Global and Continuous Loudness Estimation of Time-Varying Levels

Patrick Susini, Stephen McAdams, Bennett K. Smith
Institut de Recherche et Coordination Acoustique/Musique (IRCAM-CNRS), 1 place Igor Stravinsky,
F-75004 Paris, France

Summary
Conventional psychoacoustic methods are not sufficient to produce instantaneous judgments or real-time perceptual tracking of nonstationary sounds. In order to evaluate continuously or globally the loudness of pure tones of variable duration and time-varying level, two cross-modal matching methods were used, one with continuous force feedback and another without force feedback but using a continuous analogical/categorical judgment scale. The global loudness of various predefined acoustic level profiles was estimated under two experimental conditions: one with continuous estimation during the sound sequence and the other without. The results show that the continuous judgment profiles transcribe quite well the stimulus contours, although a temporal lag on the order of 1 s between stimulus contour and response profile is observed, as is an asymmetry between increasing and decreasing profiles. Global judgments are influenced by the rate or duration of level change and by the level at the end of the signal. A recency effect, similar to that observed in auditory memory research using an immediate recall task, is thus revealed for loudness estimates on nonstationary sounds lasting a few tens of seconds. Finally, global judgments are generally higher without preceding continuous evaluation for both matching methods.

PACS no. 43.66.Cb, 43.66.Mk, 43.66.Yw

1. Introduction

Kuwano and Namba [1] and Fastl [2] have done work on the subjective evaluation of long-duration sound sequences extracted from urban environments. They have shown that sound events that are prominent in level strongly influence the global impression of loudness reported by listeners. The physical measures performed in these two studies show in one of them that instantaneous judgments and global judgments were well correlated with acoustic level in dBA and L_{eq}, respectively [1], while in the other one [2] the global impression was best predicted by a specific value of loudness, N_4 (ISO 532B, [3]).

Those acoustic energy integration algorithms do not take into account the temporal distribution of sound events. However, various experiments performed in the realm of memory research reveal a primacy and a recency effect. For example, in an immediate serial recall experiment in which subjects are asked to reproduce a series of items (words, letters, or numbers) in their order of appearance immediately following presentation, a considerable advantage is found for the first and the last items in the series [4]. The recall curve for auditory presentation is U-shaped. The two branches of the curve correspond to the first and last items of the series. The two processes that give rise to this result are called primacy and recency effects, respectively. According to Crowder and Morton [5], two processes are brought into play giving access on the one hand to interpretations of the first stimuli that characterize their conceptual and abstract properties (categorical memory), and on the other hand to acoustic properties of the most recent stimuli (auditory sensory memory or precategorical acoustic storage) [6, 7]. The latter process, recency effect, is particularly characteristic of auditory memory.

Two distinct hypotheses concerning auditory integration of long-duration sequences thus emerge. On the one hand, the dominant events are responsible for the global impression, whatever their temporal distribution. On the other hand, research on auditory memory shows two characteristic results in recall tasks: primacy and recency effects arising from two types of storage, long- and short-term storage, respectively.

In the study presented in this paper, the different experimental conditions and the choice of stimuli were designed to test whether a recency effect is present in a global loudness judgment task and how the characteristics of temporal variation (rate of change, increase vs decrease) influence the loudness judgments. The experiment was not designed to study the primacy effect; indeed the stimuli used have no meaning for listeners and could not thus be categorized easily. Three groups of stimuli were created. Two were principally designed to study the effect of rate and duration of temporal variation, whereas the purpose of the third

Received 16 February 2001, revised 4 April 2002,
accepted 10 April 2002.

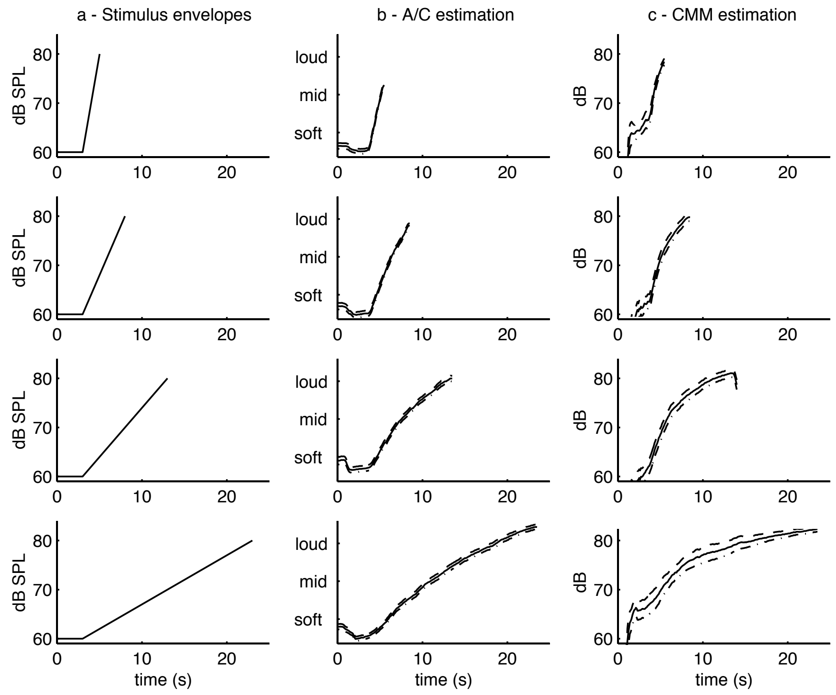

Figure 1. a) Acoustic level profiles in dB of RAMP stimuli (R2, R5, R10, and R20). Mean profiles across subjects of continuous judgments: b) by the analogical/categorical method (A/C) and c) by cross-modal matching (CMM). The dashed-lines represent ± one standard-error of the mean (solid lines) at each instant.

was to study the influence of temporal order of higher- and lower-level peaks.

Two cross-modal matching methods were used: one with continuous force feedback and another without force feedback but using a continuous analogical/categorical judgment scale. The subjects used both methods to give the continuous and global judgments. The global judgment at the end of each stimulus presentation is estimated under two experimental conditions, with and without continuous judgment. The two methods and two experimental conditions were combined to test whether the results were independent of the experimental protocol.

2. Method

2.1. Stimuli

Stimulus sequences consisted of 1-kHz pure tones with time-varying levels. Three groups of temporal profiles were used. For each, the onset and offset ramps were 50 ms in duration. In the first group, the signals had a 3-s plateau at 60 dB SPL, followed by a linear increase in level on a decibel scale to 80 dB SPL over durations of 2, 5, 10 or 20 s (Figure 1a). This class of contours will be labeled RAMP, with individual contours notated R2, R5, R10 and R20, respectively, for the four ramp durations. In the second group, the contours were composed of increasing (60 to 80 dB) and then decreasing (80 to 60 dB) ramps of identical duration, similarly to the single ramps of the first group, but of oppositely signed slopes, with 3-s plateaux at 60 dB at the beginning and end. The duration of increasing and decreasing ramps were 2, 5, 10 or 20 s (Figure 2a). This class of contours is labeled 1PEAK with individual contours denoted 1P2, 1P5, 1P10, and 1P20, respectively. The contours of the third group correspond to six combinations of three peaks, the maximum levels of which were 75, 80, and 90 dB SPL, and which are denoted, L (Low), M (Medium), and H (High), respectively. The increasing and decreasing ramps forming each peak were 5-s in duration (Figure 3a). The plateaux between peaks had a duration of 10 s and a constant level of 60 dB SPL. The six combinations correspond to the different permutations of the three peaks: HML, HLM, MHL, LHM, MLH and LMH, in the order presented in Figure 3a. This class of contours is labeled 3PEAK. Each contour started with a 3-s plateau at 60 dB SPL.

2.2. Apparatus

Several studies have examined psychophysical methods with short-duration, stationary stimuli (cf. [8] for a review). However, few studies have examined continuous ratings of time-varying stimuli. Table I summarizes the methods proposed by different authors over the last fifteen years [1, 2, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20]. The present study uses two methods. One method was chosen on the basis of a review of the literature [17], whereas the other was developed in our laboratory [8].

Two judgment methods were used, each involving a separate device: a cross-modal matching device (CMM) with force feedback and an analogical/categorical scaling device (A/C) without force feedback.

In the CMM procedure, the subject associated an equivalent muscular force with the loudness of a stimulus by moving the lever of the device about an axis of rotation. The technical and functional characteristics of the device have been described in a previous article [8]. Briefly, the

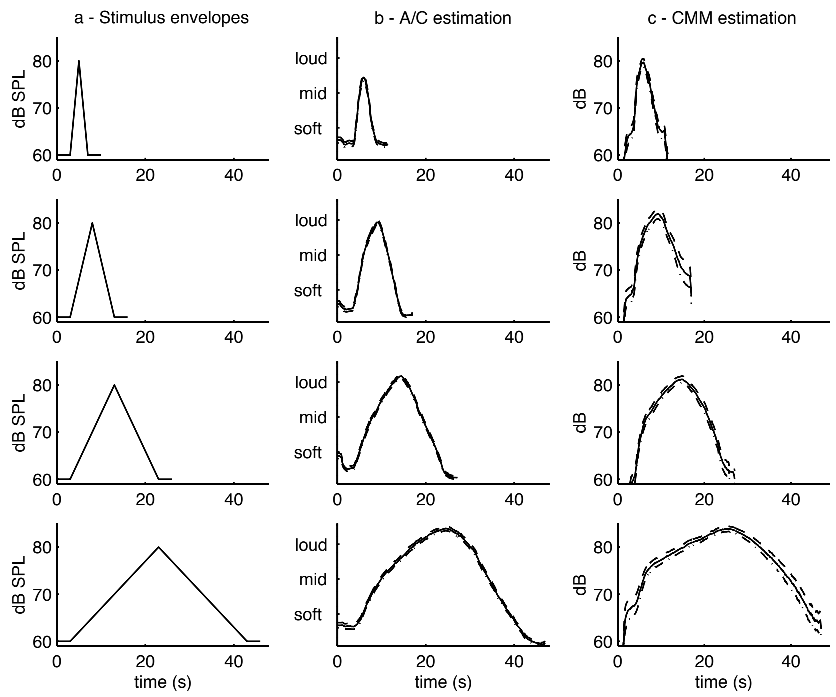

Figure 2. a) Acoustic level profiles in dB of 1PEAK stimuli (1P2, 1P5, 1P10, and 1P20). Mean profiles obtained: b) by the analogical / categorical method (A/C) without force feedback and c) by cross-modal matching (CMM) with force feedback. The dashed-lines represent ± one standard-error of the mean (solid lines) at each instant.

Table I. Summary of judgment methods and their applications.

Rating	Sounds used	References
Categories	Road traffic	Kuwano, Namba [1, 9]
	Trains	Namba *et al.* [10]
	Helicopter	Kuwano, Namba [11]
	Car acceleration	Kuwano *et al.* [12]
Analogical	Road traffic	Fastl [2, 13] Gottschling [14]
	Music	Madsen [15], Schubert [16]
Analogical/ categorical	Road traffic	Weber [17] Hedberg, Jansson [18]
	”Artifical” sounds	Hedberg, Jansson [18]
	Speech quality	Hansen, Kollmeier [19]
Semantic	Music	Namba *et al.* [20]

lever is fixed at a rotation point. The upper part is used as a handle by the subject and the lower part plays the role of a counterweight, the force of which depends on the displacement angle. The lever, by exerting a resistance as a function of the angular displacement, thus creates a force feedback that continuously informs the subject about the judgment in progress. At the level of the axis of rotation, a potentiometer measures an electrical voltage associated with the angle of displacement. As such, the intensity of the muscular force applied by the subject is determined as a function of the angle and the resistance of the system, the latter being adjustable for each individual. The resistance of the device depends on the angle of displacement of the lever, the mass at the opposite end of the lever and its position with respect to the axis of rotation. The proprioceptive function associated with the device corresponds to a power function of the force applied to the lever [8]. The device can be calibrated according to the individual sensitivity of the subject and to the range of variation of the stimuli tested by independently modifying two parameters: the mass and its distance from the axis of rotation. This consists of evaluating the matching function so that the amplitude of angular displacement of the device is maximal at the upper end of the range of stimulus variation for a given subject. The calibration phase and the individual scale transformation are described in Appendix A1.

The analogical/categorical device (A/C) is similar to the one developed by Weber [17]. A cursor connected to a potentiometer and mounted on a small rectangular box is displaced in a continuous manner by the subject along this scale. An output voltage corresponds to the position of the cursor and allows the continuous recording of the listener’s judgment. The device combines an analogical measure with several discrete category labels. In the version used in this study, the range of variation of the cursor was subdivided into seven categories [9]. The labels correspond to very, very loud – very loud – loud – medium – soft – very soft – very, very soft, but were presented in French (*très, très fort* to *très, très faible*). The scale is thus divided into six equal intervals considered to be perceptually equivalent. Weber had used five rather than seven categories, but it was felt that a greater resolution of the category scale would be useful in the present experiment. Indeed, the range of variation of the stimuli and the number of categories are factors upon which the stability of

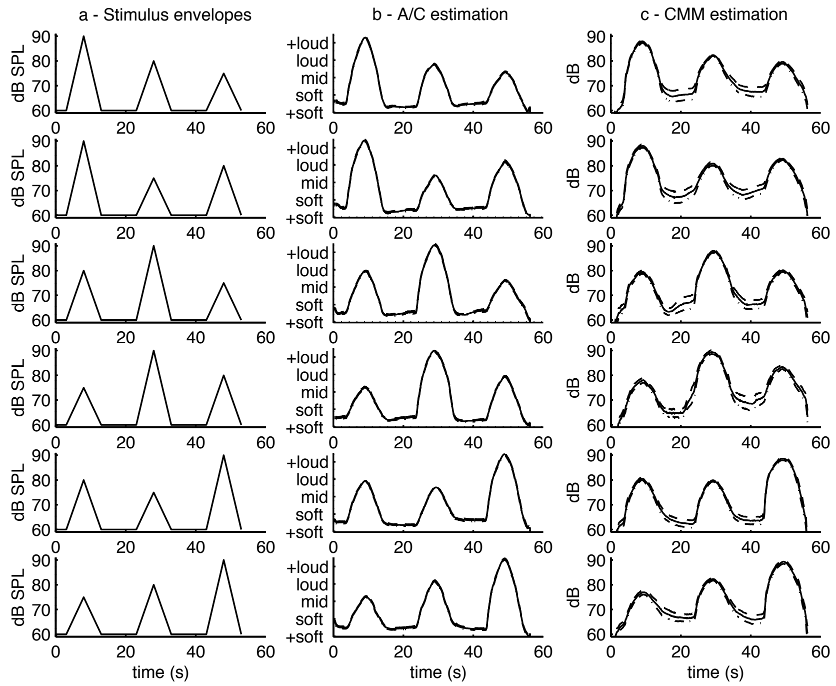

Figure 3. a) Acoustic level profiles in dB of 3PEAK stimuli (from top to bottom: HML, HLM, MHL, LHM, MLH, and LMH). Mean profiles obtained: b) by the analogical / categorical method (A/C) and c) by cross-modal matching (CMM). The dashed-line profiles represent ± one standard-error of the mean (solid lines) at each instant.

the results depends. If the signal varies between two successive categories, for example medium and loud, the subject must perform a mental division of the corresponding “semantic” distance in order to refine his or her response. On the other hand, this gap of uncertainty corresponding to the distance between two categories can be diminished by increasing the number of categories [21, 22].

In each of the experiments, the stimuli used were generated at a sampling rate of 44.1 kHz with 16-bit resolution by a NeXT workstation equipped with IRCAM’s ISPW digital signal processing card and Max software [23]. The sounds were converted by ProPort D-A converters before being amplified by a Canford stereo amplifier and presented over AKG 1000 open-air headphones. Subjects were seated in a Soluna S1 double-walled sound booth. Levels were calibrated using a Brüel and Kjær 2209 sound-level meter. The experiment was run using the PsiExp v2.5 experimentation environment including stimulus control, data recording, and graphical user interface [24].

2.3. Procedure

After having estimated the individual cross-modal psychophysical functions (see Appendix A1), the experiment took place in two stages corresponding to the two task conditions (with and without continuous judgment). For each condition, the experiment was performed with both CMM and A/C devices. The order of use of the two devices was counterbalanced across listeners.

Condition 1: Continuous evaluation plus global judgment. The subjects listened attentively to the sound sequences and continuously estimated the temporal evolution of their loudness with the device being used, associating at each moment a muscular force or position along the analogical/categorical scale equivalent to the perceived level. Once the sequence was finished, they performed an evaluation of the global loudness over its whole duration with the device by positioning the lever or cursor at an appropriate position and pressing a key on the computer keyboard. After the global evaluation, the subject triggered the presentation of the next trial by pressing a key. Stimuli were presented once each in random order.

Condition 2: Global judgment only. The subjects simply listened to the sound sequences and judged their global impression of loudness over the entire duration at the end of the sequence.

In each condition, the subjects performed six training trials. For each subject, two stimuli from each of the three contour types were chosen at random. Within each condition with a given device the sounds corresponding to the three contour types with different durations (RAMP and 1PEAK) or configurations (3PEAK) were presented in random order. Condition 1 was always presented before Condition 2 for subjects performing both.

2.4. Subjects

A group of 19 subjects participated in the experiment (11 men, 8 woman, mean age = 28, SD = 3). One subject performed the A/C method for both Conditions, but did not complete the CMM method. Another subject did not complete the CMM method and the A/C method for Condition 2 (see Table II). Data analyses are performed on all subjects within tasks and conditions, but when comparisons across Conditions are performed, only those subjects that completed both Conditions were included. No subject reported having hearing problems.

Table II. Subjects participating in the different experimental conditions.

	CMM	A/C
Condition I	S1–S18	S1–S19
Condition II	S1–S17	S1–S18

2.5. Data analyses

Several repeated-measures analyses of variance (ANOVA) were performed. To control for violations of sphericity that may arise with the use of repeated measures analyses (the same listener performs several comparisons across different stimulus conditions), the Greenhouse-Geisser ε was used to correct the degrees of freedom in the F test in order to determine the corrected probability that the comparison corresponded to the null hypothesis. F tests are cited with their original degrees of freedom, but if ε is less than one, its value is cited along with the corrected probability.

3. Results

3.1. Continuous judgments

The continuous judgment profiles obtained with the two devices (CMM, A/C) are presented in Figures 1–3 for RAMP, 1PEAK and 3PEAK stimuli, respectively. The profiles correspond to the mean of the individual judgments expressed as the position of the potentiometer along the 7-category scale for the A/C device and in equivalent dB (Appendix A1, equation A3) for the CMM device (see Note 1 below).

Globally, the profiles obtained for RAMP stimuli with the A/C device (Figure 1) have a form that is closer to the linear physical contours than those obtained with the CMM device, which have a convex curvature. The 3-s plateau at the beginning of the signal is generally absent with the CMM device. The same observation can be made for the 1PEAK stimuli. A notable asymmetry is found between increasing and decreasing ramps for both devices. This asymmetry is stronger for longer durations and is more pronounced for the CMM device. With the A/C device, the plateau at the end is more and more underestimated with respect to the starting plateau as the ramp duration increases (Figure 2). Almost no end plateau is observed with the CMM device. The judgment profiles for 3PEAK stimuli (e.g. MLH) are closer in form to the physical contours with the A/C device than with the CMM device. Indeed the plateaux obtained with the A/C device are equivalent, whereas those measured with the CMM device vary over time and are estimated with a greater variability across subjects than are the peaks of the stimuli (see Figure 3). The estimated difference between plateaux can attain 5 dB.

The individual reaction times were determined for the continuous judgment task. To analyze the reaction times, the continuous judgment profile and the stimulus contour were sampled at 100-ms intervals. The cross-correlation was calculated on these series with temporal lags at multiples of 100 ms from 0 to 3 s. The lag corresponding to the peak in the cross-correlation function was taken as the reaction time for a given subject. The calculation was performed for each stimulus and each subject. The global individual reaction time was obtained by averaging the set of reaction times across stimuli for a given subject. These reaction times varied from 0.3 to 1.4 s for the A/C device and from 0.6 to 2.0 s for the CMM device. Globally, for the A/C device the average reaction time across subjects is 0.9 s, and that for the CMM device is 1.1 s. The difference between A/C and CMM reaction times reveals an advantage for the former device, especially in the case of rapid fluctuations. The inertia of the CMM device is surely the cause of the longer delay observed. Another problem associated with the inertia of the system is fatigue due to the continuous muscular effort that must be exerted over a long duration. However, no characteristic effect related to muscle fatigue could be observed in a continuous task lasting several minutes (see Note 2 below).

3.2. Global judgments

Figures 4-6 present the global judgments obtained for RAMP, 1PEAK, and 3PEAK stimuli, respectively. In order to simplify the presentation, we have adopted the following notation: GJ1 and GJ2 are the global judgments for Conditions 1 and 2, respectively; Mean(CJ) is the arithmetic mean of the entire continuous judgment profile and Max(CJ) is the maximum value of the profile.

3.2.1. Condition 1: Continuous evaluation plus global judgment

Sustained attentive listening was required to perform the continuous loudness judgment task in this condition. For RAMP stimuli, the mean and maximum values of the continuous judgment profiles and the global judgment increase as a function of the duration of the level ramp (Figure 4). For example, although the acoustic range (60-80 dB) is identical for all durations, the maximum value of the continuous judgments with the CMM device expressed in equivalent dB progresses from around 79 to 83 dB as the ramp durations increase from 2 to 20 s. In an ANOVA comparing Mean(CJ) and GJ1 across the four ramp durations, the effect of ramp duration is highly significant for both devices (A/C: $F(3,54) = 26.2$, $p < 0.0001$, $\varepsilon = 0.84$; CMM: $F(3,51) = 11.7$, $p < 0.0001$, $\varepsilon = 0.75$). The same holds for a comparison of Max(CJ) and GJ1 (A/C: $F(3,54) = 15.9$, $p < 0.0001$, $\varepsilon = 0.82$; CMM: $F(3,51) = 6.8$, $p < 0.005$, $\varepsilon = 0.76$). In neither case did the duration effect interact with the type of judgment. A deviation of the range of the continuous judgments is thus revealed for both devices as a function of ramp duration for single-ramp stimuli. For the A/C device, neither the maximum nor the mean of the continuous judgments is equivalent to the global judgment which is always situated between the two: both Mean(CJ) and Max(CJ) are significantly different from GJ1 for the A/C device ($F(1,18) = 19.9$, $p < 0.0005$, and $F(1,18) = 44.4$, $p < 0.0001$, respectively). For the CMM

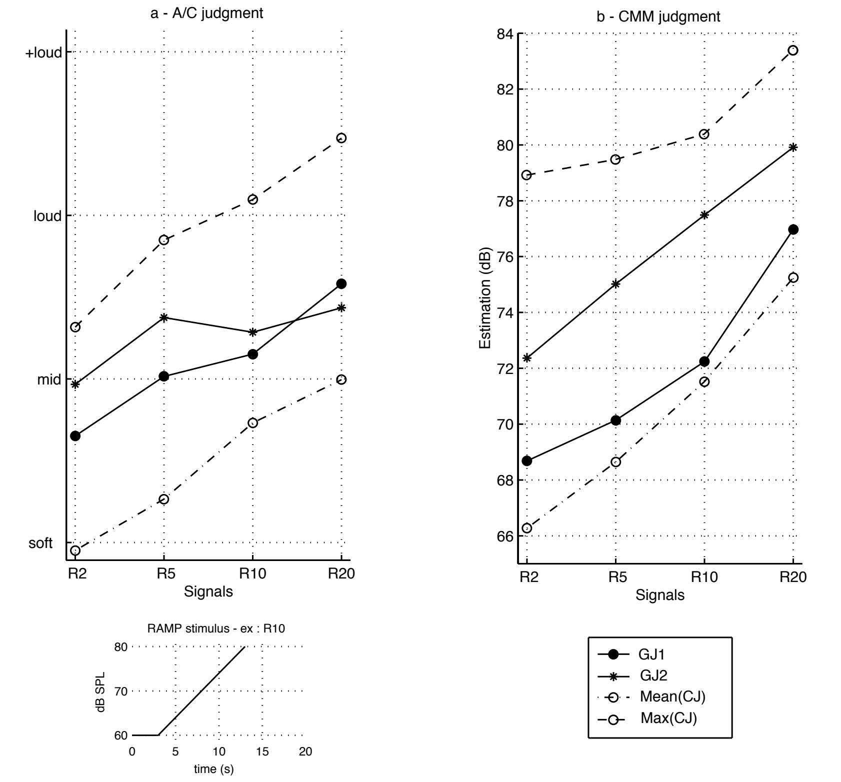

Figure 4. Summary data for RAMP stimuli (R2, R5, R10, and R20) obtained with the A/C device (left panel) and the CMM device (right panel). The average individual means (Mean(CJ)) and maxima (Max(CJ)) of the continuous judgment profiles, as well as the global judgments for Conditions 1 (GJ1: continuous plus global judgment) and 2 (GJ2: global judgment alone), are shown for both devices.

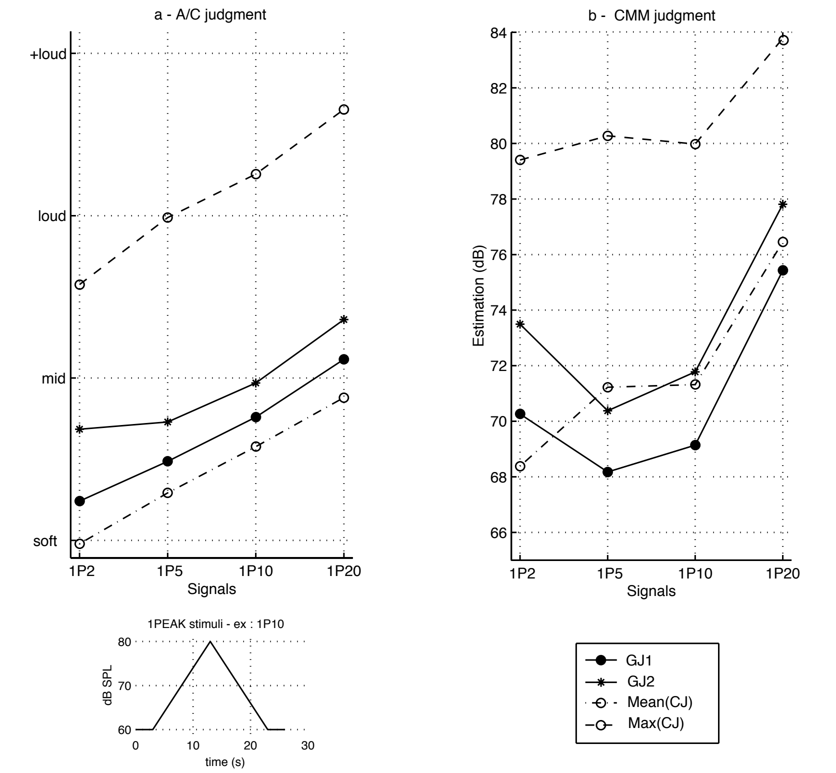

Figure 5. Summary data for 1PEAK stimuli (1P2, 1P5, 1P10, and 1P20) obtained with the A/C device (left panel) and the CMM device (right panel). See Figure 4 caption.

device, GJ1 is globally less than Max(CJ) (F(1,17) = 34.4, $p < 0.0001$), but is equivalent to Mean(CJ) (F(1,17) = 1.8, NS).

The results obtained for 1PEAK stimuli are globally similar in nature to those for RAMP stimuli (Figure 5) in the sense that Max(CJ), Mean(CJ), and GJ1 increase

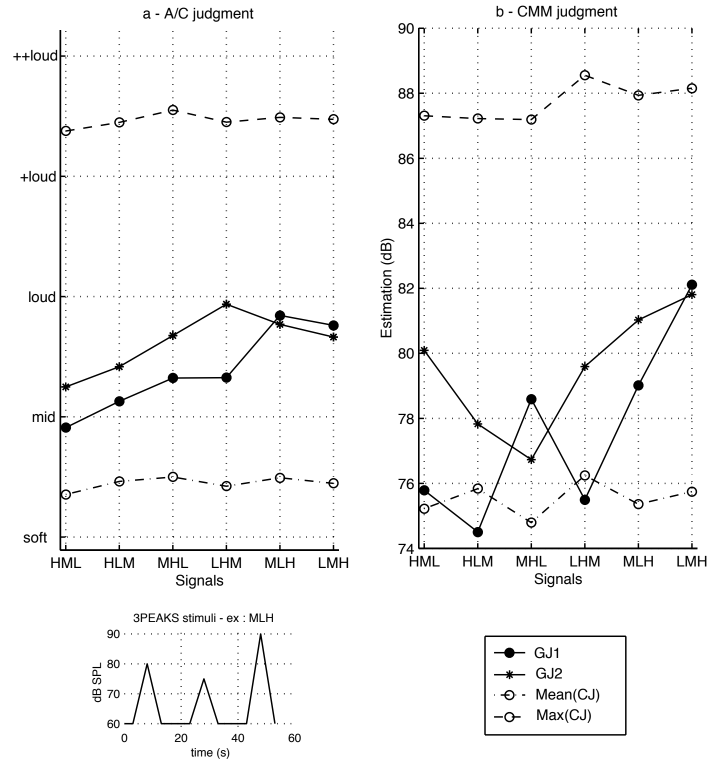

Figure 6. Summary data for 3PEAK stimuli (HML, HLM, MHL, LHM, MLH, and LMH) obtained with the A/C device (left panel) and the CMM device (right panel). See Figure 4 caption. Although Configuration is not a continuous variable, the points are connected to facilitate visual comparison of the relations between measures and conditions.

overall as a function of the duration of the ramps. The exceptions to this trend are a plateau for 1P5 and 1P10 in the Mean(CJ) data and an inexplicably higher estimate for 1P2 than for 1P5 and 1P10 in the GJ1 data. In a comparison of Mean(CJ) and GJ1, the global effect of duration is significant for both devices (A/C: $F(3,54) = 19.0$, $p < 0.0001$, $\varepsilon = 0.95$; CMM: $F(3,51) = 5.9$, $p < 0.01$, $\varepsilon = 0.63$) and no interaction with the judgment type is revealed. The same pattern is found for the comparison of Max(CJ) and GJ1. The global judgment for each stimulus is not significantly different from Mean(CJ) with both devices (A/C: $F(1,18) = 2.2$, NS; CMM: $F(1,17) < 1$, NS), but is significantly lower than Max(CJ) (A/C: $F(1,18) = 59.3$, $p < 0.0001$; CMM: $F(1,17) = 30.0$, $p < 0.0001$).

The 3PEAK stimuli all have identical Leq values of 74 dB SPL over the whole duration of the sound sequence and are composed of three peaks of differing maximum levels in all possible permutations. The Mean(CJ) and Max(CJ) values remain fairly constant for all six configurations (Figure 6).

An overall significant difference among configurations for global judgments following continuous estimation is found only for the A/C device (A/C: $F(5,90) = 7.9$, $p < 0.005$, $\varepsilon = 0.62$; CMM: $F(5,85) = 2.1$, NS, $\varepsilon = 0.36$). Planned contrasts on the effect of configuration with GJ1 as dependent variable reveal that stimuli with the highest peak (H) at the last position are judged as having a global loudness significantly higher than those with H in the second position for the A/C device ($F(1,90) = 13.9$, $p < 0.005$, $\varepsilon = 0.62$) but not for the CMM device ($F(1,85) = 3.2$, NS, $\varepsilon = 0.36$). For both devices, H in last position gives higher estimates than H in first position (A/C: $F(1,90) = 37.3$, $p < 0.0001$, $\varepsilon = 0.62$; CMM: $F(1,85) = 7.5$, $p < 0.05$, $\varepsilon = 0.36$). The difference between stimuli with H in the first and second positions is not significant for the CMM device ($F(1,85) < 1$, NS, $\varepsilon = 0.36$), but is significant for the A/C device ($F(1,90) = 5.6$, $p < 0.05$, $\varepsilon = 0.62$). There are no secondary effects due to the relative position of M and L peaks for a given position of the H peak ($p > 0.1$ in all cases). It thus appears that the global judgment varies as a function of the position of the highest peak in the sequence with more marked differences appearing for the A/C device compared to the CMM device.

However, as Figure 7 shows, this lack of effect for the CMM device is primarily due to unusually low values in three configurations for two subjects (S12, S14), as well as an unusually high value for S12 in one configuration. The results of all the other 16 subjects follow a more coherent pattern consistent with the recency effect mentioned in the introduction (higher global estimation with higher levels near the end of the sequence). A parallel set of ANOVAs was thus performed without the data of these two subjects for the CMM device. In the new analyses, the effect of configuration is significant ($F(5,75) = 6.2$, $p < 0.001$, $\varepsilon = 0.66$), and planned contrasts now reveal significant differences between configurations with the high peaks in first and second position ($F(1,75) = 9.5$, $p < 0.01$, $\varepsilon = 0.66$) and in second and third positions ($F(1,75) = 5.7$, $p < 0.05$, $\varepsilon = 0.66$). There are still no secondary effects of the relative position of medium and low peaks (all p's > 0.35).

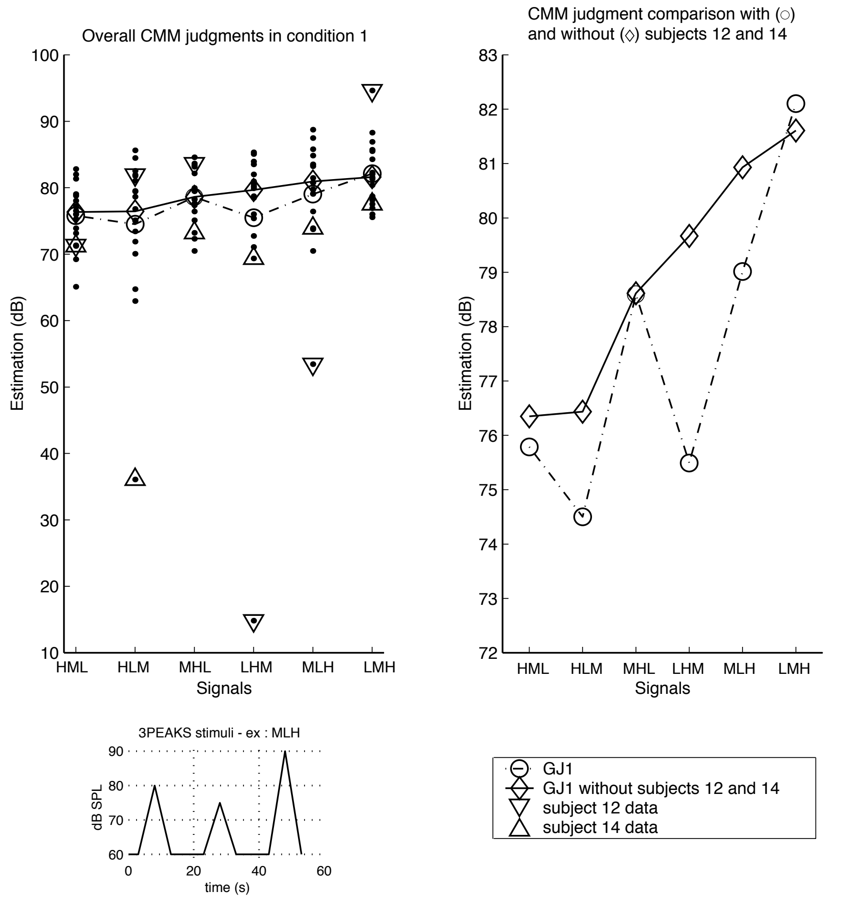

Figure 7. Data for 3PEAK stimuli with the CMM device. Left panel: data for individual subjects, with means for all subjects (open circles) and excluding S12 and S14 (open diamonds, the data for these subjects are indicated with open triangles). Right panel: mean data with and without S12 and S14 on a smaller-ranged ordinate to show the emergence of the recency effect (higher estimates when the H peak is near the end of the sequence).

The effects of configuration on measures derived from the continuous judgments and on the global judgment are significant for the Mean(CJ)/GJ1 comparison for the A/C device ($F(1,18) = 53.0$, $p < 0.001$), but not for the CMM device ($F(1,17) = 2.5$, NS) when the data for S12 and S14 are included. Without these subjects the Mean(CJ) is lower on average than GJ1 ($F(1,15) = 11.6$, $p < 0.005$). Max(CJ) values are higher on average than GJ1 values for both devices (A/C: $F(1,18) = 186.0$, $p < 0.0001$; CMM: $F(1,17) = 45.2$, $p < 0.0001$). The difference between the global judgments on stimuli HML and LMH is equivalent to 6 dB with the CMM device and to one category with the A/C device. The value of the global judgment for stimuli with the highest peak at the beginning of the sequence (HML and HLM) is equivalent to Mean(CJ) for the CMM device. At the other extreme, the stimuli with the highest peak at the end of the sequence (LMH and MLH) give rise to global judgments that are between the mean and maximum of the continuous judgments for this device.

3.2.2. Condition 2: Global judgment alone

In Condition 2, the subject only gave a global judgment after listening to the entire sequence, not being constrained to pay particular attention to the instantaneous variations in the signal. The data are included in Figures 4–6. In general, the judgments obtained are slightly higher than or equal to those for Condition 1. A repeated-measures ANOVA comparing GJ1 to GJ2 was performed for all three stimulus classes. For RAMP stimuli, GJ1 was lower than GJ2 with the CMM device ($F(1,16) = 6.5$, $p < 0.05$), but equivalent with the A/C device ($F(1,17) = 3.2$, NS). For 1PEAK stimuli, GJ1 was equivalent to GJ2 for the CMM device ($F(1,16) = 1.9$, NS), but was lower for the A/C device ($F(1,17) = 6.2$, $p < 0.05$). For 3PEAK stimuli, the data seem quite different in form, but the differences are not reliable statistically (CMM: $F(1,16) < 1$, NS; A/C: $F(1,17) = 4.1$, $p = 0.059$). In no case was there a significant interaction with duration or configuration. In summary, the results for Condition 2 without continuous evaluation are equivalent to (although at times higher than) those found for Condition 1, although the differences are somewhat less contrasted, and the inter-subject variability is somewhat larger. (On average with the CMM device, the standard deviation is 1.3 and 2.2 dB for Conditions 1 and 2, respectively).

3.2.3. Summary of principle findings

The data obtained for RAMP stimuli tend to increase in value with duration. The relations among the measures derived from the continuous judgment profiles and the global judgments are Mean(CJ) < GJ1 < GJ2 < Max(CJ) for both devices.

The tendency for higher estimates to accompany longer ramp durations is globally similar for 1PEAK stimuli, although there are a few inversions and equivalences with the CMM device. A higher degree of variability across subjects is found with the CMM device than with the A/C device. The relations among measures are Mean(CJ) $<$ GJ1 $<$ GJ2 $\ll$ Max(CJ) with the A/C device and Mean(CJ) $\approx$ GJ1 $<$ GJ2 $\ll$ Max(CJ) with the CMM device.

The values obtained for 3PEAK stimuli tend to increase as a function of the position of the highest-level peak (H) in the sequence, this trend being quite clear if the data of two subjects with erratic data are removed. This effect is more pronounced for Condition 1 than for Condition 2 and more pronounced for the CMM device than for the A/C device. While the relative temporal position of the low and medium peaks seems to have an effect on the global judgment for a given position of the high peak for Conditions 1 and 2, the effect is not statistically reliable. Whatever the device used, the maximum and mean values of the continuous judgment profiles vary little across stimulus configurations indicating little sensitivity of these measures to the temporal order of events, but also an accurate tracking of the instantaneous loudness of the time-varying signal. In Condition 2 with the A/C device, the values are lower when H is in first position, but then level off at a higher value for the second and third positions. The relations among measures across conditions are Mean(CJ) $<$ GJ1 $\leq$ GJ2 $\ll$ Max(CJ) with the A/C device and Mean(CJ) $\leq$ GJ1$\approx$GJ2 $\ll$ Max(CJ) with the CMM device.

4. Discussion

The cross-modal matching (CMM) method with force feedback, as well as the analogical/categorical (A/C) method without force feedback were used to estimate the loudness judgment profiles associated with temporal fluctuations of level in a pure tone. Continuous and global estimates of the loudness of three classes of sound sequences with time-varying levels of a 1-kHz pure tone were made with two methods under two judgment conditions. The level contours of the stimuli included upward ramps, single peaks (upward then downward ramps) and triple peaks. The ramp duration was varied for the ramp and single peak stimuli as were the temporal configurations of stimuli with three peaks of differing maximum levels. In one judgment condition, continuous evaluation was made followed by a global estimate at the end of the sound sequence. In the second condition, only a global estimate was made at the end. A comparison of the devices, based on the data obtained, is found in Appendix A2, since the main thrust of the discussion here concerns the effects of judgment condition and the effects of temporal structure on global judgments.

Generally, a small temporal lag between the physical level contours and the judgment profiles was found and was more or less pronounced for different subjects. Taking into account the lags of 1.1 s and 0.9 s, the correlations between mean judgment profiles and level contours are 0.94 and 0.97 for the CMM and A/C devices, respectively. The results obtained by Weber [17] showed globally that these kinds of lags vary from 0.5 to 1.5 s. These values are similar to those for individual subjects found in this study using the same type of method, as well as the reaction time estimated by Kuwano and Namba [1] at 1 s. The reaction time values found in this study are also similar to the 1-s (±0.15-s) reaction time obtained recently by Hansen and Kollmeier [19] in continuous evaluation of time-varying speech quality.

Whatever the method, the correspondence varies considerably with duration. Indeed the maximum and mean values of the continuous judgments increase with the duration of the linear level ramp for an identical dynamic range (60–80 dB SPL). A sort of level drift in the continuous judgments appears as a function of the ramp duration for both devices. The global judgment values also increase in this way for all experimental conditions tested. This result holds whether or not a continuous judgment task precedes the global judgment. This latter point indicates that the effect of drift in the global judgment values is not a consequence of the continuous judgment task. However, the nature of the mechanism responsible for this drift in the auditory integration processes involved in making a global judgment remains to be determined.

One of the principle objectives of the experiments performed was to verify, in a global judgment task, the hypothesis concerning the existence of a memory effect similar to the recency effect observed in serial recall tasks in memory research. The highest values drawn from the recall curve in an immediate recall task after presentation of a list of items (letters, words or numbers, in general), corresponds to the first and last items of the list. The two processes that underly this result give rise to the primacy and recency effects, respectively. A partial explanation of the primacy effect is that the first items in the list are reinjected by subvocalization (a mechanism of articulatory self-repetition) a greater number of times than the others. This rehearsal is thought to allow a transfer of information from short-term to long-term memory. The nature of the stimuli used in our experiments do not reveal a primacy effect related to the beginning of the signal, most likely because such effects are generally related to long-term semantic memory rather than to a short-term acoustic store. However, the data collected for the 3PEAK stimuli show significant differences among the stimuli as a function of the temporal position of the highest-level peak in line with the predictions of the recency effect: the global judgment is greater if the H peak moves toward the end of the stimulus sequence. The results are less pronounced with the A/C device, perhaps because this device requires a continuous translation of the visually presented A/C scale in order to make a correspondence between the sensory impression and its rating, making the relation between sensation and judgment more indirect. Nonetheless, the position of the H peak in the sequence seems globally to be a dominant factor producing the recency effect. This result confirms the hypothesis, proposed by Hoeger *et al.* [25], of the recency effect in loudness judgments. In their experiment, they used 1kHz pure tones with different time structures and with the same equivalent sound levels. Their results showed that the the information at the end of the sound is much more salient than the sound events heard earlier. This result is similar when the sound's duration is 54 s and 7.2 min.

On the other hand, Springer *et al.* [26] have recently performed an experiment in which high-level events were concentrated, according to three distributions, at the beginning, in the middle, or at the end of a 10-minute pink-noise sequence. However, the results showed no significant difference between the stimuli for the global judgment, suggesting that no effect of recency was brought into play in the experimental task. In contradiction to the recency effect, it may be that listeners could have identified the different configurations and understood that their only distinguishing factor was a temporal displacement of the energy peaks. As such, they would make similar judgments independently of the configuration.

The data obtained for the RAMP stimuli show an ”attraction” effect of the global judgment for the final maximum of the continuous judgment profiles. The data for 1PEAK stimuli are closer to the mean of the continuous judgment profiles for the symmetric level contours. Finally, it should be emphasized that the variations of the judgments (”attraction” effect, recency effect) are mostly independent of the experimental protocol.

It should be noted that Kuwano and Namba [1] on the one hand, and Fastl [2] on the other, have proposed two procedures for calculating an index that conveys the global judgment values they obtained for sound sequences derived from urban soundscapes. The calculation of the index is based on the hypothesis that the events having the predominant sound level determine the global judgment which partially fits with our results. Indeed, the global judgments of the 3PEAK stimuli depend on the dominant event in loudness (H) but also depend significantly on its temporal position in the sequence. In our case, the measures obtained with both indices would predict identical global judgments for all six configurations of the 3PEAK stimuli, since their indices do not take into account the temporal distribution of the energy in the sound sequence. This difference is perhaps related, on the one hand, to the duration of the signals, which are clearly shorter than those studied by Kuwano and Namba [1] and Fastl [2] (20 min), and on the other hand, to the nature of the signals which have meaning for the listeners in the cited studies and thus call more on long-term memory mechanisms. The explanation related to the duration is not satisfactory because it seems that in serial recall experiments, when the list of items to be recalled is lengthened, the primacy effect disappears but the recency effect remains [27]. The recency effect thus seems to be independent of the length of the list and most likely of the stimulus duration as well.

5. Conclusions

The stimuli, judgment methods, and judgment conditions revealed two specific effects in loudness estimates of nonstationary sounds. Firstly, the results bring to light shifts in instantaneous and global judgments that depend on the duration of the signals, with globally higher estimates being made for longer-duration signals. Secondly, a non-negligeable recency effect is clearly present as evidenced by differences in the temporal distribution of energy over sound sequences of identical total energy. The latter effect seems primarily to be related to the temporal position of the highest contour peak, higher loudness estimates resulting from later presentation of the highest peak. In elaborating a model of global loudness judgment of long- duration, nonstationary sound sequences, it will be necessary to take into account a combination of the highest levels, their temporal position with respect to the moment of global judgment, and their duration of emergence.

Notes

1. With the force feedback (CMM) device, the applied force depends both on the angular displacement and angular acceleration. The force applied to the mass-rod system by the subject in the case of a dynamic movement is expressed by the sum of a displacement term and a term related to the inertia of the system that is proportional to the acceleration:

$$\begin{aligned} \left|R(t)\right| &= F_1(t) + F_2(t), \\ F_1(t) &= \frac{mgl}{a}\sin\alpha(t), \\ F_2(t) &= \frac{ml^2}{a}\ddot{\alpha}(t), \end{aligned} \tag{1}$$

with a being the distance between the axis of rotation and the lever handle, l the distance between the axis of rotation and the mass m, and α the displacement angle. An examination of the acceleration curves in the temporal profiles obtained in the present study reveals a peak at the beginning of a level increase that corresponds to a sudden variation in the matching function for a quick change in intensity. This effect brings up a potential disadvantage of the CMM device for rapid variations. However, the amplitude of the peak diminishes as a function of the duration of the ramp. In general, for the stimulus conditions under study here, the curves reveal that the acceleration term (F_2) of the movement is negligeable compared with the displacement term (F_1), and all the more so as the ramp duration increases.

2. An experiment was performed to estimate the effect of fatigue over 7.5 minutes. The sound level was varied over 90 plateaux of 5 s each. Six levels between 60 and 75 dB SPL were presented and repeated randomly 15 times each over the whole sequence. Individual results sometimes show an intra-subject variability at the lower levels (60 dB) that could attain force matching differences equivalent to 10–15 dB for some subjects as compared to differences of 3–5 dB for the higher levels (75 dB). The data did not, however, reveal any systematic change over time that could be interpreted as an effect of fatigue, at least over the 7.5-minute duration of this control experiment.

Appendix

A1. Individual calibration of the CMM device

To compare the judgments obtained across subjects, individual psychophysical functions for the CMM device were determined for each subject. In the calibration phase, the stimuli used were 11 stationary 1-kHz pure tones the levels of which varied from 60 to 90 dB SPL in 3 dB steps. The duration of each sound was 1 s with 40-ms linear attack and decay ramps.

The apparatus was the same as that of the main experiment. The subject triggered the beginning of the experiment. A sound of constant level was presented once every two seconds. The subject moved the lever to a position at which he or she judged the force to have an intensity equivalent to that of the sound's loudness and then entered the response by hitting the ”V” key on the keyboard with the other hand. The level of each sound and the corresponding angle of the lever were recorded by the program. The sounds were presented in random order for each subject. A session lasted one minute on average. It was repeated at least two to three times in the presence of the experimenter in order to familiarize the subject with the procedure and to adjust the device's resistance (mass and/or distance of the mass from the axis of rotation of the lever) as a function of the subject's reactions. The adjustment phase consisted of observing whether the subject used a large range of angular displacement of the lever as a function of the stimulus range. An attempt was made to associate the highest level with a displacement near the maximum of the lever (90°). When the subject felt at ease with the calibrated device, a last series was performed in the presence of the experimenter. The data obtained from this series were used to determine the individual psychophysical function of the subject relating force to acoustic pressure for stationary sounds.

The psychophysical power function obtained by S. S. Stevens [28], expressed in its logarithmic form for loudness (l) and apparent force (f), are given in equations (A1) and (A2), respectively:

$$\log_{10}\psi_l = a_l \log_{10}\phi_l + \log_{10} k_l, \tag{A1}$$

$$\log_{10}\psi_f = a_f \log_{10}\phi_f + \log_{10} k_f, \tag{A2}$$

where ψ corresponds to subjective intensity, ϕ corresponds to physical intensity, k is an arbitrary constant to adjust the scale, and a is the exponent that depends on the sensory modality and conditions of stimulation. This function seems to be valid for a large set of sensory modalities [28, 29].

Matching apparent force (ψ_f) and perceived level (ψ_l) corresponds to the following relation:

$$\log_{10}\phi_f = a \log_{10}\phi_l + k, \tag{A3}$$

where $a = \dfrac{a_l}{a_f}$ and $k = \dfrac{\log_{10} k_l - \log_{10} k_f}{a_f}$.

So a linear relation is obtained that expresses the log of the force in Newtons as a function of the log of the acoustic pressure in μPa. For each subject i, the individual matching function is obtained by the procedure described above. The form of the individual functions and their respective regression coefficients show a good correspondance between the data and the associated power function for each subject. Mean a values are 0.46 ± 0.17 (standard deviation). Mean k values are -1.6 ± 1.7. The variance explained by the power functions varies from 67% to 98% (mean 90%). The (a, k) pair for each subject allows the definition of an individual scale transformation between force and acoustic pressure. The judgments obtained in the form of “equivalent” forces in Newtons can thus be transformed into acoustic pressure and represented in dB for each subject in the data analyses in the main experiment.

A2. Comparison of the CMM and A/C methods

Globally, the two methods, based on different kinds of mechanical devices, provide fairly similar profiles over the set of stimuli tested. The global judgment values show similar tendencies as a function of the stimulus configurations tested irrespective of the device and the listening condition under consideration. Nonetheless, a closer examination reveals differences that are inherent in each device and in individual judgment strategies. Further, the form of the profiles is asymmetric for continuous estimations of increasing and decreasing levels. These differences will be discussed below. In Table A1, we summarize different comparisons between the two methods used for real-time and global evaluation of loudness contours in the present data.

Two other characteristics distinguish the two methods: the response type, depending on the manipulation of the device, and the data representation scale. Concerning the A/C device, the manipulation of the potentiometer introduces negligeable force feedback, which particularly facilitates its use for sounds with rapid fluctuations. However, for slow fluctuations, between 10 and 20 s (e.g. stimuli R10, R20, 1P10, and 1P20), the individual profiles split into two groups. Some profiles correspond to continuous curves whereas others are in the form of a staircase. The interpretation made by the subject is a function of the acoustic levels of the stimuli associated with the semantic descriptors of the scale. This latter result indicates the disadvantage of the combination of the two kinds of scales on the same device for very slowly changing sounds. It should be noted that the same type of characteristic shows up only very weakly in the responses of certain subjects with the CMM device.

Finally, the results reveal a limit in the CMM method in this continuous judgment task depending on the duration of fluctuation of the signals to be estimated. For variations that are too quick, the matching function obtained in the calibration phase does not convey exactly the continuous judgments. In this case, an adjustment of the func-

Table A1. Comparison between analogical/categorical and cross-modal matching methods.*: Responses derived in part from interviews with subjects at the end of the experiment

Characteristics	A/C Analogical/categorical device	CMM Proprioceptive device
Possibility for individual calibration	No	Yes (duration ≈ 5 minutes)
Type of continuous judgement	Analogical (at times staircase)	Analogical
Representation scale	Semantic scale interpreted as a function of acoustic level in dB	Equivalence between estimated level in dB and acoustic level in dB
Quality of reproduction of profiles	Good	Medium
Variability of profiles	Low	Variable in stationary parts
Mean reaction time (s)	0.9	1.1
Difference between rising and falling levels	Increases with duration	Increases with duration
Resolution of the scale	Low to high depending on use of scale (cf. Type of continuous judgment)	High
Disadvantages	Uncertainty between two adjacent judgment categories. Two types of reponses: analogical and staircase	Calibration and learning phases necessary. Possibility of fatigue for long-duration tests.
Advantages	Rapid use without apparent fatigue	Direct correspondence between the measured sensation and the response
Subjects appreciation*	Good – Not enough information between categories	Good - Intuitive
Type of sequences*	Rapid – Urban sound sequences	Slow – Vehicle acceleration

tion needs to be considered. For signals of longer duration, the subjects' responses are asymmetric with respect to the direction of variation of the device. This asymmetry increases with duration. Another type of adjustment thus needs to be considered as well. Note that this asymmetry also appears in the curves obtained with the A/C device, although it is stronger with the CMM method. Consequently, it seems that this effect persists, whatever the method employed. Indeed, the matching function is not identical depending on whether the force is applied to the device in pushing or retaining mode [8]. In the latter case, it produces an underestimation of the acoustic level. The results thus show that a pure tone with a continuously decreasing level is underestimated compared to the same sound with an increasing level. A similar effect has been observed in comparing a pure tone presented at different levels with a similar sound but whose level decreases continuously. For an equivalent level, the continuous sound is underestimated compared with the stationary sound. This effect is called auditory "decruitment" [30, 31]. It would thus seem possible that the observed asymmetry is not specific to the methods used here, but is the consequence of an inherent auditory asymmetry.

An important advantage of the CMM method is that it allows one to obtain a data representation on a dB scale and thus provides an immediate correspondence between the stimulus level tested and the continuous judgment profiles. Over the set of stimuli in this study, a good correspondence between the equivalent level in dB derived from the force exerted and the the acoustic level in dB SPL was found.

Acknowledgement

We are grateful to Anne Bardot (PSA Peugeot Citroën) and Etienne Parizet (formerly of Renault) for their collaboration in this project. We thank the two reviewers for their fruitful comments on an earlier version of this article.

References

[1] S. Kuwano, S. Namba: Continuous judgment of level-fluctuating sounds and the relationship between overall loudness and instantaneous loudness. Psychol. Res. **47** (1985) 27–37.

[2] H. Fastl: Evaluation and measurement of perceived average loudness. – In: Fifth Oldenburg Symposium on Psychological Acoustics. A. Schick, J. Hellbrück, R. Weber (eds.). BIS, Oldenburg, 1991, 205–216.

[3] E. Zwicker, K. Deuter, W. Peisl: Loudness meters based on ISO 532 B with large dynamic range. InterNoise 85, 1985; Proceedings of the 1985 International Conference on Noise Control Engineering, 1985, 1119–1122.

[4] J. B. B. Murdock: The serial position effect of free recall. Journal of Experimental Psychology **64** (1962) 482–488.

[5] R. Crowder, J. Morton: Precategorical acoustic storage (PAS). Percept. Psychophys. **5** (1969) 365–373.

[6] M. Anisfeld, M. E. Knapp: Association, synonymity, and directionality in false recogntion. J. Exp. Psychol. **77** (1968) 171–179.

[7] A. D. Baddeley: The influence of acoustic and semantic similarity on long-term memory for word sequences. Quart. J. Exp. Psychol. **18** (1966) 302–309.

[8] P. Susini, S. McAdams: Psychophysical validation of a proprioceptive device by cross-modal matching of loudness. Acustica/Acta Acustica **86** (2000) 515–525.

[9] S. Kuwano, S. Namba: On the loudness of road traffic noise of longer duration (20 min) in relation to instantaneous judgment. J. Acoust. Soc. Am. **64** (1978) 127–128.

[10] S. Namba, T. Kato, S. Kuwano: Long-term evaluation of the loudness of train noise in laboratory situation. – In: 15th International Congress on Acoustics. M. Newman (ed.). Trondheim, Norway, 1995, 215–218.

[11] S. Kuwano, S. Namba: Continuous judgement of loudness and annoyance. Fechner Day, Proceedings of the 6th Annual Meeting of the International Society for Psychophysics, Würzburg, Germany, 1990, 129–134.

[12] S. Kuwano, S. Namba, T. Hato, M. Matui, K. Miura, H. Imai: Psychological evaluation of noise in passenger cars: Analysis in different groups of subjects in nationality, age and gender. – In: A. Schick. S. O. S. on Psychological Acoustics (ed.). BIS, Oldenburg, 1993, 521–536.

[13] H. Fastl: Average loudness of road traffic noise. International Conference on Noise Control Engineering, Inter Noise 89, 1989.

[14] G. Gottschling: On the relations of instantaneous and overall loudness. Acustica/Acta Acustica **85** (1999) 427–429.

[15] C. K. Madsen: Empirical investigation of the ”aesthetic response” to music: Musicians and nonmusicians. Fourth International Conference on Music Perception and Cognition, Montreal, Canada, 1996, 103–110.

[16] E. Schubert: Continuous response to music using the two dimensional emotion space. Fourth International Conference on Music Perception and Cognition, Montreal, Canada, 1996, 263–268.

[17] R. Weber: The continuous loudness judgement of temporally variable sounds with an ”analog” category procedure. – In: Fifth Oldenburg Symposium on Psychological Acoustics. A. Schick, J. Hellbrück, R. Weber (eds.). BIS, Oldenburg, 1991, 267–294.

[18] D. Hedberg, C. Jansson: Continuous rating of sound quality. Technical Audiology, Karolinska Institutet. Report TA 134, 1998.

[19] M. Hansen, B. Kollmeier: Continuous assessment of timevarying speech quality. J. Acoust. Soc. Am. **106** (1999) 2888–2899.

[20] S. Namba, S. Kuwano, T. Hatoh, M. Kato: Assement of musical performance by using the method of continuous judgement by selected description. Music Perception **8** (1991) 251–276.

[21] A. Parducci: Category ratings: Still more contextual effects. – In: Social Attitudes and Psychological Measurement. B. Wegener (ed.). 1982.

[22] A. Parducci, D. H. Wedell: The category effect with rating scales: Number of categories, number of stimuli, and method of presentation. J. Exp. Psychol.: Human Percept. Perf. **12** (1986) 496–516.

[23] E. Lindemann, M. Puckette, E. Viara, M. De Cecco, F. Dechelle, B. K. Smith: The architecture of the IRCAM musical workstation. Comput. Mus. J. **15** (1991) 41–49.

[24] B. Smith: Psiexp: an environment for psychoacoustic experimentation using the IRCAM musical workstation. Society for Music Perception and Cognition Conference’95, University of California, Berkeley, 1995.

[25] R. Höger, E. Matthies, E. Letzing: Physikalische versus psychologische Reizintegration: Der Mittelungspegel aus wahrnehmungspsychologischer Sicht. Zeitschrift für Lärmbekämpfung **35** (1988) 163–167.

[26] N. Springer, A. Schick, R. Weber: Instantaneous and overall loudness of temporally variable pink noise. – In: Seventh Oldenburg Symposium on Psychological Acoustics. A. Schick, M. Klatte (eds.). BIS, Oldenburg, 1997, 91–98.

[27] M. Glanzer, A. Schwartz: Mnemonic structure in freerecall: Differential effects on STS and LTS. Journal of Verbal Learning and Verbal Behavior **10** (1971) 194–198.

[28] S. S. Stevens: On the psychophysical law. Psychol. Rev. **64** (1957) 153–181.

[29] S. S. Stevens: Psychophysics: Introduction to its perceptual, neural and social prospects. Wiley, New York, 1975.

[30] G. Canévet, B. Scharf: The loudness of sounds that increase and decrease continuously in level. J. Soc. Am. **88** (1990) 2136–2142.

[31] R. Teghtsoonian, M. Teghtsoonian, G. Canévet: The perception of waning signals: Decruitment in loudness and perceived size. Perception and Psychophysics **62** (2000) 637–646.