Temporal weights in the level discrimination of time-varying sounds[a]

Benjamin Pedersen[b] and Wolfgang Ellermeier[c]
Sound Quality Research Unit (SQRU), Department of Acoustics, Aalborg University, Fredrik Bajers Vej 7-B5, 9220 Aalborg Øst, Denmark

(Received 12 July 2006; revised 17 November 2007; accepted 19 November 2007)

To determine how listeners weight different portions of the signal when integrating level information, they were presented with 1-s noise samples the levels of which randomly changed every 100 ms by repeatedly, and independently, drawing from a normal distribution. A given stimulus could be derived from one of two such distributions, a decibel apart, and listeners had to classify each sound as belonging to the "soft" or "loud" group. Subsequently, logistic regression analyses were used to determine to what extent each of the ten temporal segments contributed to the overall judgment. In Experiment 1, a nonoptimal weighting strategy was found that emphasized the beginning, and, to a lesser extent, the ending of the sounds. When listeners received trial-by-trial feedback, however, they approached equal weighting of all stimulus components. In Experiment 2, a spectral change was introduced in the middle of the stimulus sequence, changing from low-pass to high-pass noise, and vice versa. The temporal location of the stimulus change was strongly weighted, much as a new onset. These findings are not accounted for by current models of loudness or intensity discrimination, but are consistent with the idea that temporal weighting in loudness judgments is driven by salient events. © *2008 Acoustical Society of America.*

[DOI: 10.1121/1.2822883]

PACS number(s): 43.66.Fe, 43.66.Cb, 43.66.Ba, 43.66.Mk [RAL] Pages: 963–972

I. INTRODUCTION

A. Weighting level information in auditory discrimination tasks

When discriminating or evaluating complex sounds, the auditory system may be assumed to integrate information both across spectral regions and over time. A powerful tool to study such integration processes has been the *analysis of weights* given to the stimulus components defined in the experiment. Pioneered by COSS analysis (i.e., analyzing responses "Conditional On a Single Stimulus" or stimulus component; Berg, 1989), a number of related methodologies have evolved (e.g., Lutfi, 1995), all of which have in common that the listener does not have to be explicitly queried as to his or her weighting of the informational elements. Rather, all but a *global* judgment of pitch (Berg, 1989), loudness (Willihnganz *et al.*, 1997), or lateralization (Saberi, 1996; Stecker and Hafter, 2002) is required, from which, via statistical analysis or the construction of psychometric functions, its relation to the particular informational components is derived.

1. Spectral weights

Most of the few studies applying the analysis-of-weights methodology to the auditory system's use of level information have been concerned with the determination of *spectral* weights in level-discrimination tasks (Doherty and Lutfi, 1996, 1999; Kortekaas *et al.*, 2003; Willihnganz *et al.*, 1997). To that end, in a two-interval, forced-choice paradigm, random, independent level perturbations were added to each of a number of tonal components of different frequency, and the effect of these frequency-specific perturbations on the listener's overall decision yielded the spectral weights in question. Typically, the average weighting functions were found to be relatively flat, though sometimes with greater emphasis given to the highest or lowest frequency components (see Kortekaas *et al.*, 2003).

2. Temporal weights

There have been hardly any studies on the weighting of level information as a function of time (for a review see Stellmack and Viemeister, 2000). Buus (1999) investigated the detectability of a series of six adjacent 25-ms, 1-kHz tone pulses in masking noise. By adding independent level perturbations to the pulses, he was able to construct conditional psychometric functions relating detectability to the random level variations, separately for each of the six temporal pulse locations. From the slopes of these psychometric functions, much like in COSS analysis, relative weights were derived specifying the contribution of each temporal position in the pulse sequence to overall detectability. Analyzing three listeners in a number of experimental conditions, Buus found their weighting functions to be nearly optimal, i.e., giving

[a] Parts of this work were presented at the 149th meeting of the Acoustical Society of America, Vancouver, Canada, May 2005 and at the joint meeting of the German and the French acoustical societies (CFA/DAGA), Strasbourg, France, March 2004.

[b] Now at Atlantic Technologies, Jónas Broncks gøta 35, FO-110 Torshavn, Faroe Islands. Electronic mail: ja_min_pedersen@hotmail.com.

[c] Now at Technische Universität Darmstadt, Germany. Electronic mail: ellermeier@psychologie.tu-darmstadt.de.

equal weight to each of the (equally informative) six pulses, with small, but statistically significant departures favoring the middle portion of the pulse sequence (see his Fig. 3).

Lufi's (1990) studies of sample discrimination contained one condition in which sequences comprised of up to 12 tones had to be discriminated on the basis of an overall level difference between target and standard sequence. COSS analysis (performed on the data of a single listener, see Lufi's Fig. 9) showed the weights assigned to the elements in the sequence to be approximately equal.

In a study involving one of the present authors (Ellermeier and Schrödl, 2000), using a 2IFC paradigm, on each trial listeners compared two 1-s samples of broadband noise (one of which was incremented relative to the other by 1 dB) with respect to their overall loudness. The noise samples were divided into ten segments of 100 ms each onto which small, random level perturbations were imposed. Using COSS analysis (Berg, 1989), weights were derived for the ten temporal segments. They exhibited a bowl-shaped pattern with the beginning of the noise sequence, and (to a lesser extent) the end being emphasized.

B. Memory effects

Further evidence for an unequal weighting as a function of time comes from studies investigating performance effects supposedly related to the functioning of auditory memory. These studies, however, looked at the discriminability of tone patterns in which *frequency* (or pitch) changes rather than level changes had to be tracked. McFarland and Cacace (1992) found strong primacy and recency effects in tone patterns being between seven and thirteen elements long, i.e., significantly better discrimination at the beginning or end of the sequence.

Surprenant (2001) varied the interstimulus interval (ISI) between the sequences to be discriminated, and found strong recency effects, with additional primacy effects emerging as the ISI was increased. Whether such memory effects are obtained for the discrimination of level changes as well remains an open question.

C. Rationale

Given the scarce and equivocal evidence regarding temporal weighting in level discrimination, it appears worthwhile to reinvestigate the issue. In contrast to earlier investigations that shall be done using a *one-interval task* much like in the original study illustrating the weights technique (Berg, 1989). In the present implementation, subjects will be presented with a single stimulus on each trial, and will simply have to classify it as belonging to the "loud" or "soft" set defined by the experiment. This task is conceptually much simpler than a 2IFC task (see Kortekaas *et al.*, 2003), and it does not require assumptions about within-trial memory processes involved, such as making different predictions depending on the length of the ISI (Surprenant, 2001).

Furthermore, since it is conceivable that the contradictory outcomes of some of the studies of temporal weighting may be due to different degrees of practice with the task, or to different strategies used, in Experiment 1, the opportunity to acquire an optimal weighting (with respect to using the physical level information) shall be experimentally manipulated by giving one group of listeners explicit trial-by-trial feedback as to the "correct" response alternative, while another group receives no such feedback, and thus no chance to optimize their strategy.

Finally, since those authors motivated by theories of memory have speculated on the "distinctiveness" of certain events in the temporal sequence, such as the beginning and end of a sound (Neath *et al.*, 2006; Surprenant, 2001), in Experiment 2 additional distinct events shall be experimentally induced by abruptly changing the spectral content of the sound to be judged. In particular, noise sequences will be designed that instantaneously shift from a low-pass to a high-pass characteristic (and vice versa) in the middle of the temporal sequence. Potentially, the spectral shift might constitute a new "distinct" event, e.g., signaling a new "onset," and thereby altering the weight pattern when compared to a control sequence of nonchanging broadband noise.

By virtue of the use of trial-by-trial feedback based on the physical sound generation (in Experiment 1), the task becomes one of *intensity discrimination* (albeit based on multiple channels). It is reasonable to assume, however, that—no matter whether they receive feedback, or not—the subjective quality listeners base their decisions on is related to some "internal" computation of instantaneous or overall *loudness*. The advantage of this view is that it brings models of time-varying loudness and of loudness integration to bear on the behavior observed.

II. EXPERIMENT 1: LEVEL-FLUCTUATING SOUNDS

A. Method

1. Listeners

Ten listeners (one female, nine male) including the authors ("WE" and "BP" in the figures) participated in the experiment. The mean age of the participants was 26 years (range: 18–46 years). All were audiometrically screened, and no one was found to have significant hearing loss (more than 20-dB hearing loss at more than one frequency of 0.125, 0.25, 0.5, 0.75, 1, 1.5, 2, 3, 4, 6, and 8 kHz). Except for the authors, the participants were students with little or no experience in listening experiments.

2. Apparatus

Stimuli were generated digitally on the PC controlling the experiment. A Tucker Davis Technologies System 3 was used for digital-to-analog conversion (RP2.1 unit), setting appropriate levels (two PA5 attenuators), and for powering the headphones (HB7 unit). Signals were presented diotically via headphones (Beyerdynamic DT 990 PRO), at a sample rate of 50 kHz and with 24 bit resolution.

The listeners were seated in a double walled listening cabin during the experiment and made responses using two buttons marked "soft" and "loud" on a special button box connected to the Tucker Davis RP2.1 unit. The box was also used for providing feedback using red and green lights.

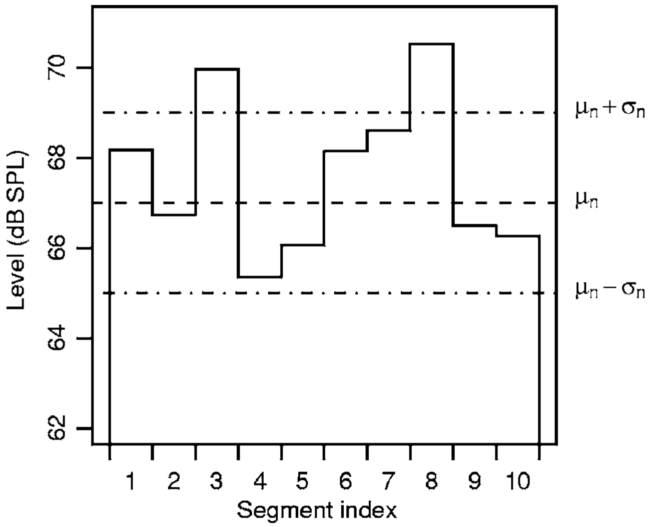

FIG. 1. Temporal envelope of a sound sample (here, "noise").

3. Stimuli

The sounds used in the experiment were samples of white noise having 1-s duration. Their overall level was randomly varied every 100 ms, thus producing a stepwise level-fluctuating sound consisting of ten segments (see Fig. 1). The overall level of each segment was picked randomly from one of two normal distributions denoted "signal" and "noise," with the signal distribution having a higher mean value. The signal distribution had mean value μ_s=68 dB SPL and a standard deviation of σ_n=2 dB. The noise distribution had a mean value μ_n=67 dB SPL and a standard deviation of σ_s =2 dB. Consequently, approximately 95% of the segment levels for each distribution fall in the range $\mu \pm 4$ dB.

Further, the noise and signal distributions overlapped considerably, such that the mean of the ten segments of a given noise sound was sometimes higher than the overall mean (67.5 dB) and vice versa for signal sounds. How often that is expected to happen can be estimated using the properties of the normal distribution. The standard deviation of the mean (given ten segments) is: $\sigma_{10}=\sigma_n/\sqrt{N}$, where N is the number of segments per stimulus (ten). Thus there is approximately a 21% chance for the mean of the ten noise segments to exceed the midpoint between the noise and signal distributions (67.5 dB).

The setup was calibrated using an artificial ear (Brüel & Kjær 4153) with a microphone (Brüel & Kjær 4134). When sound pressure levels are used throughout this article, they refer to the rms sound pressure level of a continuous broadband noise as would be measured in the artificial ear at the given presentation level.

4. Experimental procedure

Participants were instructed that the sounds "were randomly generated," and came from a "soft" or a "loud" set of levels with equal probability. A one-interval two-alternative forced-choice paradigm was used. On each trial, the listener heard a single sound and was asked to judge it as being either soft or loud. In the sequence of trials, noise and signal sounds were presented in random order.

Listeners were divided into two groups in one of which the listeners received trial-by-trial feedback. If the generated sound was from the noise distribution and the listener responded soft or if the sound was from the signal distribution and the response was loud the feedback was a green light, in the other cases it was a red light. No such feedback was given to the other group.

After the completion of each block of 130 trials, overall feedback was given by telling the participants the percentage of "correct" responses they had obtained, i.e., responses which agreed with the noise or signal property of the stimulus. This type of overall feedback was given to all listeners. It helped to motivate the listeners, however based on this type of feedback, it was impossible to change a decision strategy based on trial-by-trial learning.

The first two and a half sessions were used for training. During training the difference between the noise and signal means, μ_s and μ_n, was successively decreased from 3 dB over 2 dB to a final 1-dB difference.

5. Data collection

The experiment was arranged in blocks of 130 trials of which only trials 10–130 were analyzed, leaving the first 9 trials for building up a decision criterion. Five such blocks made up one session, which lasted approximately 40 min. Each listener proceeded through 10 sessions.

6. Determination of temporal weights

In making an overall judgment, listeners are assumed to base their responses on a decision variable, D, defined as

$$D(\mathbf{x}) = \left(\sum_{i=1}^{10} w_i x_i\right) - c, \qquad (1)$$

where $\mathbf{x}$ is a vector of the ten segment levels constituting a given sound. x_i refers to the sound pressure level in decibels of each of the ten segments and w_i is a perceptual weight given to the ith segment. It is assumed that the weighted sum of the segment levels is compared to a fixed decision criterion c. So the strength of the decision variable is given by the difference between the magnitude of the weighted sound levels and the fixed decision criterion.

A logistic function was employed to statistically relate the binary dependent variable (judgments of loud and soft) to the strength of the decision variable:

$$\Psi(D) = p(\text{"loud"}) = \frac{e^D}{1+e^D} = \frac{1}{1+e^{-D}}, \qquad (2)$$

where Ψ describes the probability, p, of a loud response. Note that sometimes other functions (e.g., normal ogives, Berg, 1989) are used to characterize Ψ, but it has been shown, and is true for the present data, that the estimated weights are to a great extent insensitive to the choice of function (Tang *et al.*, 2005).

Insertion of Eq. (1) in Eq. (2) gives

$$\Psi(\mathbf{x}) = p(\text{loud}|\mathbf{w},c,\mathbf{x}) = \frac{1}{1+e^{c-\Sigma_i w_i x_i}}. \qquad (3)$$

TABLE I. Performance of all listeners in Experiment 1 as the percentage of trials which were correctly identified according to the distribution of origin (DIST rows) and according to the mean sound pressure level of the ten segments of a given sound (SPL rows). Signal detection theory sensitivity (d') and bias (β) scores also indicated in separate rows. Listeners in the no-feedback condition (NF) are in the top half and listeners in the feedback condition (FB) are in the bottom half.

		BB	BJ	BP	CP	JJ	Mean
NF	DIST	63	66	69	65	66	**66**
	SPL	68	72	76	71	71	**72**
	d'	0.67	0.86	1.00	0.78	0.85	**0.83**
	β	0.93	0.86	0.91	0.89	1.06	**0.93**
		BL	EH	JV	LH	WE	Mean
FB	DIST	64	72	73	62	69	**68**
	SPL	68	83	81	66	76	**75**
	d'	0.74	1.17	1.21	0.61	0.97	**0.94**
	β	1.05	0.92	0.90	0.85	0.91	**0.93**

The outcome of the experiment is a sequence of loud and soft responses with associated values for $\mathbf{x}$. The values of $\mathbf{w}$ and c which are most likely to yield the results, under the given model, can be estimated by maximum likelihood optimization. For the logistic function, as applied here, this is also known as logistic regression. Standard test statistics for the validity of the model can be applied and furthermore the logistic regression has the benefit of being directly applicable to binary (loud and soft) data (see, for example, Cohen, 2003). These are the main reasons for choosing logistic regression over alternative methods used in other studies estimating weights (for example, Berg, 1989; Ellermeier and Schrödl, 2000; Lutfi, 1995). Though conceptually different, the various methods at hand give very similar estimates for perceptual weights in practice.

It is seen from Eq. (3) that the regression coefficients ($\mathbf{w}$ and c) are not linearly related to the predicted probability of making loud response. The nonlinear relationship is generally true for logistic regression. In this work however, the logistic function is used as a psychometric function, and the regression coefficients are linearly related to the strength of the underlying decision variable as stated in Eq. (1).

In Eq. (1), a linear relationship between the decision variable, D, and the segment levels, $\mathbf{x}$, is assumed. Generally, however, the loudness of steady-state sounds is not linearly related to the sound pressure level in decibels, but within the range of levels used in the present experiment (approximately 60 dB to 75 dB SPL) the relationship is close to linear (see Moore, 2003).

B. Results of Experiment 1

In Table I the performance of the listeners is evaluated via four different measures (DIST, SPL, d', and β in Table I). The DIST score indicates the percentage of trials on which the listeners correctly identified whether the sound originated from the signal or noise distribution. Thus, it evaluates performance in the same way as the feedback given during the experiment. Based on this statistic, performance is very similar in the no-feedback (66%) and feedback (68%) conditions. An alternative is to compute d' and β as defined in signal detection theory. The basis for the two measures is the percentage of trials where the signal distribution was correctly identified (hit-rate) and the percentage of trials where noise sounds were incorrectly identified as loud (false alarm rate). It appears that d' is slightly higher (0.94 vs 0.83) in the feedback condition, but due to the interindividual variance this difference does not reach statistical significance. The bias is nearly identical in the two conditions, marginally favoring loud judgments (β=0.93). It might be argued that the distribution-based performance measures (DIST and d') are unfair, because it is impossible to get all trials correct, since, by chance, high levels can originate from the noise distribution, and vice versa. Therefore, another performance measure (termed "SPL" in Table I) was computed based on the trial-by-trial mean sound pressure level of the ten sound segments. If this mean was higher than 67.5 dB (the midpoint between the two distributions), a loud response was considered correct and when lower than the overall mean, a soft response was considered correct. It appears that performance measured in this way is only slightly higher in the feedback group (75%) when compared to the no-feedback group (72%). The interindividual variance in performance is significantly larger than the mean difference between the two experimental conditions, ranging from 66% for listener "LH" to 83% for listener "EH." But note that this performance measure may not constitute a "fair" comparison, since it favors "flat" weighting curves and a decision criterion close to the overall mean.

In total, 4598 trials per listener (38 blocks × 121 trials) were used to derive weighting curves. The individual weighting curves are seen in Fig. 2. The weights are the scaled regression coefficients of the logistic regression [w_i in Eq. (1)], which provided the most likely fit to the listeners responses given the segment levels (x_i). The coefficients (w_i) are scaled by a factor so the sum of the ten weights is 1. This normalization makes the *relative* importance of each segment (the weighting curve) comparable across listeners. Different scaling values for different listeners reflect individual differences in sensitivity to level changes, which imply that the overall sensitivity is not reflected in the scaled weighting curves.

Figure 2 shows the derived weighting curves for listen-

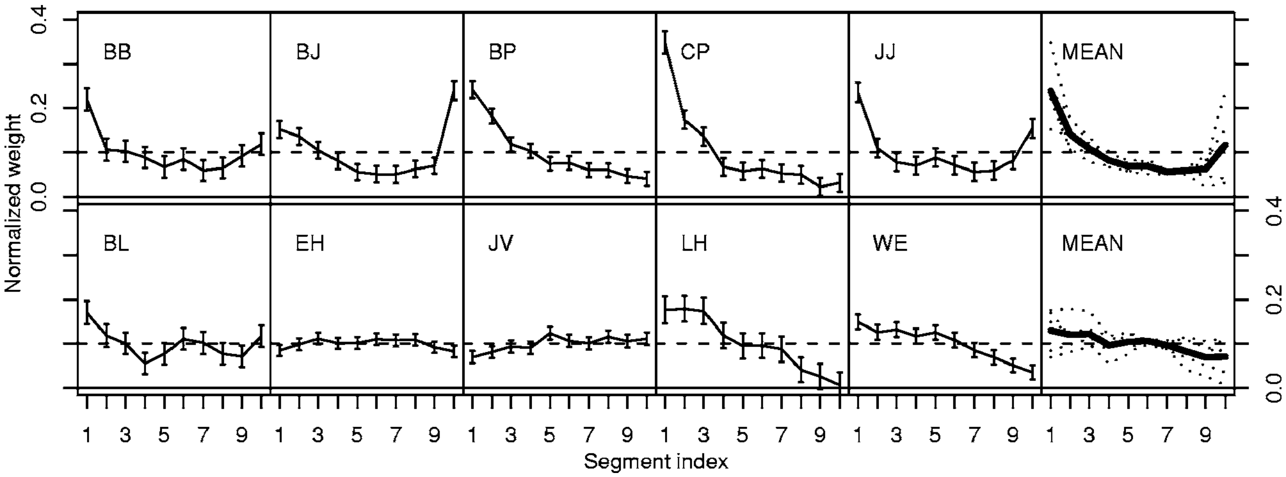

FIG. 2. Temporal weights for Experiment 1. Top row: Without trial-by-trial. Bottom row: Feedback. The error bars indicate the 95%-confidence intervals for the weights as calculated from the logistic regression. Average weights for the two conditions are depicted in the right column.

ers receiving feedback (bottom row) and those of listeners not receiving feedback (top row). For each segment weight, the error bar indicates the 95%-confidence interval. The end points of each interval were calculated prior to normalization and afterwards scaled by the same normalization factor as the weights. Comparing the size of the error bars to the weight differences between segments, it is clear that the shape of the weighting curve is meaningful for a given listener and not a product of random processes. It is also clear that the weighting curves are highly individual, consider “BJ” versus “CP” for example: CP heavily weights the beginning of the sound, while BJ put most weight on the end. For most listeners either the beginning or ending of the sound is weighted more heavily. Exceptions from this are “EH” and “JV,” who do not show pronounced weighting of specific segments.

The effect of feedback can be inspected by comparing listeners in the upper row of Fig. 2 to those in the lower one. Comparing the two mean weighting curves it looks as if feedback did influence the overall shape of the weighting curves. The tendency to emphasize the beginning or the end of a sound seems to be more pronounced in the group of listeners who did *not* receive feedbacks.[1]

An estimate of the statistical significance of this apparent influence of feedback is not easily made, since (a) all weights are normalized to sum to 1, and (b) the weights for the ten segments are not statistically independent for a given listener. Therefore the following testing strategy is suggested: If a listener does not receive feedback, either the beginning or the ending of the sound is weighted more heavily. In any case (beginning, ending or both being weighted more heavily), the central part of the sound must receive less weight due to the normalization. A score for each listener’s weighting of the central part of the sound can be obtained by calculating the sum of the “central” weights 4–8. One score is thus obtained for each of the listeners in each group, and the scores in the groups can be compared using a two-sample t-test. It turned out to be highly significant, $t(7.16)=5.30$;[2] $p=0.001$ indicating that the central weights in the no-feedback group were lower than in the feedback group. This in turn means that the curves in the feedback and no-feedback conditions do indeed have different shapes. If non-normalized weights are used, the certainty is even greater, however, this merely implies that listeners receiving feedback perform better than those not receiving feedback.

The same approach can be used to compare the mean weighting curve in each group to “flat” weights (all weights being equal to 0.1). When, in the no-feedback group, the central weights are compared to a value of 0.1, a one sample t-test results in $t(4)=10.01$; $p<0.001$, and in the feedback group: $t(4)=0.54$; $p=0.62$. That is, the central part of the mean curves is significantly different from optimal weighting for the no-feedback group only. However, from the 95%-confidence intervals in Fig. 2 it is clear that some weights are significantly different from the optimal 0.1 for individual listeners both in the feedback and no-feedback group.

C. Discussion

Global loudness judgments of level-fluctuating noise samples produced evidence for a nonoptimal temporal weighting in that onsets (and to a lesser extent offsets) were weighted more heavily in contributing to overall loudness. A similar, u-shaped weighting pattern as a function of time was observed by Sadralodabai and Sorkin (1999), though for an entirely different task (detecting temporal-pattern changes in a sequence of sinusoids). Furthermore, in the present experiment, trial-by-trial feedback—the presence of which turns the loudness classification into an intensity discrimination task—significantly reduced this emphasis, effectively resulting in an approximately equal (i.e., optimal) weighting of all segments of the sounds.

The present experiment thus provides support both for equal (as in Buus, 1999; Lutfi, 1990) and unequal (as in Ellermeier and Schrödl, 2000) temporal weights, and though all previous studies used some form of feedback it may be speculated that it may have been implemented more or less efficiently. The fact, however, that those participants receiving feedback in the present study were able to “optimize” their performance (with respect to correctly identifying noise and signal sounds) to approximate ideal weights, suggests that there is considerable potential for "perceptual learning" in the temporal weighting patterns.

Earlier investigations have shown that feedback affects the weighting pattern in complex detection or discrimination tasks, e.g., by (a) shifting attention to different spectral regions (Doherty and Lutfi, 1999), (b) focusing on different temporal components (Plank and Ellermeier, 2003), or (c) using different physical cues altogether (Richards, 2002). Note that in all of these examples, however, the task per se was changed (e.g., feedback was made contingent on a different signal property; Richards, 2002), while in the present experiment the mere presence or absence of feedback altered the weighting pattern.

It thus appears, as has been shown for spectral weights (Lutfi, 1995; Southworth and Berg, 1995), there is considerable liberty in how listeners weight the components of perceptual information available, and that, depending on the task requirements, different weighting patterns may emerge. The considerable individual differences evident in the present data also argue for a certain flexibility in the assignment of weights.

One might speculate whether the overall pattern of weights decreasing with segment number (see Fig. 2) is due to earlier noise segments masking the later ones (forward masking). This is unlikely due to several reasons: First, the relative level differences between adjacent segments are too small (a few decibels), and the segment duration is too long (100 ms) to expect much forward masking. Second, the fact that essentially flat patterns (Subjects EH and JV), or a pattern that emphasizes the end (Subject "BJ") are observed, argues against a peripheral process such as masking being causal. Third, one would not expect feedback to produce a release from forward masking. Fourth, and finally, simulations using the loudness model by Glasberg and Moore (2002), which takes masking into account, failed to predict a decaying pattern of weights (Pedersen, 2007, Chap. 3).

The outcome of Experiment 1, however, does not specify the nature of the processes very well. It remains open, for example, whether the emphasis of beginning and ending observed in the unbiased listening condition is due to memory effects (primacy and recency), or simply to the perceptual salience of onsets and offsets.

III. EXPERIMENT 2: TWO-EVENT SOUNDS

To further clarify the issues raised by Experiment 1, a second experiment was performed, in which sounds of the same duration and temporal structure as those used in Experiment 1 were subjected to a sudden spectral change in the middle of the temporal sequence. The spectral change thus constitutes a salient event which is not tied to primacy or recency, and the effect of which on the temporal weighting pattern may be observed.

A. Method

1. Listeners

Six naive listeners took part in the experiment, none of whom had participated in Experiment 1. Their hearing was screened, and no one was found to have significant hearing loss (more than 20-dB hearing loss at more than one frequency of 0.125, 0.25, 0.5, 0.75, 1, 1.5, 2, 3, 4, 6, and 8 kHz). The participants were five males and one female with an average age of 24 years (range: 22–28 years).

2. Apparatus

In Experiment 2, different hardware was used for signal generation: Signals were digitally generated using a sound card (RME HDSP9632) and subsequently converted to an analog signal via a digital-to-analog converter (Tracer Technologies Big DAADI), using 16-bit resolution and a sample rate of 44.1 kHz. The resulting signal was fed to a headphone amplifier (Behringer HA4400) and diotically played over headphones (Beyerdynamic DT 990 PRO).

3. Stimuli

As in Experiment 1, all sounds were of 1-s duration and the levels of the ten temporal segments were chosen from random distributions having the same parameters as in Experiment 1. The only difference was the spectral content of the sounds. In one condition of Experiment 2 the first half of the sound (i.e., the first five segments) was low-pass filtered and the last part (the last five segments) high-pass filtered. This type of sound is denoted "LH," indicating the change from low to high frequency content. In a different condition the segments were filtered in the opposite order, denoted "HL," i.e., changing from high-pass to low-pass filtered noise. The cut-off frequency was 1 kHz for both high- and low-pass filters. The filtering was done using digital finite impulse response filters (FIR of order 501), for which the attenuation was more that 50 dB in the nonpass section at a distance of more than 150 Hz from the cut-off frequency (1 kHz). The phase response of each filter was linear. The two filtered blocks were aligned so no silent interval occurred. A third condition, where no spectral change occurred, was included for comparison with Experiment 1. In this condition white noise was used as in Experiment 1 (denoted "WN").

4. Experimental procedure

The listeners' task was the same as in Experiment 1. After hearing a single sound, the listener responded whether it was loud or soft. No trial-by-trial feedback was given. After each block of 200 trials the percentage of "correct" responses based on the distribution from which the sounds were drawn was communicated to the participants. Because of the difference in quality of the filtered blocks, the listeners were specifically instructed to judge the composite sound as one whole.

Before data collection started all listeners learned the task in a similar way as in Experiment I. The difference in mean between the noise and signal distributions was slowly decreased (from 4 to 1 decibels). The training blocks contained fewer trials (50) and LH, HL, and WN blocks were included. Feedback on the percentage of correct responses helped listeners to realize whether they were on the right track.

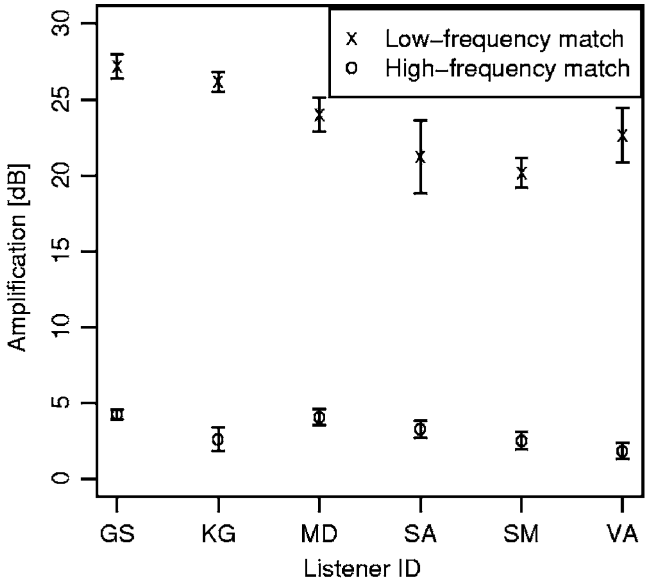

FIG. 3. Loudness matches for Experiment 2. Amplification required for the low-pass and high-pass noises to match a 67.5 dB SPL white-noise reference. Individual outcomes for each of the six listeners are depicted. The error bars indicate the 95%-confidence intervals.

5. Data collection

The experiment was arranged in blocks of 200 trials each. A given block contained either filtered noise (both LH and HL) or broadband noise (WN). In blocks containing filtered noise, LH and HL trials were presented in a random sequence. A total of 1200 trials per condition (LH, HL, or WN) was presented. Each session, lasting approximately 40 min, contained three blocks, one in which unchanging white-noise stimuli were presented (WN), and two containing spectral changes (LH and HL). The order of the blocks was counterbalanced within listeners, and across the six sessions used for data collection. 1140 trials were used per condition and listener in the regression analysis, since the first 9 trials in each block were discarded for practice.

6. Loudness calibration

In order to present the filtered noises at equal loudness, all listeners initially performed individual loudness matches before proceeding to the experiment proper. An adaptive two-interval forced choice one-up/one-down paradigm was used to match samples of either low-pass or high-pass filtered noise to the fixed white-noise reference at 67.5 dB SPL. All sounds had a duration of 0.5 s and there were no random fluctuations of the segment levels.

The resulting loudness matches varied somewhat across listeners (up to ~7 dB for the low-pass, and 3 dB for the high-pass noise, see Fig. 3). They required the low-pass noise to be raised in level by approximately 23 dB on average to be equally loud as the broadband noise. The high-pass noise required 3-dB amplification on average to achieve the same loudness. In the experiment proper *individual* matches were used for calibration of the filtered blocks.

B. Results of Experiment 2

As in Experiment 1, performance measures were calculated for each listener in each condition. On average, performance was almost identical in the three experimental conditions (WN, LH, and HL), amounting to approximately 64% correct when based on the SPL criterion. The SPL score varied across listeners with a minimum of 54% for listener VA in the LH condition and maximum of 74% for listener MD in the WN and LH conditions. The value of β was very close to 1.0 for all listeners in all conditions, indicating an equal balance between loud and soft judgments in all conditions. Generally, the performance was worse in Experiment 2 than in Experiment 1, which may for example be caused by the extreme weighting applied by some listeners (see for example GS in Fig. 4) or because of the listeners being less consistent in their judgments (large error bars for VA in Fig. 4). However, there are no indications that spectral-change condtions are harder than the condition with no spectral change.

As in Experiment 1, weighting curves were derived for each individual listener, using logistic regression, separately for the white noise (WN), low-high (LH), and high-low (HL) conditions. The estimated weights are depicted in Fig. 4.

The results of the white-noise condition may be compared to those of Experiment 1, in which identical stimuli were used. A similar trend as in the "no feedback" condition of Experiment 1 is found (compare top rows of Figs. 4 and 2), with relatively greater weights being assigned to the initial sound segments. The results of the two experiments are similar, except that the emphasis on the initial segments is even greater, and there is no evidence for higher weighting of the ending of the sound in the new experiment. As in Experiment 1, the weighting patterns vary across listeners.

When a spectral change is introduced in the middle of the sound (LH and HL in the center and bottom rows of Fig. 4), the weighting curves show distinctly different patterns. For most listeners the sixth segment (for which the spectral change occurs) receives greater weight in the LH and HL conditions. It also appears that the order of the high- and low-pass filtered blocks makes a difference for the weighting strategy applied by the listeners, though in idiosyncratic ways, consider GS for example: In the HL condition his decision is based almost exclusively on the first segment (beginning of low-frequency block), whereas in the LH condition both the first and the sixth segment contribute significantly to the decision. Thus, the start of the low-frequency block is always heavily weighted by this listener, but the beginning of the high-frequency block is only weighted heavily if it is also the onset of the entire sound. Listener SM almost shows the reverse behavior with respect to the weighting in the two spectral conditions. Finally, SA almost seems to ignore the high-frequency part of the sound in both the LH and HL conditions.

As can be seen from the size of the 95%-confidence intervals depicted in Fig. 4, some listeners were clearly more consistent in their weighting than others. Nevertheless, all listeners performed significantly better than chance.

A statistical test as to whether the spectral change (LH or HL) made a difference compared to the nonchanging

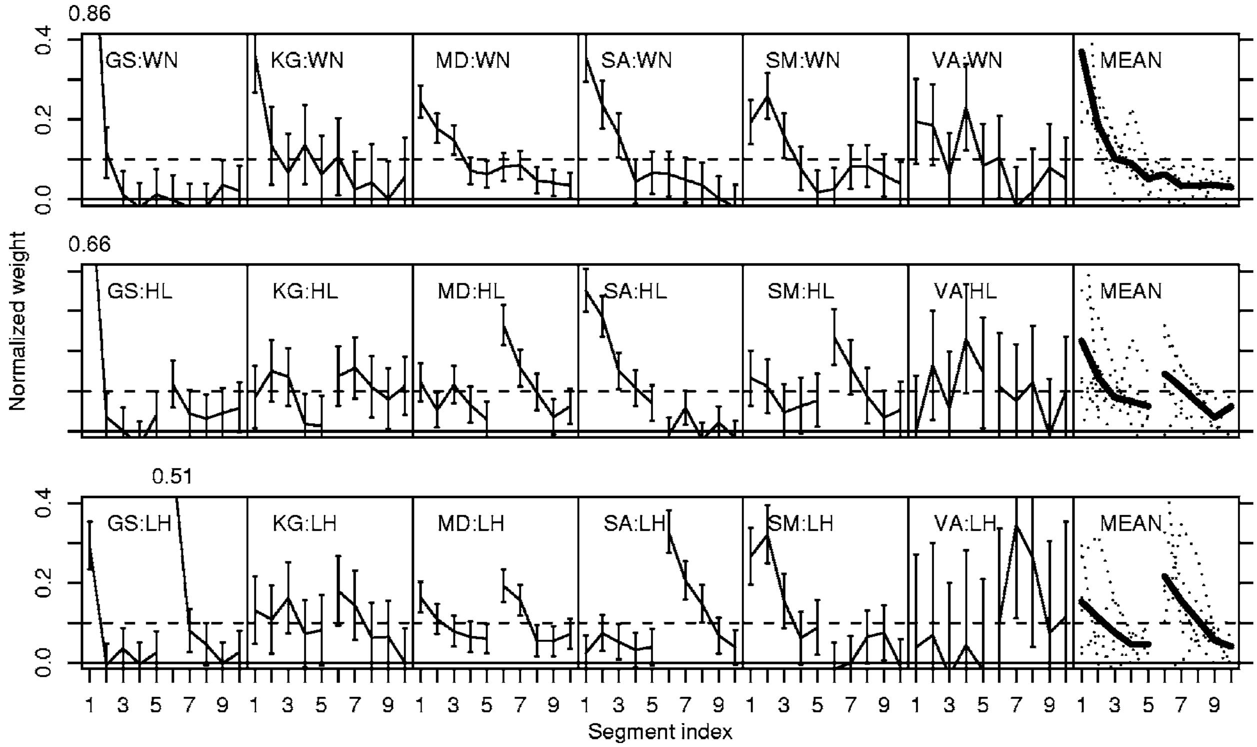

FIG. 4. Weighting curves for all listeners in Experiment 2. Columns represent listeners while rows contain the different experimental conditions. First row: Broadband noise with no frequency change. Second and third rows: Two-event conditions; high-low and low-high changes, respectively. The onset of the spectral change is indicated by a break in the weighting curve. The error bars indicate 95%-confidence intervals for the weights as calculated from the logistic regression. Average weights for the three spectral conditions are depicted in the right column.

(WN) condition was performed in the following way: The sixth and seventh segments were defined as reflecting the onset of the spectral change. By summing each listener's weights for these two segments, a score for the weighting of the spectral change was calculated for each listener in each condition. Using these scores, two-tailed, repeated measures t-tests were performed, between the spectral-change conditions and the nonchanging condition. They revealed the weights for the critical segments (6 and 7) to be significantly greater in the spectral-change conditions, both when comparing LH with WN: $t(5)=3.02$, $p=0.03$, and when comparing HL with WN: $t(5)=2.65$, $p=0.045$. Thus, the increased weighting given to the onset of a new spectral event (see Fig. 4) appears to be statistically significant.

C. Discussion

Experiment 2 showed the temporal location at which a spectral change occurred to receive about as much weight as the initial onset of the composite sound. This is consistent with the idea of perceptual weighting being guided by salient events. These may be onsets, offsets, spectral shifts, or qualitative changes yet to be investigated such as changes in spatial location, etc.

The results of Experiment 2 are not easily reconciled with a memory explanation based on primacy and recency effects, at least not one that requires the entire sound to be stored in memory in a simple sequential way. Whether assumptions about "resetting" the onset detector, or separate storage of the two spectral events might remedy the situation, is doubtful.

IV. FINAL DISCUSSION AND CONCLUSION

That condition of Experiment 1 in which listeners received trial-by-trial feedback based on the generation of the physical levels making up the sound sequence may be considered a straightforward intensity discrimination task. In that task, listeners were encouraged to optimize their performance with respect to inferring which of two level generators produced the auditory percept they had. The experiment showed that—given feedback—the participants could indeed accomplish that and were using the evidence of ten successively presented levels almost optimally, i.e., with a nearly "flat" weighting characteristic (see the bottom part of Fig. 2).

When—by contrast—the trial-by-trial feedback was omitted, as for the other group of listeners participating in Experiment 1, and in Experiment 2, additional effects emerged, such as a higher weighting given to the beginning (and end) of the sound sequence (see the top part of Fig. 2), or an increased emphasis on those portions of the signal that were temporally close to a spectral change (see Fig. 4).

This may be summarized as stating that *salient events* are perceptually emphasized in natural, unbiased listening, as it occurs when no particular feedback scheme is implemented. It may be further hypothesized that that kind of listening is close to our everyday loudness perception. Only when feedback suggests otherwise (as in the pertinent condition of Experiment 1), are the temporal loudness weights adapted to maximize correct performance.

This kind of reasoning, and the fact that most of the data of the present work were collected in an intensity discrimination paradigm that—for the participants—was framed as a loudness classification task without feedback, suggests to investigate how well current loudness models can explain the peculiar temporal weighting patterns observed. It will be argued that most of the results of the present experiments are incompatible with the notion of an automatic, accumulative integration process as hypothesized by most loudness models (e.g., Glasberg and Moore, 2002; Zwicker, 1977). A major outcome of these loudness models is to generate a continuous loudness curve, which is to account for the results of, e.g., temporal masking experiments and subjective loudness matches of modulated sounds. But to predict how listeners arrive at global loudness judgments requires further stating how this continuous curve is "integrated" to produce a single judgment. The present data address both of these stages.

In the calculation of a continuous loudness curve, all current models operate with some sort of temporal summation with a critical time coefficient in the range from 20 to 50 ms depending on whether the loudness curve is rising or falling (Glasberg and Moore, 2002; Grimm *et al.*, 2002; Zwicker, 1977). It is therefore impossible for loudness determined by these models to fluctuate any faster than the time coefficients allow. The fact that in the present experiment, for some listeners, adjacent segments were weighted very differently (see Fig. 2) implies that their "continuous loudness" must fluctuate at least as rapidly as the segment duration of the sounds (100 ms) or else a particular segment could not be "singled out" receiving extra weight. Thus, though the time coefficients of the models are not in direct contradiction with the observed weighting patterns, there is some indication that the integration taking place is not a simple "smoothing" process. In their loudness model, Glasberg and Moore (2002) introduce a further stage of determining "long-term" loudness, with integration coefficients of approximately 100 ms for rising and 2000 ms for falling loudness. These long time coefficients are not compatible with the results of the present experiments. Other researchers have also found it hard to reconcile the outcome of listening experiments with the predictions of loudness models when different forms of temporal variation were examined (e.g., Grimm *et al.*, 2002; Stecker and Hafter, 2000).

When it comes to integrating the sensory information into a loudness judgment, the present experiments provides further evidence against the operation of simple loudness integration:

(1) Weights derived for the ten temporal segments defined were not uniform, but rather, in the unbiased, nonfeedback conditions of Experiment 1 and 2, provided evidence for perceptual emphasis of onsets and offsets. That is not predicted by any of the current loudness models. Nor is it predicted by practical measurement rules (e.g., Zwicker and Fastl, 1999) that assume values close to the maximum (e.g., the fourth percentile; Grimm *et al.*, 2002; Zwicker and Fastl, 1999) to determine the loudness of a time-varying sound. All of these rules would, for the randomly varying sounds used in the present experiments, imply "flat" weighting curves to result.

(2) When trial-by-trial feedback was provided in Experiment 1, listeners adapted their temporal weights to approach an optimal, uniform weighting of all stimulus segments. Such a "learning effect" is hard to reconcile with the notion of an automatic integration process operating in the auditory periphery with a relatively long time coefficient. Rather, the listener must have access to some representation of the segment loudnesses (prior to integration) with a finer resolution than the segment duration in order to modify weights to maximize the percentage of correct responses.

(3) When a change was introduced into the noise sequence by switching the spectrum from a low-pass to a high-pass characteristic (or vice versa) in Experiment 2, listeners strongly weighted the onset of the "new" sound feature, thus boosting weights in the central portion of the composite stimulus. That is inconsistent with temporal wide-band energy integration which would be "blind" to the spectral change; it is also inconsistent with a memory explanation based on a "primacy" and "recency" advantage.

(4) All of the weighting patterns observed exhibited considerable interindividual variability. That in itself argues against a low-level integration mechanism, which one would not assume to leave degrees of freedom for individual idiosyncrasies. Rather it suggests some cognitive process to be involved, which can be controlled by the listener to some extent.

What then are the alternatives for understanding the weighting of level information, and its adaptability to various listening conditions? It appears that, in the time range of several hundred milliseconds investigated here, different stimulus segments must be individually accessible, granting the listener "multiple looks" (Viemeister and Wakefield, 1991) on a temporal loudness pattern. Depending on the task requirements (Experiment 1) or on stimulus features (the spectral changes in Experiment 2), these "looks" may be weighted differently, under implicit control by the listener. The particular salience of onsets and offsets, as well as qualitative changes in the stimulus, may be due to mechanisms of memory, or more likely to the "distinctiveness" (Neath, 1993; Neath *et al.*, 2006) of these events in relation to other stimulus components, thereby attracting greater perceptual weight.

How could these hypotheses be put to further tests? If memory was a factor, one might expect the timing of the event sequence to play a crucial role. Furthermore, to explore the distinctiveness concept, salient changes other than spectral ones (e.g., spatial lateralization) might be explored, or an event could be generated by switching from coherent to incoherent noise samples of the carrier signal across the two ears. Potentially, the segment levels could also be different across the ears, providing a means to examine both temporal and binaural loudness summation.

Hopefully, based on such research, a clearer picture will emerge, on how perceptual and cognitive processes interact when listeners discriminate time-varying sounds differing in level.

ACKNOWLEDGMENTS

We would like to thank Florian Wickelmaier for helpful hints regarding the statistical analysis of the results. This research was carried out as part of the "Centercontract on Sound Quality" which establishes participation in and funding of the "Sound Quality Research Unit" (SQRU) at Aalborg University. The participating companies are Bang & Olufsen. Brüel & Kjær, and DELTA Acoustics & Vibration. Further financial support comes from the Ministry for Science, Technology, and Development (VTU), and from the Danish Research Council for Technology and Production (FTP).

[1] The "stability" of the individual weighting patterns was examined by analyzing the data in ten blocks of increasing practice (details in Pedersen, 2007). There was no indication that listeners altered their temporal weighting in the course of the experiment. The observed difference between listeners in the feedback and no-feedback group may thus already emerge in the training trials prior to the experiment proper.

[2] Noninteger degrees of freedom result since a Welch–Satterthwaite approximation was used, which means that equal variance of the weights in the feedback and no-feedback group need not be assumed.

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

Buus, S. (**1999**). "Temporal integration and multiple looks, revisited: Weights as a function of time," J. Acoust. Soc. Am. **105**, 2466–2475.

Cohen, J. (**2003**). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, 3rd ed. (Lawrence Erlbaum, Mahwah, NJ).

Doherty, K. A., and Lutfi, R. A. (**1996**). "Spectral weights for overall level discrimination in listeners with sensorineural hearing loss," J. Acoust. Soc. Am. **99**, 1053–1058.

Doherty, K. A., and Lutfi, R. A. (**1999**). "Level discrimination of single tones in a multitone complex by normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. **105**, 1831–1840.

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 Psychophsics*, edited by C. Bonnet (Université Louis Pasteur, Strasbourg), pp. 169–173.

Glasberg, B. R., and Moore, B. C. J. (**2002**). "A model of loudness applicable to time-varying sounds," J. Audio Eng. Soc. **50**, 331–342.

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

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

Lutfi, R. A. (**1990**). "Informational processing of complex sound. II. Cross-dimensional analysis," J. Acoust. Soc. Am. **87**, 2141–2148.

Lutfi, R. A. (**1995**). "Correlation-coefficients and correlation ratios as estimates of observer weights in multiple-observation tasks," J. Acoust. Soc. Am. **97**, 1333–1334.

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

Moore, B. C. J. (**2003**). *An Introduction to the Psychology of Hearing*, 5th ed. (Academic, San Diego).

Neath, I. (**1993**). "Distinctiveness and serial position effects in recognition," Mem. Cognit. **21**, 689–698.

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.

Pedersen, B. (**2007**). "Auditory temporal resolution and integration: Stages of analyzing time-varying sounds," Ph.D. thesis, Aalborg University, Aalborg University, Denmark.

Plank, T., and Ellermeier, W. (**2003**). "Discrimination of temporal loudness profiles," in *Fechner Day 2003. Proceedings of the 19th Annual Meeting of the International Society for Psychophysics*, edited by B. Berglund and E. Borg, Stockholm, Sweden, pp. 241–244.

Richards, V. M. (**2002**). "Varying feedback to evaluate detection strategies: The detection of a tone added to noise," J. Assoc. Res. Otolaryngol. **3**, 209–221.

Saberi, K. (**1996**). "Observer weighting of interaural delays in filtered impulses," Percept. Psychophys. **58**, 1037–1046.

Sadralodabai, T., and Sorkin, R. D. (**1999**). "Effect of temporal position, proportional variance, and proportional duration on decision weights in temporal pattern discrimination," J. Acoust. Soc. Am. **105**, 358–365.

Southworth, C., and Berg, B. G. (**1995**). "Multiple cues for the discrimination of narrow-band sounds," J. Acoust. Soc. Am. **98**, 2486–2492.

Stecker, G. C., and Hafter, E. R. (**2000**). "An effect of temporal asymmetry on loudness," J. Acoust. Soc. Am. **107**, 3358–3368.

Stecker, G. C., and Hafter, E. R. (**2002**). "Temporal weighting in sound localization," J. Acoust. Soc. Am. **112**, 1046–1057.

Stellmack, M. A., and Viemeister, N. F. (**2000**). "Observer weighting of monaural level information in a pair of tone pulses," J. Acoust. Soc. Am. **107**, 3382–3393.

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

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

Viemeister, N. F., and Wakefield, G. H. (**1991**). "Temporal integration and multiple looks," J. Acoust. Soc. Am. **90**, 858–865.

Willihnganz, M. S., Stellmack, M. A., Lutfi, R. A., and Wightman, F. L. (**1997**). "Spectral weights in level discrimination by preschool children: Synthetic listening conditions," J. Acoust. Soc. Am. **101**, 2803–2810.

Zwicker, E. (**1977**). "Procedure for calculating loudnesss of temporally variable sounds," J. Acoust. Soc. Am. **62**, 675–682.

Zwicker, E., and Fastl, H. (**1999**). *Psychoacoustics: Facts and models* (Springer, Berlin).