy effect and a recency effect. How could these different temporal weighting strategies be explained? One potential explanation is based on the difference between the stimuli presented in the two tasks. As explained above, we deliberately presented a much larger variation in mean level in the AME task than in the LD task, to ensure that subjects could easily do the task. The means of the four level distributions were separated by 7 dB, which is more than the *SD* of 5 dB, and covered a range of 21 dB. In contrast, in the LD task the means of the two level distributions differed by only 3 dB. For this reason, one could argue that the participants were influenced by the range of level variations of the stimuli. To test the hypothesis that the

TABLE III. Comparison between full and restricted models at each mean level in the AME task of experiment 1. The full models contained temporal weights inferred from the separate analyses of the estimates obtained at each mean level (54, 61, 68, and 75 dB SPL), in addition to the L_{eq}. The last column indicates the number of subjects for which the full model explained a significantly higher proportion of the variance (R^2) than the restricted model.

Level	Model	R^2 Mean	SD	min	max	Full vs rest (cases out of 7 where $p < 0.05$)
L1 (54 dB SPL)	restricted	**0.068**	0.041	0.013	0.119	
	full	**0.139**	0.071	0.067	0.226	3
L2 (61 dB SPL)	restricted	**0.043**	0.030	0.011	0.095	
	full	**0.127**	0.055	0.056	0.211	3
L3 (68 dB SPL)	restricted	**0.138**	0.052	0.082	0.207	
	full	**0.217**	0.074	0.120	0.331	5
L4 (75 dB SPL)	restricted	**0.217**	0.103	0.065	0.363	
	full	**0.285**	0.126	0.117	0.472	3

different weighting patterns can be attributed to the different range of segment levels presented in the two tasks, we conducted a second experiment presenting exactly the same stimuli in the AME as in the LD task.

III. EXPERIMENT 2

A. Materials and method

1. Participants

This experiment was conducted on a new group of seven subjects (4 women, 3 men; age 20–31 years), who participated voluntarily and were naive with respect to the hypotheses under test. All participants reported normal hearing. They gave their informed written consent according to the Declaration of Helsinki prior to the experiment and were paid for their participation.

2. Stimuli and apparatus

All stimuli presented in this experiment were constructed exactly as those of the LD task of experiment 1 (see above), both for the AME and the LD task. Thus, the segment levels of the fluctuating stimuli were always drawn from normal distributions with means equal to 63.5 dB SPL ("low" distribution) or 66.5 dB SPL ("high" distribution). Stationary noises were also presented at the end of the experiment at four different mean levels: 60.0, 62.5, 65.0, and 67.5 dB SPL. The apparatus was the same as in experiment 1.

3. Procedure

The experiment comprised six 1-h sessions scheduled on different days. One session consisted of four 90-trial blocks, where two blocks were assigned to the AME task and the two remaining blocks to the LD task. The blocks were presented such that participants alternated between the two tasks (e.g., block1 = LD, block2 = AME, block3 = LD and block4 = AME). The type of the first block was also alternated between sessions. On each trial, a sound was presented with the segment levels drawn either from the "low" or the "high" distribution (see above), randomly chosen with *a priori* equal probability. The procedures were the same as in experiment 1: in the LD blocks, the participant had to decide whether the sound was "soft" or "loud" while in the AME blocks, they had to give a number best representing the *global loudness* of each sound. Before the experiment, subjects were specifically informed that the variation in loudness between the stimuli would not be very large, so that in the AME task, they should select as many different numbers as possible to accurately capture the small variation in loudness from trial to trial. In order to have similar experimental conditions between the two tasks, the participants did not receive feedback on a trial-by-trial or on a block-by-block basis. The results of the first session, which served as a training session, were removed from the analysis. Thus, a total of 900 trials were collected per listener and per task. At the end of the last session, constant broadband noises were presented to the participants in an AME task, which consisted of 40 trials (10 repetitions of each of the four levels) to measure their loudness function with stationary stimuli. In this final part, participants were asked to use the same scale as they used during previous sessions.

4. Fitting loudness functions and decision models

The same fitting procedure as in experiment 1 was used to fit loudness functions, both for constant and fluctuating noises. The decision models were strictly identical to those of experiment 1.

B. Results

1. Loudness functions

Overall, the data show that the subjects had no problems to do the AME task and produced many different numbers to evaluate the global loudness of fluctuating noises with very similar mean levels. As an example, the magnitude estimates given by one participant (S8) in the AME task (for both fluctuating and constant noises) are plotted in Fig. 4. For every participant, the parameters of the loudness functions fitted to his estimates are reported in Table IV.

Loudness exponents α ranged between 0.30 and 0.80 for both stationary and level-fluctuating noises (see Table IV). The slope was not significantly different between stationary and level-fluctuating noises [$t(6) = 0.909$, $p = 0.399$]. On average, these exponents were very close to those measured in the first experiment, with a mean value close to 0.5. As in experiment 1, the values of the parameter k were very different between participants (who were allowed to use their own scale). The proportion of variance accounted for R^2 was significantly higher for stationary noises than for level-fluctuating noises [$t(6) = 4.523$, $p = 0.004$].

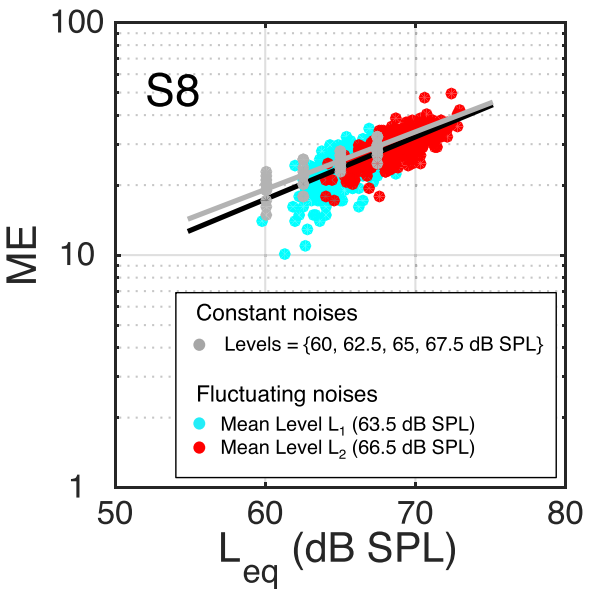

FIG. 4. (Color online) Raw magnitude estimates produced by one subject (S8) in the AME task of experiment 2. Same format as Fig. 1.

TABLE IV. Experiment 2, same format as Table I.

		Experiment 2								
		S8	S9	S10	S11	S12	S13	S14	Average	SD
Constant noises	k	133.11	100.34	375.5	35.42	21.18	89.85	21.28	**110.95**	124.4
	α	0.50	0.30	0.60	0.57	0.45	0.63	0.43	**0.49**	0.11
	R^2	0.74	0.81	0.66	0.47	0.73	0.77	0.18	**0.62**	0.22
Fluctuating noises	k	141.76	105.45	668.15	23.48	20.72	72.54	21.67	**150.54**	233.00
	α	0.54	0.35	0.80	0.48	0.47	0.58	0.48	**0.52**	0.14
	R^2	0.65	0.49	0.43	0.31	0.42	0.41	0.15	**0.40**	0.16

2. Temporal weighting patterns

The averaged weighting patterns obtained for the two tasks, shown in the left panel of Fig. 5, were similar to what was observed in the first experiment, with a clear primacy effect in the AME task while in the LD task, slightly higher weights were assigned to both the first and the last segments compared to the segments in the middle. An rmANOVA with the within-subjects factors segment and task showed no significant effect of segment [$F(11, 66) = 2.418$, $p = 0.11$, $\eta_p^2 = 0.287$, $\tilde{\varepsilon} = 0.22$]. However, the significant segment $\times$ task interaction [$F(11, 66) = 4.480$, $p = 0.026$, $\eta_p^2 = 0.427$, $\tilde{\varepsilon} = 0.22$] indicated that a different weighting was applied in the two tasks.

On average, the temporal weighting patterns were rather flat: as in experiment 1, small CVs were measured both in the AME task ($M = 0.40$, SD $= 0.17$) and in the LD task ($M = 0.65$, SD $= 0.53$). There was no significant difference between the two tasks [$t(6) = 1.144$, $p = 0.296$]. One subject (S9) adopted a very specific strategy in the LD task by exclusively considering the three last segments of the stimuli to judge global loudness (for S9 in the LD task, CV $= 1.792$). This weighting pattern differed strongly from the weights assigned by the remaining participants, as indicated by the large confidence interval obtained for the weight on the last segment in the LD task (see Fig. 5, left panel). After the end of the experiment, this participant told the experimenter that he had consciously and deliberately used this particular strategy in the LD task, although he was aware that the task was to judge the *global loudness* of the sounds similarly in the two tasks. However, he had no clear explanation why he adopted this strategy. The averaged weighting patterns when this participant was excluded are presented in the right panel of Fig. 5, showing weighting patterns even more similar to the weights obtained in experiment 1. In order to check whether the difference between the weighting patterns in the two tasks mainly relied on listener S9, we performed a second rmANOVA without this participant. The analysis still provided a significant segment $\times$ task interaction [$F(11, 55) = 5.269$, $p = 0.001$, $\eta_p^2 = 0.43$, $\tilde{\varepsilon} = 0.45$]. Also, the effect of segment was now significant [$F(11, 55) = 3.770$, $p = 0.012$, $\eta_p^2 = 0.513$, $\tilde{\varepsilon} = 0.58$].

Additional rmANOVAs conducted on each task separately without the data of S9 confirmed the presence of non-uniform temporal weights, with a significant effect of segment in the AME task [$F(11, 55) = 7.80$, $p < 0.001$, $\eta_p^2 = 0.61$, $\tilde{\varepsilon} = 0.40$], while in the LD task the effect just failed to reach significance [$F(11, 55) = 2.530$, $p = 0.051$, $\eta_p^2 = 0.34$, $\tilde{\varepsilon} = 0.485$].

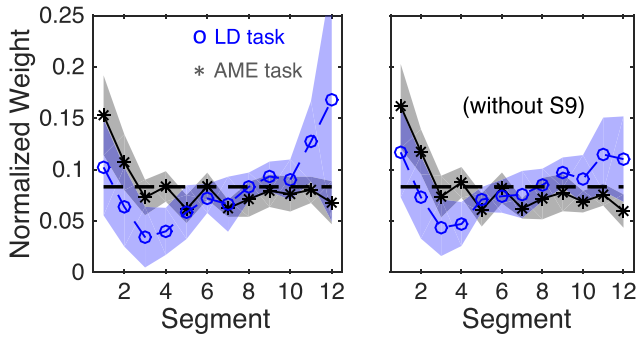

FIG. 5. (Color online) Mean normalized temporal weights for both tasks of experiment 2, when all the subjects are considered (left panel) or when subject S9—who differed from the remaining participants in exclusively considering the three last segments of the stimuli in the LD task—is excluded (right panel). Same format as Fig. 3.

3. Predictive power of the decision models

Fair to good model predictions were found for the LD task, except for subject S14. In the LD task, AUC ranged from 0.69 (S14) to 0.92 ($M = 0.85$, SD $= 0.08$). In the AME task, R^2 was quite low on average ($M = 0.40$, SD $= 0.15$); it ranged from 0.14 (S14) to 0.64. These values are comparable to what we observed for experiment 1 in the separate analyses of the AME data per mean level.

4. Increased predictive power by including temporal weights

Full and restricted models were compared as in the first experiment to evaluate the benefit of adding temporal weights to a loudness model (for details about this model, see experiment 1). The results are reported in Table V.

In the LD task, the AUC of the full model was not significantly higher than the AUC of the restricted model, [$t(6) = 1.839$, $p = 0.115$]. Cohen's d_z indicates a moderate effect size ($d_z = 0.695$), lower than what was measured in the first experiment ($d_z = 0.985$). However, likelihood-ratio tests indicated a significantly better goodness-of-fit ($p < 0.05$) with the full model compared to the restricted model for all subjects but S8 and S14. In the AME task, the R^2 of the full model ranged from 0.17 to 0.66 (M $= 0.43$, SD $= 0.15$) and was significantly higher than the R^2 of the restricted model [$t(6) = 4.673$, $p = 0.003$]. The effect size was large,

TABLE V. Experiment 2, same format as Table II.

Experiment 2										
Goodness-of-fit index (Task)	Model	S8	S9	S10	S11	S12	S13	S14	Average	SD
AUC (LD)	restricted	0.935	0.749	0.897	0.876	0.854	0.870	0.687	**0.838**	0.089
	full	0.939	0.885	0.911	0.891	0.862	0.911	0.697	**0.871**	0.081
	p-value	0.279	<0.001	0.001	0.001	0.023	<0.001	0.627		
R^2 (AME)	restricted	0.645	0.486	0.423	0.308	0.415	0.404	0.141	**0.403**	0.155
	full	0.655	0.503	0.474	0.333	0.440	0.417	0.172	**0.428**	0.150
	p-value	0.018	0.005	0.001	0.001	0.001	0.077	0.002		

$d_z = 1.7662$, similar to what was observed in the first experiment. A significantly better goodness-of-fit was obtained with the full model for all subjects but S13. Overall, this model comparison indicates a large and significant benefit of adding temporal weights to predict loudness in the AME task but a smaller benefit for the LD task.

C. Discussion

On average, the shape of the temporal weighting pattern in the AME task was strikingly similar to the pattern observed in the first experiment, with a significant primacy effect but no recency effect. In the LD task, there was a small primacy effect, but also a small recency effect, even if S9 was excluded from the analysis. The temporal weighting patterns differed significantly between the two tasks, similar to what was found for the first experiment. Thus, the specific "damped" trend of the temporal weighting pattern for the AME task (i.e., with primacy) was also observed in experiment 2, where the range of levels was considerably smaller. As for experiment 1, our results show that considering temporal weights in a model can yield significant improvements in the prediction of the loudness of level-fluctuating sounds.

IV. GENERAL DISCUSSION AND CONCLUSION

The present paper presents the results of two experiments employing psychophysical reverse-correlation to compare the temporal weighting of loudness between a traditional absolute magnitude estimation task (AME) and a level-discrimination task (LD).

In the LD task, we observed "u-shaped" weighting patterns in both experiments, with higher weights assigned to the first and final temporal segments of the sounds than to the temporal segments situated in the middle. In experiment 1, the effect of segment was not significant, however, and in experiment 2 it just failed to reach significance ($p = 0.051$ when one subject was excluded who had deliberately used a different strategy in the LD task than in the AME task). It should be noted that we tested only seven subjects. Thus, the statistical power to detect effects was comparably small, and at least for experiment 2 one could thus expect to find a significant effect of the segment factor if a larger group of listeners had been tested. Previous studies for level-discrimination tasks usually reported much stronger primacy effects (sometimes, weaker recency effects as well) (e.g., Pedersen and Ellermeier, 2008; Oberfeld and Plank, 2011; Oberfeld, 2015). In part, these differences are likely due to stimulus dissimilarities. Probably most important, in the present study, the stimuli were considerably longer (3 s) than in previous experiments on temporal weights, which mostly presented sounds of 1 s or less in duration. Thus, our data show that listeners judging the global loudness of longer sounds also show a trend towards assigning non-uniform temporal weights. Second, the temporal weights inferred from sample discrimination tasks could in fact reflect effects related to memory processes (Dittrich and Oberfeld, 2009; Oberfeld and Plank, 2011). There is some evidence for weaker serial position effects in memory at slow presentation rates (Wickelgren, 1970; Dosher, 1999). Such an effect might have contributed to the weaker primacy effects observed in the present study, because the stimuli were varying in level at a slower rate (4 Hz) than in previous studies (~10 Hz). Another stimulus difference that we should mention here is the variance of the random distributions from which the segment levels were drawn, which was substantially larger than in previous studies ($SD = 5$ dB vs 2 dB). It is not unlikely that the higher "modulation depth" resulted in a percept that was qualitatively different from the "flat-profile" stimuli with comparably small level variations presented in earlier studies. It is not obvious, however, why a different percept should result in different temporal weights.

Furthermore, in the AME tasks of both experiments, we found evidence for significant primacy effects, i.e., the first few temporal segments of the sounds received greater weights compared to the following segments. However, there were no recency effects. The second experiment allowed us to rule out that the much larger level range presented in the first experiment in the AME task, compared to the LD task, was at the origins of the temporal weighting dissimilarities between the two tasks. Taken together, the data of both experiments show that the temporal weighting patterns obtained in the AME task were statistically different from those obtained in the LD task, which clearly contradicts the initial hypothesis proposed in the Introduction.

In this study, we derived the temporal weights in the two tasks using segment loudness as predictors, while the analyses in previous studies were based on the sound pressure levels of the segments (e.g., Ponsot *et al.*, 2013). One may thus ask whether similar weighting patterns are found with segment levels as predictors or if the compressive relation between level and loudness affects the estimated weights. To answer this question, the data were re-analyzed

using segment levels as predictors and we found that the temporal weights were virtually unaffected by this modification. Therefore, the present results are not due to the novel form of analysis introduced in this study; similar conclusions would have been reached if segment levels had been used as predictors in the decision models instead.

Thus, the present study suggests a difference in the temporal weighting processes that depends on whether listeners are asked to do a binary loudness classification (LD task) or to evaluate the global loudness of time-varying sounds by assigning a number (AME task). Although the causes of the dissimilar weighting strategies presently observed are not obvious, one direction could be proposed to explain these results. This explanation relies on potential attentional differences and differences in the allocation of processing resources in the two tasks. Indeed, the AME task requires a more complex evaluation process as compared to the LD task: selecting a number from a potentially infinite set of numbers, or selecting one of only two possible responses. Thus, subjects probably need to invest higher effort in the AME task, which might lead to a more "economic" type of listening as compared to the other task. In addition, the fact that subjects repeated the tasks over hundreds of trials might have emphasized the need to adopt a strategy minimizing both the cost of the evaluation process and their response time in order to handle the task more easily. One way to achieve this in the AME task would be to start the process of selecting a number corresponding to the loudness of the sound soon after the sound has started rather than to wait until the end of the sound. The assignment of attention to the temporal dynamics of the sound would then be different between the two tasks, resulting in different temporal weighting patterns. Also, if our assumption is correct that listeners adopt "economic" strategies in the AME task, then the difference between the weights observed in an AME and in an LD task should be reduced at shorter sound durations, as for example, the 1 s durations used in many previous studies. However, we have no direct experimental evidence to support these arguments yet. An interesting perspective would be to further investigate this aspect and, more generally, to evaluate the extent to which the temporal weighting strategies are conditioned by attention. For instance, it could be hypothesized that visual distractors or background stimuli presented at the beginning or the end of the stimuli to be evaluated would change the magnitude of primacy and recency effects.

In regards to loudness coding and evaluation more generally, the present results confirm the view that time-varying stimuli are not weighted uniformly. We found evidence for primacy effects underlying the judgments (i.e., the first portions playing a greater role) when global loudness was measured either in an AME or in an LD task. Our data indicate that for sounds clearly longer than 1 s, a recency effect emerges in a LD task but not in an AME task. It would be interesting to explore which temporal weights are assigned in even longer sounds. One could speculate that with for example, a 10-s sound, the primacy effect would be further reduced while the recency effect would increase (e.g., Susini *et al.*, 2002). Although such a result should indeed be observed in an LD task, which is assumed to directly reflect temporal weighting processes, we would also speculate that the temporal weighting patterns inferred from an AME task will still show primacy effects for longer sounds, compatible with our idea that subjects always focus on the beginning of the sounds in this task. Additional research is needed to identify the respective contributions of spontaneous temporal weighting and task-specific processes.

The amount of interindividual variation in the temporal weights observed in our experiments was somewhat higher than in most previous studies measuring temporal weights for shorter sounds with durations up to 1 s, where typically very consistent primacy effects were found. The benefit of using individual temporal weights to predict loudness judgments was found to differ between tasks and listeners, but in the majority of cases it yielded significant improvements, which emphasizes the importance of considering non-uniform temporal weighting processes that underpin global loudness evaluation.

The main finding of the present study was the dissimilarity in temporal loudness weighting strategies found between global loudness judged in a sample discrimination task and in a magnitude estimation task. This effect was found even when the range of level variation did not differ between the two tasks. As discussed above, we however argue that this result does not invalidate temporal weighting processes typically inferred from sample discrimination tasks, but rather suggests that magnitude estimations repeated over hundreds of trials might lead people to adopt specific listening strategies in order to handle the required task in an economic fashion. Further work on this issue could help to gain better insight into the mechanisms underlying global loudness evaluation. To understand why and how listeners deploy specific temporal weighting strategies depending on the psychophysical task is certainly a promising direction for future research on loudness. In particular, it would be interesting to examine the proposed hypothesis that the dissimilar temporal weighting between different psychophysical tasks can be attributed to the amount of attention/processing resources distributed between the process of listening and forming a "sensory" representation of its *global loudness*, and the process of selecting a response (e.g., selecting a number in the AME task). Further studies are required before a precise implementation of a typical "universal" non-uniform temporal weighting function in future loudness models. In particular, it remains to be determined which weighting strategy closely reflects actual decisional processes used by listeners in more realistic (non-laboratory) loudness evaluation situations. The precise shape of the weighting function, its dependence on stimulus parameters (e.g., number of segments, variance of each segment) and its robustness to different stimulus configurations and psychophysical tasks still have to be addressed in more detail.

ACKNOWLEDGMENTS

This work was supported by the project *LoudNat* funded by the French National Research Agency (Grant No. ANR-11-BS09-016-01). We would like to thank Elizabeth Strickland and one anonymous reviewer for useful comments on this manuscript.

[1]We were not interested in the values of this parameter.

Ahumada, A. J. (**2002**). "Classification image weights and internal noise level estimation," J. Vision **2**(1), 121–131.

Ahumada, A. J., and Lovell, J. (**1971**). "Stimulus features in signal detection," J. Acoust. Soc. Am. **49**(6B), 1751–1756.

ANSI (**1983**). S1.4, *American National Standard—Specifications for Sound Level Meters* (Acoustical Society of America, New York).

Arieh, Y., and Marks, L. E. (**2011**). "Measurement of loudness, Part II: Context effects," in *Loudness* (Springer, New York), pp. 57–87.

Beard, B. L., and Ahumada, A. J. (**1998**). "A technique to extract relevant image features for visual tasks," in *Human Vision and Electronic Imaging III*, edited by B. E. Rogowitz and T. N. Pappas (SPIE, Bellingham), pp. 79–85.

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

Berg, B. G., and Robinson, D. E. (**1987**). "Multiple observations and internal noise," J. Acoust. Soc. Am. **81**, S33–S33.

Braida, L. D., and Durlach, N. I. (**1972**). "Intensity perception: II. Resolution in one-interval paradigms," J. Acoust. Soc. Am. **51**, 483–502.

Canévet, G., Teghtsoonian, R., and Teghtsoonian, M. (**2003**). "A comparison of loudness change in signals that continuously rise or fall in amplitude," Acta Acust. Acust. **89**(2), 339–345; available at URL: http://openurl.ingenta.com/content?genre=article&issn=1610-1928&volume=89&issue=2&spage=339&epage=345.

Cohen, J. (**1988**). *Statistical Power Analysis for the Behavioral Sciences* (L. Erlbaum Associates, Hillsdale, NJ).

Dai, H., and Micheyl, C. (**2010**). "Psychophysical reverse correlation with multiple response alternatives," J. Exp. Psychol. Hum. Percept. Perform. **36**(4), 976–993.

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

Dosher, B. A. (**1999**). "Item interference and time delays in working memory: Immediate serial recall," Int. J. Psychol. **34**, 276–284.

Durlach, N. I., and Braida, L. D. (**1969**). "Intensity perception: I. Preliminary theory of intensity resolution," J. Acoust. Soc. Am. **46**, 372–383.

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), pp. 169–173.

Green, D. M., and Swets, J. A. (**1966**). *Signal Detection Theory and Psychophysics* (Wiley, New York).

Hellman, R. P. (**1976**). "Growth of loudness at 1000 and 3000 Hz," J. Acoust. Soc. Am. **60**, 672–679.

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. (**1963**). "Monaural loudness function at 1000 Cps and interaural summation," J. Acoust. Soc. Am. **35**, 856–865.

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.

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

Jesteadt, W., Valente, D. L., Joshi, S. N., and Schmid, K. K. (**2014**). "Perceptual weights for loudness judgments of six-tone complexes," J. Acoust. Soc. Am. **136**(2), 728–735.

Leibold, L. J., Tan, H., and Jesteadt, W. (**2009**). "Spectral weights for sample discrimination as a function of overall level," J. Acoust. Soc. Am. **125**(1), 339–346.

Leibold, L. J., Tan, H., Khaddam, S., and Jesteadt, W. (**2007**). "Contributions of individual components to the overall loudness of a multitone complex," J. Acoust. Soc. Am. **121**(5), 2822–2831.

McCullagh, P., and Nelder, J. A. (**1989**). *Generalized Linear Models* (Chapman and Hall, London).

Murray, R. F. (**2011**). "Classification images: A review," J. Vision **11**, 1–25.

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

Oberfeld, D. (**2015**). "Are temporal loudness weights under top-down control? Effects of trial-by-trial feedback," Acta Acust. Acust. **101**(6), 1073–1250.

Oberfeld, D., Heeren, W., Rennies, J., and Verhey, J. (**2012**). "Spectro-temporal weighting of loudness," PLoS One **7**, e50184.

Oberfeld, D., and Plank, T. (**2011**). "The temporal weighting of loudness: Effects of the level profile," Atten. Percept. Psychophys. **73**, 189–208.

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

Ponsot, E., Susini, P., Saint Pierre, G., and Meunier, S. (**2013**). "Temporal loudness weights for sounds with increasing and decreasing intensity profiles," J. Acoust. Soc. Am. **134**, EL321–EL326.

R Core Team (**2015**). "R: A language and environment for statistical computing [computer program]," R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org/ (Last viewed 04/08/15).

Rennies, J., and Verhey, J. L. (**2009**). "Temporal weighting in loudness of broadband and narrowband signals," J. Acoust. Soc. Am. **126**, 951–954.

Stevens, S. S. (**1956**). "The direct estimation of sensory magnitudes—Loudness," Am. J. Psychol. **69**, 1–25.

Susini, P., McAdams, S., and Smith, B. K. (**2002**). "Global and continuous loudness estimation of time-varying levels," Acta Acust. Acust. **88**(4), 536–548.

Suzuki, Y., and Takeshima, H. (**2004**). "Equal-loudness-level contours for pure tones," J. Acoust. Soc. Am. **116**, 918–933.

Teghtsoonian, R., Teghtsoonian, M., and Canévet, G. (**2005**). "Sweep-induced acceleration in loudness change and the 'bias for rising intensities'," Percept. and Psychophys. **67**(4), 699–712.

Wickelgren, W. A. (**1970**). "Time, interference, and rate of presentation in short-term recognition memory for items," J. Math. Psych. **7**(2), 219–235.

Zwicker, E., and Fastl, H. (**1999**). *Psychoacoustics: Facts and Models*, 2nd ed. (Springer, Berlin).