Overall judgment of loudness of time-varying sounds

Josef Schlittenlacher[1]

Department of Experimental Psychology, University of Cambridge, Downing Street, Cambridge CB2 3EB, England

Takeo Hashimoto

Seikei University, 3-3-1 Kichijoji-Kitamachi, Musashino-shi, 180-8633 Tokyo, Japan

Sonoko Kuwano and Seiichiro Namba

Osaka University, 2-1 Yamadaoka, Suita, 565-0871 Osaka, Japan

Submitted to the Journal of the Acoustical Society of America on 08/15/2017

Running heading: Overall judgment of loudness

[1] js2251@cam.ac.uk; Part of this work was performed at his previous affiliation, Institut für Psychologie, Technische Universität Darmstadt, Alexanderstraße 10, 64283 Darmstadt, Germany

Abstract

Listeners can judge the overall loudness of time-varying sounds quite easily, i.e. assign a single value that represents the loudness of the entire sound. This holds even if the duration is long and the judgment includes memory effects.

Different metrics for calculating overall loudness have been developed. They agree that overall loudness is higher than the mean of loudness over time. Percentiles like the N_5, the loudness being exceeded 5 percent of the time, are adopted by ISO 532-1.

In the present study the concept of an energy mean known from level measurements (ISO 1996-1) was applied to the loudness domain. This equivalent continuous loudness level, LL_P, was compared to the N_5 using a set of real-world sounds that was orthogonal between the two metrics over a wide dynamic range of 30 phon. Cross-modality matching with line length was used in three experiments with a focus on either the overall judgment of loudness, continuous judgment while a sound was played, or both. The LL_P showed considerably higher correlations with overall judgments than N_5. Comparing continuous instantaneous judgment with calculated instantaneous loudness suggests that the participants might have focused on the sounds' prominent portions.

I. INTRODUCTION

The present study aims to examine the most appropriate metric to evaluate the overall loudness of temporally varying sounds. In particular, it compares the N_5, a percentile, to the LL_P, which is the energy mean of loudness levels.

Auditory perception originated from recognizing the world with temporal variation. An object can be recognized and differentiated from other objects since every object has its own characteristic temporally varying pattern. We can make instantaneous judgments on loudness and pitch, which are attributes of auditory sensation, by following the momentary variation. From this variation, we can recognize the movement and the position of a sound source. In the case of music, momentary changes in loudness can also be an expression of performance. There are many research publications concerning such variations over time by measuring instantaneous loudness (e.g. Kuwano and Namba, 1985).

Although real-world sounds usually vary significantly over time, it is quite easy for a listener to assign a single representative loudness rating to them. This single value is called overall loudness. An example from daily life is adjusting the loudness of a music-playing device. Overall loudness can also be extremely useful in technical applications, for example when comparing noisy environments or when measuring sound quality.

The international standard for the evaluation of environmental noise, ISO 1996-1:2016, has been used to quantify the long-term effects of loudness and annoyance of environmental noises. This standard uses the adjusted A-weighted equivalent continuous sound pressure level (L_{Aeq}), which is the A-weighted mean energy level. The choice of A-weighted level rather than a loudness-based metric was explained by the state of technology when the standard was first published. The L_{Aeq} was also supported by several psychoacoustic experiments. These used sounds of various durations and temporal patterns as stimuli, and found that the L_{Aeq}

consistently showed higher correlation with loudness than other statistical values such as percentile values or means of sensory-scale values did (e.g. Namba et al., 1972; Namba et al., 1978; Kuwano et al., 1978; Namba and Kuwano, 1982; Kuwano and Namba, 1985).

Although the L_{Aeq} correlates highly with the loudness of many sounds, it does not consider the effects of duration on loudness. It always averages across the actual duration of the sound. However, there is a critical duration of loudness where the effect of the integration of sound energy changes (Scharf, 1978; Namba et al., 2008). Beyond this critical duration, loudness does not depend on duration. When the duration of a sound is shorter than the critical duration, the loudness increases with duration and correlates well with the sound exposure level (L_{AE}), which has a fixed integration time of 1 second. Nonetheless loudness corresponds to the mean energy level in a first approximation.

When the duration of a sound is beyond the frame of sensation and perception, loudness judgments may contain memory effects. There is no evidence that the mean energy level corresponds to the loudness that is stored in memory. However, as far as the signal-response relationship between a sound and its overall loudness is concerned, a high correlation is maintained between L_{Aeq} and loudness even when durations are longer than 10 minutes (Namba and Kuwano, 1980; Namba et al., 1997). Social surveys have also found high correlation between L_{Aeq} (or also day-night average sound level L_{dn}) and the residents' responses to noise, considering long intervals like one year (e.g. Kaku et al., 2007). This is evidence that we can easily judge the overall impression of sounds with a fairly long duration. We may well compose the overall impression by recalling and editing past experience (Kuwano et al., 2003).

More sophisticated loudness models have overcome some disadvantages of A-weighted sound pressure levels. For example, A-weighted levels are a poor estimator for the loudness of sounds that include prominent pure tone components in wide-band noise and/or strong low frequency components (Zwicker and Fastl, 1990; Kuwano et al., 1989). Loudness models for

stationary sounds were first internationally standardized in ISO 532:1975. Kuwano et al. (1978) proposed new metrics that combined the merits of both ISO 532 and ISO 1996, i.e. transforming a spectrum to loudness according to ISO 532 and considering the effect of temporal variations on loudness according to ISO 1996. They named these metrics LL_E and LL_P, representing loudness levels with the unit phon.

For LL_E, third-octave band levels are measured every 100 milliseconds. The energy means across time are taken for each third-octave band. This overall third-octave spectrum is taken to calculate the overall loudness level. For LL_P, first loudness levels are calculated every 100 milliseconds. Second, these loudness levels are converted to an intensity-like quantity as if they represented values in decibels. Finally, the intensity values are averaged over time and converted back to phon, which yields the overall loudness level. Because of the computational limitations in 1978, Kuwano et al. used Stevens's method (ISO 532:1975 Part A) and calculated the loudness level each 100 milliseconds. It should be noted that ISO 1996-1:2016 mentions a "loudness-based method" in its informative appendix. This method is the same as LL_P. However, no practical examples are introduced in that appendix.

Both technology and loudness models have improved since. One of them is the recently published ISO 532-1:2017. It is very similar to its national predecessor, DIN 45631/A1 (2010), and based on models developed by Zwicker (1977) and Chalupper and Fastl (2002). DIN 45631/A1 was used in the present study because it was already available when the experiments were conducted. Kuwano et al. (2011, 2013) and Namba et al. (2011-a,b) already applied LL_P and LL_E to the Zwicker method. They found that LL_E and LL_P show good agreement with each other. However, ISO 532-1 proposes the N_5 as the measure for overall loudness. That is the loudness level that is exceeded in 5 percent of the time.

Both N_5 and LL_P are of statistical origin, and they agree with each other and experimental data (e.g. Kuwano and Namba, 1985, Fastl, 1991) in the fact that the overall loudness is higher than the arithmetic mean of loudness over time.

However, the concepts of LL_P and N_5 are entirely different. N_5 is an ordinal value, i.e. it does not change if the lower 95 percent or the upper 5 percent of loudness values change, as long as the changes do not affect that percentile. In contrast, LL_P is based on the entire distribution of loudness values and is a kind of weighted mean, giving higher weight to higher loudness levels. However, N_5 and LL_P produce rather similar outputs close to the maximum for many measurements. The effect of temporal masking is similar to smoothing the loudness-time function and leads to smaller differences between N_5 and LL_P. In the measurement of N_5, a non-linear circuit is included to implement time constants of hearing (DIN 45631/A1, 2010). Typical real-world sounds are estimated somewhat louder by N_5 than by LL_P, but the correlation between the two is high. For this reason it was not easy to find a suitable set of exceptions, which are needed to test the differences between the two concepts. Ideally, such an experiment consists of a set of orthogonal conditions over a wide dynamic range.

Furthermore, it was important to find a good choice for the stimulus duration. We chose 10 seconds as it is long enough to include several changes in loudness over time but still short enough to be repeated in many trials. However, the N_5 implies that 5 percent of a 10-second stimulus is represented by only 500 milliseconds. If the softer 95 percent are absolute silence, LL_P is reduced by 13 phon while N_5 remains the same. LL_P could be infinitely higher than N_5 if there are short periods of very high loudness that occur less than 5 percent of the time.

Apart from experiments employing artificial time-varying stimuli like amplitude-modulated sounds (e.g. Moore et al., 1999), there are numerous studies which have assessed the overall loudness of real-world sounds, for example speech (Fastl, 1976; Moore et al., 2003; Rennies et al., 2013), road traffic noise (Kuwano and Namba, 1985; Fastl, 1991; Hellbrück,

2000; Namba et al., 2008), aircraft noise (Namba and Kuwano, 1980; Namba et al., 1993; Kuwano and Namba, 1996), technical sounds (Rennies et al., 2015), music (Laumann et al., 2007) or a variety of other types of sounds (Kuwano et al., 1978; Namba and Kuwano, 1982; Kuwano et al., 1988; Skovenborg et al., 2004), including some very specific noise sources like a tennis court (Stemplinger, 1999). However, to our knowledge none of these studies compared a percentile to an energy mean based on a loudness model, and only a few of them used instantaneous judgment in addition to overall judgment.

In the present experiments we used a set of real-world sounds for which N_5 and LL_P give very different predictions by being orthogonal over a dynamic range of 30 phon. This allows a clear comparison about which of the two metrics performs better.

II. METHOD

Three experiments using the same sounds but slightly different approaches in methodology were conducted. In Experiment 1, participants were asked for the overall judgment (OJ) of loudness only. Experiment 2 used both OJ and instantaneous judgment (IJ) to provide further data for OJ, to relate OJ to IJ, and to test whether the additional IJ influenced OJ. Experiment 3 used IJ only to allow a focus on IJ with less interruptions, i.e. all stimuli were concatenated to a stream. By using the same participant group as in Experiment 2, the results of Experiment 3 still could be compared to OJs, and it could be tested if IJ was influenced by concurrently thinking about OJ in Experiment 2.

A. Participants

Eleven participants, six of them females and five males, aged 21 to 30 years (median 23) were tested in Experiment 1. Twenty-one different participants were tested in Experiment

2. They were 17 females and four males, aged 18 to 25 years (median 21). Twenty participants, 16 of them females and four males, aged 18 to 25 years (median 21) were tested in Experiment 3. The participants of Experiments 1 and 2 were two distinct groups to yield a bigger total sample size for the overall judgment. The participants of Experiments 2 and 3 were the same, except one female who had moved to another city in the meantime. Because the main focus of this study lay on OJ, it seemed more desirable to allow a comparison of the IJs to OJs made previously by the same participants rather than to increase the sample size for IJ. All participants had thresholds equal or better than 20 dB HL, measured in octave steps from 125 Hz to 8 kHz.

B. Apparatus

The experiments took place in a double-walled sound-proof booth. Stimuli were presented via an audio interface (RME Hammerfall DSP Multiface II), an amplifier (TDT HB7) and headphones (Beyerdynamics DT-48.00). Free-field equalization was implemented in Matlab, simulating the passive network introduced by Zwicker and Maiwald (1963).

C. Stimuli

All stimuli were recordings of real sounds, made using a binaural headset (Head acoustics BHS I) in Tokyo, Japan. They were converted to diotic sounds using the transfer function provided by the manufacturer. They were chosen from several hours of recording made for the present study in various environments, and in a way to be roughly equidistant on the phon scale.

The sounds either consisted of a prominent portion and a background noise that both were present in the original recording, or they were constructed by adding two recordings. Two sounds fell in the first category, and six in the latter (see Table I for a description of each sound). For these six sounds, the prominent portions were hammer blows, recorded in a sound-proof

room, and the added background noise was environmental noise. The added background noise was set to a rather soft loudness level of 58 phon, and resembled environmental noise through a window or wall.

Table I: Stimuli

No	Description	Added background noise
1	Hammer, 1 blow	Construction noise
2	Hammer, 1 blow destroying acrylic glass	Train noise
3	Wooden hammer, 2 blows	Shopping arcade
4	Hammer, 2 blows	Traffic noise
5	Hammer, 23 soft blows	Helicopter
6	Hammer, 15 blows on leather on wood	Shopping arcade
7	Inside local train, announcement	
8	Train horn at railway platform	

The largest difference between N_5 and LL_P was obtained for a single hammer blow with construction noise in the background. For this sound, LL_P predicted an overall loudness level 21 or 23 phon higher than for N_5, depending on the absolute level. The largest difference in the other direction was obtained for a recording made at a railway platform with a train horn as the prominent part. For this sound, the N_5 was 8 to 9 phon higher than the LL_P.

The eight sounds were presented at two levels. In the first set of the eight sounds, the LL_P was kept constant at 85 phon while the N_5 varied between 64 and 93 phon. In the second set, the N_5 was set constant to 68 phon while the LL_P varied between 59 and 91 phon. Thus, 16 stimuli were obtained with subsets for which either the N_5 or the LL_P was constant and the other metric varied in a wide dynamic range between approximately 60 and 90 phon. In either case the loudness level of the added background noise was 58 phon before it was added to the prominent portion. The duration of each stimulus was 10 seconds with a Gaussian rise and fall time of 5 milliseconds.

The 16 stimuli were used in their original forms in Experiments 1 and 2. In Experiment 3, they were concatenated in random order to yield streams with a duration of 160 seconds. Furthermore, 5-second segments of background noise were appended at the beginning and the end of each stream. Three such streams were generated.

D. Procedure

In all experiments loudness was judged using cross-modality matching with line length. The participants were asked to adjust the length of a line to match his or her impression of loudness. This was done continuously for instantaneous judgment made while the sound was being played or once after the end of the sound for the judgment of overall loudness, depending on the experiment. The line was presented horizontally on a screen, and its length could be modified by moving the mouse. Its minimum length was 0 pixels and its maximum was 1260 pixels, leaving a margin of 10 pixels on each side of the monitor. It was set to a length of 10 pixels before each trial, i.e. a short length at which a line was still recognizable.

In Experiments 1 and 2 the 10-second long sounds were each presented six times in a random order but with blocking. These six blocks of 16 trials each were connected seamlessly so that the participant did not notice when a new block started. In Experiment 1, the participants were asked to focus on the sound while it was played and to judge its overall loudness afterwards. In Experiment 2, the participants did both IJ and OJ. In Experiment 3, the participants did IJ while a stream was played and judged the overall loudness of the entire stream after it had finished. From the viewpoint of a single 10-second long sound, the streams ensured a focus on IJ. Two streams were used in random order in Experiment 3.

The participants received one round of practice at the beginning of Experiments 1 and 2; they listened to and judged each of the 16 sounds once. They practiced with one stream before starting Experiment 3. After practice, they were told that these were all of the sounds

that would be used in the main experiment, so that they could adjust their "calibration" of line length, if necessary.

III. RESULTS

At first, the reaction time was estimated for each participant and each trial (see Kuwano, 1996, for details). These time-lag values were used to time-align the IJ values across participants. For those readers interested in reaction time, the mean was 495 milliseconds in Experiment 3, and the standard deviation between individual means 85 milliseconds. Afterwards, the geometric mean was taken to average across replications and participants as the task resembled a free magnitude estimation without standard. This way of averaging was used both for OJ and a single point of time in IJ.

Figure 1 shows the OJs made by the eleven participants in Experiment 1. Each data point represents 66 trials. The results are shown as a function of N_5 on the left hand side and as a function of LL_P on the right hand side. The abscissa shows the calculated loudness levels in phon, and the ordinate is the logarithm of the line length. The OJ varied only over a small range for the conditions with constant LL_P, depicted by red circles and ranging from 384 to 570 pixels, which is a factor of 1.5. Conditions which should have the same loudness according to N_5 (blue squares) ranged from 147 to 532 pixels, which is a factor of 3.6. The Pearson correlation between LL_P and OJ is $r_{(14)} = .944$, $p < .001$, while N_5 and OJ do not correlate significantly, $r_{(14)} = .464$, $p = .07$. Using the test of Meng et al. (1992), the difference between the correlation coefficients for LL_P and for N_5 is statistically significant, $Z = 3.44$, $p < .001$.

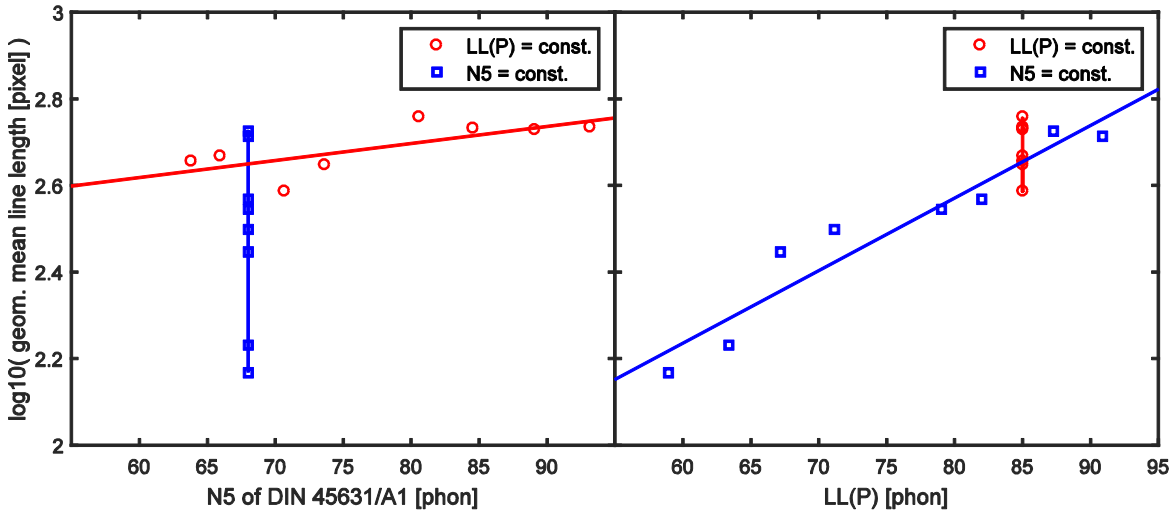

Figure 1. Overall judgments made in Experiment 1 as a function of N_5 on the left hand side and LL_P on the right hand side. Red circles indicate the set of conditions for constant LL_P, blue squares the set for constant N_5. Regression lines are shown for each set of eight stimuli. (Color online)

Figure 2 illustrates the OJs made in Experiment 2 as Figure 1 did for Experiment 1. Because of the higher number of participants, each data point represents 126 trials. The OJs for the conditions of constant LL_P range from 271 to 523 pixels, which is a factor of 1.9, those for constant N_5 from 114 to 373 pixels, which is a factor of 3.3. The Pearson correlation between LL_P and OJ is $r_{(14)} = .901$, $p < .001$. It amounts to $r_{(14)} = .606$, $p < .05$ between N_5 and OJ. The difference between the correlation coefficients is statistically significant, $Z = 2.15$, $p < .05$.

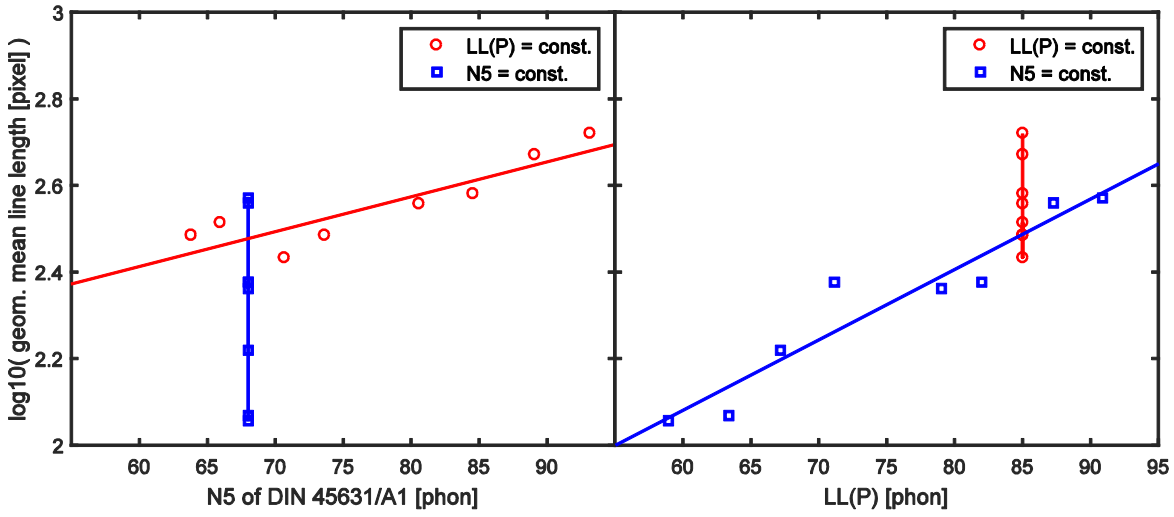

Figure 2. Overall judgments made in Experiment 2. Otherwise as Fig. 1. (Color online)

The two experiments with different participant groups further allow to test whether overall loudness was evaluated independently of doing an additional continuous judgment in Experiment 2 compared to listening to the sounds only in Experiment 1. The correlation between the OJs made in Experiments 1 and 2 is $r_{(14)} = .963$, $p < .001$.

To assess whether the OJ could be predicted simply by the geometric mean of the IJ, the geometric mean across time was calculated for each 10-second stimulus. The outcome is shown in Figure 3 (blue circles). Since 20 of the participants were the same for Experiments 2 and 3, the OJ from Experiment 2 can also be compared with the geometric means of the IJ of the 20 participants for the corresponding 10-second segments in Experiment 3. The results obtained in this way are shown in Figure 3 as red squares. All OJs are greater than the geometric mean of the IJ, i.e. the geometric mean of IJ underestimates overall loudness. A very similar outcome was obtained when taking the arithmetic mean across time (not shown here). The correlation between the geometric mean IJs of Experiments 2 and 3 is $r_{(14)} = .988$, $p < .001$. This indicates that the participants were consistent despite the different method of presentation, either isolated in a single trial or within a stream, and a time gap of several weeks between the two experiments.

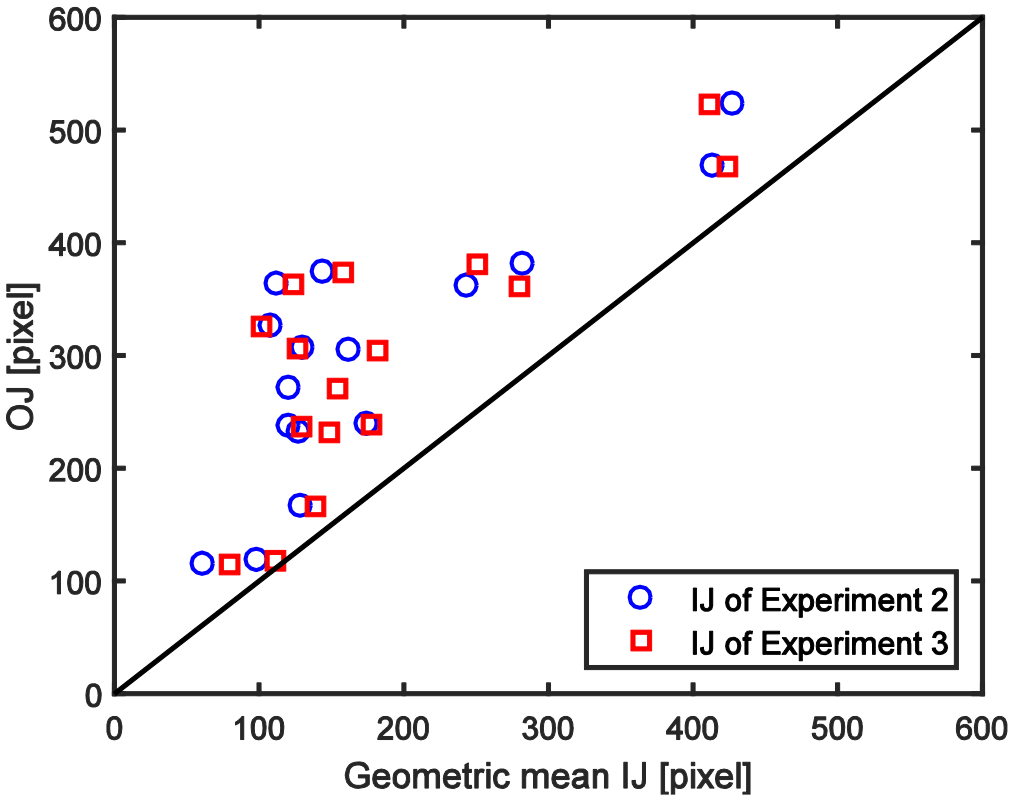

Figure 3. Overall judgment of Experiment 2 as a function of the geometric mean of instantaneous judgment made by the same participants. Blue circles represent IJ made in Experiment 2, red squares IJ made in the streams of Experiment 3. Linear scales are used for more clarity. (Color online)

Figure 4 compares IJs of Experiment 2 with the calculated loudness over time, grouped by loudness level into 1-phon wide bins. For this purpose both IJ and calculated loudness were averaged over segments lasting 100 milliseconds. Loudness levels that occurred in at least 25 segments are represented by circles, and those that occurred in at least 10 to 24 segments are represented by triangles. Error bars indicate one standard deviation into each direction. Loudness levels which occurred in less than 10 segments are indicated by crosses without error bars. A high correlation is obtained between loudness levels greater than 60 phon and their corresponding line length of the IJ, $r_{(34)} = .957$, $p < .001$. For the lower loudness levels close to

that of the added background noise, however, line length is rather constant, except perhaps when the loudness level was below 50 phon. For Experiment 3 the same analysis yields almost identical results (not shown here). The correlation between its IJ and calculated loudness level is $r_{(34)} = .945$, $p < .001$.

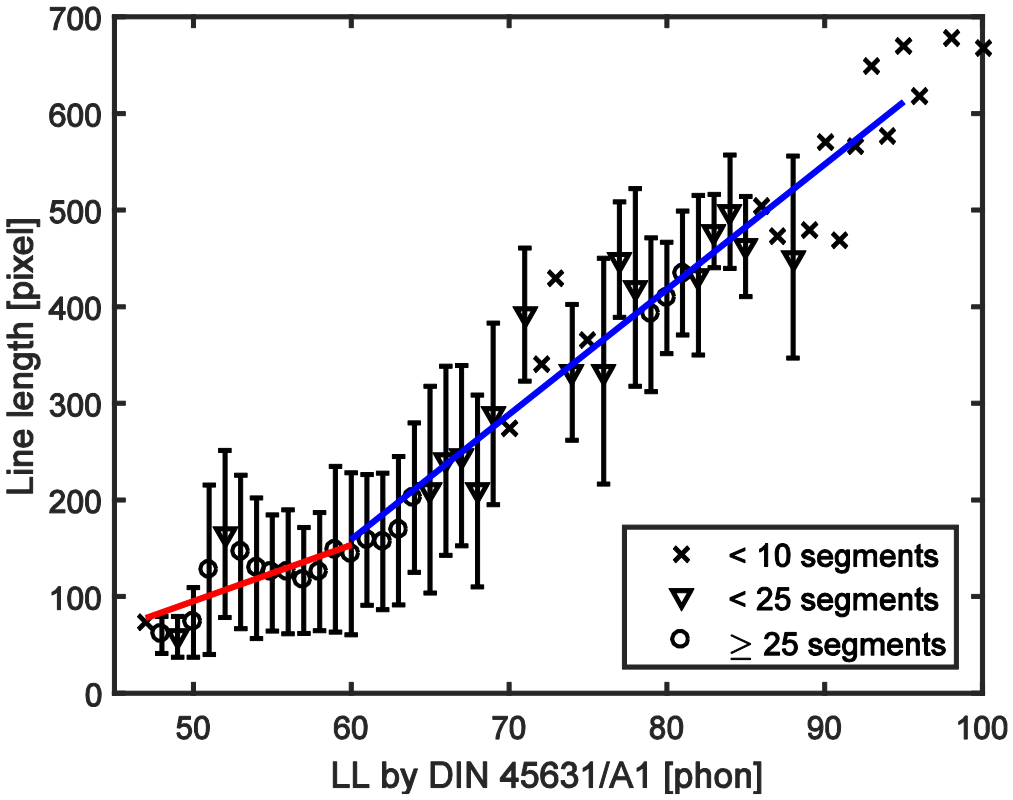

Figure 4. Instantaneous judgment as a function of calculated loudness level. Circles represent loudness levels that occur in at least 25 100-millisecond long segments, triangles those appearing in at least 10 segments and crosses those in less. Error bars indicate one standard deviation in each direction. The solid red line depicts a regression for loudness levels smaller than 60 phon, the solid blue line for loudness levels between 60 and 95 phon. (Color online)

IV. DISCUSSION

In both Experiment 1 and Experiment 2, LL_P showed considerably higher correlations

with loudness judgments than N_5 did. This is not inconsistent with studies that have shown accurate predictions for a high percentile like N_5, as these studies mostly used sounds with comparatively gradual changes in loudness. For example, Fastl (1991) used 17-minute long recordings of road traffic noise, which typically does not contain impulsive portions as a vehicle needs several seconds to pass by. Fastl found that overall loudness was well predicted by N_4, the loudness exceeded in four percent of the time. It should be noted that other studies have found high correlations between OJs of road traffic noise and the energy mean as well (e.g. Kuwano and Namba, 1985). These findings are similar for other types of environmental noise (e.g. Namba and Kuwano, 1982), though most of these earlier studies took the energy mean based on sound pressure level rather than loudness level.

When impulsive sounds were included in the set of stimuli, N_5 was not a good predictor of overall loudness. Rennies et al. (2015) found a correlation of $r = .55$ between N_5 and OJ for technical sounds such as machinery or engine noise. In that study the sounds had a duration of about 2 seconds and several of them were impulsive. This correlation coefficient is in the range of those obtained in Experiment 1 ($r = .46$) and Experiment 2 ($r = .61$). Stemplinger (1999) found that N_1, the loudness exceeded in only 1 percent of the time, was a much better predictor than N_5 for noise from a tennis court. This means that the percentile best describing overall loudness could change depending on the impulsiveness of the sound. Unfortunately, these two studies did not investigate LL_P. Taking the results of the present study into consideration too, it could be concluded that the N_5 may be a good descriptor for many sounds as long as they do not have impulsive portions. LL_P shows high correlation with the subjective evaluations of overall loudness when both impulsive sounds and more steady environmental sounds are included in the set of stimuli, i.e. for any kind of sound. ISO 532-1 is fairly flexible regarding a metric to determine the overall loudness. Although it recommends the N_5, it also allows other percentiles. Furthermore, it introduces the use of the energy mean of loudness levels in a note,

and our previous conference paper is referred to as an example (Schlittenlacher et al., 2014).

The additional consideration of IJ allows one to estimate which portions of the sound the participants actually took to form their OJ. Figure 3 shows that the OJ was systematically higher than average IJ, indicating that the louder portions of the stimuli dominated OJ. Figure 4 bolsters this assumption. It may be interpreted that the participants did not differentiate loudness in the portions of the background noise, as the IJs showed approximately the same value for loudness levels between 50 and 60 phon. By contrast, IJ grows with loudness level between 60 and 100 phon, indicating that the participants paid more attention to these portions. The energy mean seems to be a good statistical descriptor for this behavior. For example, if segment A has a loudness level being 20 phon higher than that of segment B, segment A has 100 times the energy of segment B. Thus, the energy mean is dominated by segment A and almost independent of the exact value of segment B.

The energy mean treats all temporal segments equally, meaning it does not consider a primacy effect that has often been found for stimulus durations of 1 to 2 seconds (e.g. Pedersen and Ellermeier, 2008). In their study participants gave a considerably higher weight to the first 100 milliseconds when judging the overall loudness of a 1-second long white noise whose level changed each 100 milliseconds. Also Namba et al. (1976) reported a primacy effect using a non-steady state of 700 milliseconds. These researchers found that the sound was judged louder when the intensity increment of 100 milliseconds was located at the beginning of the sound than when the increment was located in the middle. This effect was statistically significant though the amount of the effect is about 1 dB. However, Buus (1999) did not find such a primacy effect for a sequence of six pulses with a total duration of 300 milliseconds, with small but statistically significantly higher weights for the middle pulses. Oberfeld and Plank (2011) and Ponsot et al. (2013) found that the level of a segment has more impact on its weight than the temporal position has. For example, when a sound increases in level, the last segment is

weighted most highly. The finding that higher weights are assigned to higher-level segments is consistent with the present results and with the concept of an energy mean.

V. CONCLUSIONS

(1) The aim of this study was the comparison between N_5 and LL_P as a metric of non-stationary sounds. LL_P provides a good measure of overall loudness. It is highly correlated with judgments of overall loudness for a wide range of stimuli covering a wide range of levels. LL_P is based on ISO 1996 for averaging temporal level fluctuation and on ISO 532-1 (Zwicker method) for frequency weighting. LL_P is probably superior because it combines the merits of both ISO standards.

(2) The concept of an ordinal value such as N_5 is not universally acceptable. Difficulties may appear especially when a sound's loudness distribution contains a wide range of loudness levels above the percentile, or when the distribution's slope is very steep at the percentile.

Acknowledgments

We thank the German Academic Exchange Service (DAAD) for supporting this study. The first author is supported by the Engineering and Physical Sciences Research Council (UK, grant number RG78536).

References

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

Chalupper, J., and Fastl, H. (**2002**). "Dynamic loudness model (DLM) for normal and hearing-impaired listeners," Acta Acust. united Ac. **88**, 378-386.

DIN 45631/A1 (**2010**). Berechnung des Lautstärkepegels und der Lautheit aus dem Geräuschspektrum – Verfahren nach E. Zwicker – Änderung 1: Berechnung der Lautheit zeitvarianter Geräusche (Calculation of loudness level and loudness from the sound spectrum – Zwicker method – Amendment 1: Calculation of the loudness of time-variant sounds) (Deutsches Institut für Normung e.V., Beuth Verlag, Berlin).

Fastl, H. (**1991**). "Evaluation and measurement of perceived average loudness," in *Contributions to psychological acoustics – Results of the fifth Oldenburg Symposium on psychological acoustics* (Bibliotheks- und Informationssystem der Universität Oldenburg, Oldenburg), pp. 205-216.

Fastl, H. (**1976**). "Schallpegel und Lautstärke von Sprache (Sound pressure level and loudness of speech)," Acustica **35**, 341-345.

Hellbrück, J., (**2000**). "Memory effects in loudness scaling of road traffic noise – How overall loudness of short-term and long-term sounds depends on memory," J. Acoust. Soc. Jpn. (E) **21**, 329-332.

ISO 1996-1:2016 (**2016**). Acoustics – Description, measurement and assessment of environmental noise – Part 1: Basic quantities and assessment procedures. (International Organization for Standardization, Geneva).

ISO 532:1975 (**1975**), Acoustics – Methods for calculating loudness level. (International Organization for Standardization, Geneva).

ISO 532-1:2017 (**2017**). Acoustics – Methods for calculating loudness - Part 1: Zwicker method. (International Organization for Standardization, Geneva).

Kaku, J., Yokota, T., Namba, S., Ogata, S. & Yamada, I. (**2007**). "Availability of a newly developed social survey method on noise using Internet and the geographic information system (GIS)," INTER-NOISE and NOISE-CON Congress and Conference Proceedings 2007, 2793-2799.

Kuwano, S., Kato, T. and Namba, S. (**1978**). "On the loudness of level-fluctuating complex sound in relation to Leq, La, LL and PNL," Transactions of the Technical Committee on Hearing, Acoustical Society of Japan, H48-1, 6-10.

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

Kuwano, S., Namba, S., and Fastl, H. (**1988**). "On the judgment of loudness, noisiness and annoyance with actual and artificial sounds," J. Sound Vib. **127**, 457-465.

Kuwano, S., Namba, S., and Miura, H. (**1989**). "Advantages and disadvantages of A-weighted sound pressure level in relation to subjective impression of environment noises," Noise Control Eng. J. **33**, 107-115.

Kuwano, S. (**1996**). "Continuous judgment of temporally varying sounds," in *Recent trends in hearing research – Festschrift for Seiichiro Namba, edited by H. Fastl, S. Kuwano and A. Schick* (Bibliotheks- und Informationssystem der Universität Oldenburg, Oldenburg), Chap. 8, pp. 193-214.

Kuwano, S., and Namba, S. (**1996**). "Evaluation of aircraft noise: Effects of number of flyovers," Environ. Int. **22**, 131-144.

Kuwano, S., Namba, S., Kato, T. and Hellbrück, J. (**2003**). "Memory of the loudness of sounds in relation to overall impression," Acoust. Sci. Tech., 24 (4), 194-196.

Kuwano, S., Namba, S. and Kato, T. (**2011**). "Calculation of loudness level of time-varying sounds," Proceedings of Inter-noise 2011, pp. 353-358.

Kuwano, S., Hato, T., Kato, T. and Namba, S. (**2013**). "Evaluation of the loudness of stationary and non-stationary complex sounds," J. Acoust. Soc. Am. **133**, 3597-3597.

438 Laumann, K., Fastl, H., Kuwano, S., Namba, S. (**2007**). "Overall loudness versus average of
439 instantaneous loudness for excerpts of music: Effects of musical style," Fortschritte der
440 Akustik – DAGA 2007, 865-866.
441 Moore, B. C. J., Vickers, D. A., Baer, T., and Launer, S. (**1999**). "Factors affecting the
442 loudness of modulated sounds," J. Acoust. Soc. Am. **105**, 2757-2772.
443 Moore, B. C. J., Glasberg, B. R., and Stone, M. A. (**2003**). "Why are commercials so loud?
444 Perception and modeling of the loudness of amplitude-compressed speech," J. Audio
445 Eng. Soc. **51**, 1123-1132.
446 Meng, X. L., Rosenthal, R., Rubin, D. B. (**1992**). "Comparing Correlated Correlation
447 Coefficients," Psychol. Bull, **111**, 172-175.
448 Namba, S., Nakamura, T. & Kuwano, S. (**1972**). "The relation between the loudness and the
449 mean of energy of level-fluctuating noises," Japanese Journal of Psychology, **43**, 251-
450 260.
451 Namba, S., Kuwano, S. and Kato, T. (**1976**). "The loudness of sound with intensity increment,"
452 Jpn. Psychol. Res. **18**, 63-72.
453 Namba, S., Kuwano, S. and Kato, T. (**1978**). "An investigation of Leq and Lα in relation to
454 loudness," J. Acoust. Soc. Jpn. **34**, 301-307.
455 Namba, S., and Kuwano, S. (**1980**). "The relation between overall noisiness and instantaneous
456 judgment of noise and the effect of background noise level on noisiness," J. Acoust. Soc.
457 Jpn. (E) **1**, 99-106.
458 Namba, S. and Kuwano, S. (**1982**). "Psychological study on Leq as a measure of loudness of
459 various kinds of noises," J. Acoust. Soc. Jpn., **38**, 774-785.
460 Namba, S., Kuwano, S., and Koyasu, M. (**1993**). "Measurement of temporal stream of hearing
461 by continuous judgments – In the case of the evaluation of helicopter noise," J. Acoust.
462 Soc. Jpn. (E) **14**, 431-352.

463 Namba, S., Kuwano, S., Kinoshita, A. and Hayakawa, Y. (**1997**). "Psychological evaluation of
464 noise in passenger cars - The effect of visual monitoring and the measurement of
465 habituation," J. Sound Vib. **205**, 427-434.
466 Namba, S., Kuwano, S., and Fastl, H. (**2008**). "Loudness of non-steady-state sounds," Jpn.
467 Psychol. Res. **50**, 154 – 166.
468 Namba, S., Kato, T., and Kuwano, S. (**2011-a**). "Evaluation of loudness level of time-varying
469 sounds," Proceedings of inter-noise 2011, 1584-1590.
470 Namba, S., Kuwano, S. and Kato, T. (**2011-b**). "Loudness of temporally varying complex
471 sounds. J. Music Perception & Cognition," **17**, 19-39.
472 Oberfeld, D., and Plank, T. (**2011**). "The temporal weighting of loudness: effects of level
473 profile," Attent. Percept. Psychophys. **73**, 189-208.
474 Pedersen, B., and Ellermeier, W. (**2008**). "Temporal weights in the level-discrimination of time-
475 varying sounds," J. Acoust. Soc. Am. **123**, 963-972.
476 Ponsot, E., Susini, P., Saint Pierre, G., and Meunier, S. (**2013**). "Temporal loudness weights for
477 sounds with increasing and decreasing intensity profiles," J. Acoust. Soc. Am. **134**,
478 EL321-EL326.
479 Rennies, J., Holube, I., and Verhey, J. L. (**2013**). "Loudness of speech and speech-like signals,"
480 Acta Acust. united Ac. **99**, 268-282.
481 Rennies, J., Wächtler, M., Hots, J., and Verhey, J. (**2015**). "Spectro-temporal characteristics
482 affecting the loudness of technical sounds: data and model predictions," Acta Acust.
483 united Ac. **101**, 1145-1156.
484 Scharf, B. (**1978**), "Loudness," in *Handbook of perception (vol. IV), Hearing*, Carterette, E. C.
485 and Friedman, M. P. (Eds.), Academic Press, N.Y., pp 187-242.
486 Schlittenlacher, J., Hashimoto, T., Kuwano, S. and Namba, S. (**2014**). "Overall loudness of
487 short time-varying sounds," Proceedings of Internoise 2014, pp. 2666-2669.

488 Skovenborg, E. Quesnel, R., and Nielsen, S. H. (**2004**). "Loudness assessment of music and
489 speech," Audio Engineering Society Convention 116 (Berlin), pp. 1-25.
490 Stemplinger, I. (**1999**). *Beurteilung, Messung und Prognose der globalen Lautheit von*
491 *Geräuschimmissionen (Judgment, measurement and prediction of the overall loudness of*
492 *noise immissions)*, PhD Dissertation, Technische Universität München.
493 Zwicker, E., and Maiwald, D. (**1963**). "Über das Freifeldübertragungsmaß des Kopfhörers DT
494 48 (On the free-field response of the earphone DT 48)," Acustica **13**, 181-182.
495 Zwicker, E., (**1977**). "Procedure for calculating loudness of temporally varying sounds," J.
496 Acoust. Soc. Am. **62**, 675-682.
497 Zwicker, E. and Fastl, H. (**1990**). *Psychoacoustics – Facts and Models.* (Berlin, Springer), Chap.
498 8, pp. 203-238.