7 - 12 SEPTEMBER 2014, KRAKÓW
FORUM ACUSTICUM

Instantaneous and overall loudness of music

Josef Schlittenlacher, Alisa Samel, Angelo Schleussner, Katharina Rost, Özlem Çelebi, Wolfgang Ellermeier
Applied Cognitive Psychology Unit, Technische Universität Darmstadt, Germany

Summary
Although loudness is a perception which changes over time as a sound changes over time, listeners often assign one loudness value for a sound event of a longer duration such as a song, movie or a noise exposed environment. There are different approaches for modeling this overall loudness, i.e. the LL(P) as the energetic mean of instantaneous loudness and the N5 as the loudness which is exceeded in five percent of the time. In the present study, listeners were asked to judge the instantaneous loudness of music by continuous judgment using line length. After each of the 14 songs, seven of classical music and seven of rock music, they further judged the overall loudness by line length. Applying different tests, there are slight advantages for the energetic mean, however, the N5 also proves to be a good estimation of overall loudness. The results further confirm earlier studies which state that the arithmetic mean of instantaneous loudness is systematically smaller than the overall loudness.

PACS no. 43.66.+y, 43.75.+a

1. Introduction

Loudness is an important aspect of music, whether the composer or conductor of a classical piece use it as a stylistic element to cover a wide dynamic range or whether listeners control it themselves in order to reach a comfortable level or to mask their environment. Mostly, users adjust loudness subjectively while the music is played. However, there are also cases in which an objective measure for the loudness of music is desired, for example to categorize songs on the same medium or to adjust them automatically to have the same overall loudness. In these cases it is necessary not only to have a measure for the instantaneous loudness, which can be measured continuously as a function of time, but also to have a function to transform it to overall loudness, which is a single value that is representative for the whole song.

Many studies have been conducted adressing overall loudness, a large part of them evaluates unwanted sounds like road traffic noise, e.g. [1] and [2], aircraft noise [3] or railway noise [4], but some also address speech [5] or music [6]. Although the latter used musical stimuli, it used rather short segments of ten to 15 seconds instead of whole songs and it did not track instantaneous loudness. A different study, by contrast, used continuous judgment for music [7], however, it did not focus on loudness. The present paper will focus on the continuously judged loudness of music, with an additional emphasis on the overall loudness of songs.

There are several competing models which claim to be able to predict instantaneous and overall loudness of sounds in general. DIN 45631/A1 [8] is the newest revision of the Zwicker model which can also calculate the loudness of time-varying sounds. It proposes the N5 as a measure for overall loudness. That is the loudness which is exceeded in five percent of the time. The LL(P) [9] is also based on the Zwicker model, however, it proposes the energy mean of loudness levels over time as the measure for overall loudness. Of course, it could also be applied to the Moore and Glasberg model [10]. The Glasberg and Moore model for time-varying sounds itself [11] will not be evaluated in the present paper as its concept is somewhat different. It does not suggest a single value for overall loudness, but rather several loudness-time functions covering instant impressions and long-term memory effects.

Thus the present paper will investigate whether the N5 or LL(P) is more appropriate to assess the loudness of music. For this purpose, participants listened to classical music, having a wide dynamic range within a single piece, and rock songs. The estimated overall loudness level was also varied in a wide range. The participants judged instantaneous loudness while they listened to a song and its overall loudness after a song finished. In either case the length of a line was manipulated. This procedure allows various analyses: First, the overall judgment can be correlated to the model predictions. Second, the concepts of N5 and LL(P) can

also be applied to the instantaneous judgments and afterwards be compared to overall loudness, not only by means of a correlation but also in a comparison of the absolute values (see also [1]). In order not to confuse the purely calculated model predictions and their concepts being applied to the instantaneous judgment, the model predictions will be called N5 and LL(P) throughout the paper and their concepts applied to the instantaneous judgment will be called 95th percentile and energetic mean line length, respectively.

2. Method

2.1. Participants

20 listeners, eight females and twelve males, aged 18 to 50 years (median 23 years) participated in the experiment. All of them passed a hearing test meaning their threshold in quiet was not worse than 20 dB HL for any frequency between 125 Hz and 8 kHz, measured in octave steps and for each ear. They participated voluntarily without receiving any credit.

2.2. Apparatus

During the experiments, the participants sat in a double-walled sound-proof chamber manufactured by the Industrial Acoustics Company. The sounds were D/A-converted by an external audio interface (RME Hammerfall DSP Multiface II) and presented via headphones (Sennheiser HDA 200) which were connected directly to the audio interface.

2.3. Stimuli

All participants listened to 14 songs or pieces of music with durations between 142 and 251 seconds (see Table I). Seven of them can be assigned to the rock genre, seven to the genre of classical music. They were paired and adjusted in level so that one rock song and one classical piece had the same N5 loudness. These loudness levels were chosen from 70 to 94 phon in steps of 4 phon. For the present set of stimuli, LL(P) is smaller than N5 in general and LL(P) is not equal within a pair. It is systematically higher for the rock songs as they are less dynamic than the classical pieces of music.

All stimuli were stored in the wavefile format with a sampling rate of 44.1 kHz and at a resolution of 16 bit.

2.4. Procedure

The participants judged both instantaneous and overall loudness of each stimulus. For instantaneous judgment (IJ), continuous judgment by line length was employed (see [12] and [13]). This means the participant adjusted a line on the monitor by moving the mouse so that the line length corresponded to his or her perception of loudness at any time during he or she heard the sound. The maximum line length was 1260 pixels. After a song finished and an additional interval of 3 s passed by, the participant was asked to judge the song's overall loudness. Also for this purpose, he or she adjusted the length of a line on the monitor.

After the overall judgment (OJ) of each song had been made, the participant could take a break. The 14 songs were presented in random order. Before the actual experiment, the participants completed a training. It consisted of three segments of popular music, IJ and OJ were done in the same way as in the experiment. The segments had overall loudness levels of about 70, 80 and 90 phon, and a duration of 20 s each.

2.5. Analysis

In order to analyze the IJ, it is important to consider the individual reaction time of each participant or, more precisely, the individual delay in tracking perceived loudness. It can be obtained by correlating the IJ with a physical measure [14]. For the present analyses, IJ was cross-correlated to the instantaneous loudness physically measured by DIN 45631/A1, separately for each participant and each song. The time lag showing the highest correlation was taken as the reaction time, allowing a minimum of 0 s and a maximum of 3 s.

3. Results

First of all, it is interesting to see how the models predict the overall loudness of the songs. Figure 1 illustrates the geometric mean line length for each song as a function of N5 loudness. The geometric standard deviation across subjects of the free magnitude estimation by line length is shown, too. Not surprisingly, line length increases with loudness level (LL). It can also be seen that there seems to be no difference in line length between the rock songs and pieces of classical music. These effects are confirmed by a 7 x 2, LL x genre, within-subjects analysis of variance. The main effect of loudness level is statistically significant, $F(6,114) = 47.1$, $p < .001$, that of genre is not significant, $F(1,19) = 0.022$, $p = .884$. The interaction between LL and genre is not significant, $F(6,114) = 0.904$, $p = .494$.

One among several possibilities to compare the models of N5 and LL(P) is to take the correlations between their predicted overall loudness and the geometric mean line length of the overall judgment (OJ) as shown in Figure 1. To consider the power law of the loudness function, the Pearson correlation between loudness level in phon and the logarithm of the line length in pixel is taken. N5 and OJ correlate with $r(12) = .979$, LL(P) and OJ with $r(12) = .957$.

In order to investigate the concepts of LL(P) and N5, the energetic mean and 95th percentile of the instantaneous judgment (IJ) can be compared to the

Table I. Songs and parts of classical pieces used as stimuli

No.	Genre	Interpret or componist	Title	N5 [phon]	LL(P) [phon]	Duration [s]
1	Classic	Chopin	Nocturn	70.0	66.4	237
2	Classic	Beethoven	Fidelio: Prisoners' Chorus	74.0	69.4	172
3	Classic	Tchaikovsky	Swan lake	78.0	71.2	152
4	Classic	Mozart	Requiem: Hostias	82.0	77.3	209
5	Classic	Mozart	The marriage of Figaro	86.0	82.0	251
6	Classic	Orff	Carmina Burana: Estuans Interius	90.0	86.6	142
7	Classic	Orff	Carmina Burana: O Fortuna	94.0	89.2	166
8	Rock	Panic at the Disco	Nine in the Afternoon	70.0	68.5	155
9	Rock	The Rolling Stones	Start me up	74.0	72.1	209
10	Rock	Escape the Fate	Ashley	78.0	76.2	206
11	Rock	Die Toten Hosen	Der Moment	82.0	80.5	181
12	Rock	Danko Jones	Sticky Situation	86.0	84.8	156
13	Rock	Danko Jones	Baby Hates Me	90.0	88.5	208
14	Rock	30 Seconds to Mars	Closer to the Edge	94.0	92.0	271

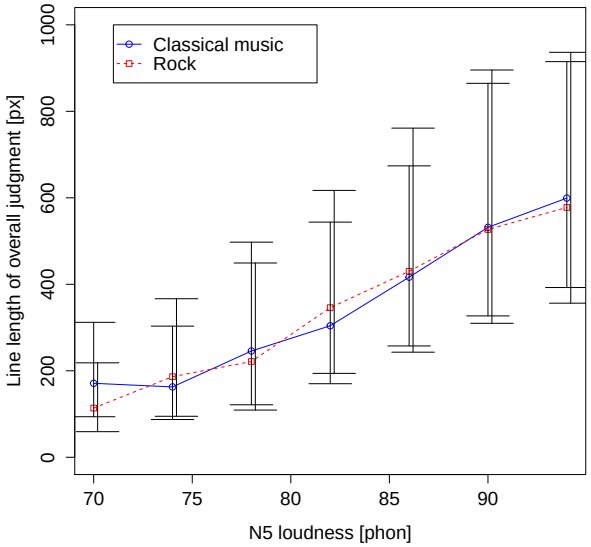

Figure 1. Line length of the overall judgment for the 14 sounds. Geometric means and geometric standard deviations across subjects are depicted.

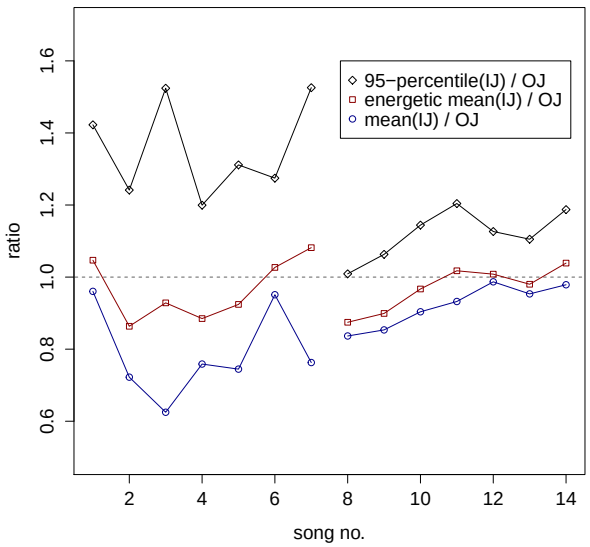

Figure 2. Ratio between different predictions based on instantaneous judgment and the overall judgment for the 14 songs studied. No. 1-7 represent pieces of classical music, no. 8-14 rock songs. Black diamonds depict the 95th percentile, red squares the energetic mean and blue circles the arithmetic mean.

OJ. For doing so, the reaction time has to be considered first. Afterwards, the prediction based on IJ is calculated for each participant and each song. Finally, the geometric mean across participants is taken. The 95th percentile of IJ and the OJ correlate with $r(12) = .978$, the energetic mean of IJ and the OJ with $r(12) = .995$ and the mean of IJ and the OJ with $r(12) = .973$. However, it is not only interesting to see how well the concepts and actual OJ correlate, but also how good they fit in their absolute values. When dividing a prediction based on IJ by the OJ, the ratio should be 1 if the prediction is perfect. Figure 2 shows these ratios to the OJ for the 95th percentile, the energetic mean and the mean of IJ.

It can be seen that all concepts show rather similar ratios for the rock songs (no. 8-14), all being in a range between 0.8 and 1.2. For the classical pieces of music (no. 1-7), only the energetic mean is close to the optimum of 1. The arithmetic mean underestimates the overall judgment, with some predictions being only slightly more than the half of the actual OJ. The 95th percentile of IJ overestimates the OJ, showing ratios between 1.2 and 1.5. This difference between the two genres can also be seen when employing 7 x 2 (LL x genre) within-subject analyses of variances for the three concepts with the predicted line length as the dependent variable. There is a significant main effect of the genre in the ANOVA for the 95th percentile, $F(1,19) = 7.50$, $p < .05$, and in

the ANOVA for the mean, $F(1,19) = 10.2$, $p < .01$. By contrast, the main effect of genre is not significant in the ANOVA for the energetic mean, $F(1,19) = 0.0099$, $p = .934$.

4. Discussion

When calculating the overall loudness of music, both models compared in this study, N5 of DIN 45631/A1 and LL(P), highly correlate with the subjective evaluations using the method of magnitude estimation by line length. For the overall judgment, a slightly higher correlation is obtained for N5. The instantaneous judgments permit us to investigate the concepts of a percentile loudness and energetic mean not only for a calculated loudness-time function but also for the actual instantaneous perception of the participants. Not surprisingly, correlations between these predictions based on the IJ and the OJ are even higher, with the energetic mean showing a perfect correlation up to a precision of two digits after the decimal point.

A detailed investigation shows that the high correlations mainly depend on the experiment's dynamic range of 24 phon. As both models are based on the Zwicker model or its first standard, ISO 532 B (1975), respectively, these 24 phon have the same effects on the predictions of both models. By contrast, the differences between the genres make up a difference of roughly 3 phon. This is caused by the different characteristics of rock and classical music. The latter typically has a large dynamic range within a single piece of music, having very soft segments like a solo and very loud segments, for example with the entire chorus. Within a rock song, loudness does not change very much over time. That is why any method for averaging the instantaneous judgments of a rock song comes close to the according overall judgment. For the classical pieces of music, only the energetic mean comes close to the overall judgment, the 95th percentile systematically overestimates it.

The latter finding is just contrary to the results of [1] who found, for road traffic with a duration of 17 minutes, that the energetic mean is systematically smaller than the overall judgment but the N4 loudness matches quite well. However, all studies agree that the mean of instantenous loudness is less than the overall loudness. Both the energetic mean and an ordinal loudness value which is exceeded in a small percentage of the time are somewhere between the mean and the maximum for typical sounds, including music. The correlations did not reveal which of the two models is better, however, the concept of LL(P) shows some advantages regarding the absolute value of overall loudness. As its predictions based on the instantaneous judgment lie within 86 and 108 percent of the overall judgment for all sounds studied, it seems to be an appropriate physical measure of loudness for different kinds of music.

Acknowledgement

The authors wish to thank Prof. Takeo Hashimoto, Prof. Sonoko Kuwano and Prof. Seiichiro Namba for their contribution in developing the software used for tracking the instantaneous and overall judgments during a joint study with the first author at Seikei University.

References

[1] H. Fastl: Evaluation and measurement of perceived average loudness. Contributions to psychological acoustics - Results of the fifth Oldenburg Symposium on psychological acoustics (1991), 205–216.

[2] S. Namba, S. Kuwano, H. Fastl: Loudness of road traffic noise using the method of continuous judgment by category. Proceedings of the International Congress on Noise as a Public Health Problem (1988), 241–246.

[3] S. Kuwano, S. Namba: Evaluation of aircraft noise: Effects of number of flyovers. Environment International **22** (1996), 131–144.

[4] H. Fastl, S. Kuwano, S. Namba: Assessing the railway bonus in laboratory studies. Journal of the Acoustical Society of Japan (E) **17** (1996), 139–148.

[5] H. Fastl: Loudness of Running Speech Measured by a Loudness Meter. Acustica **71** (1990), 156–158.

[6] E. Skovenborg, R. Quesnel, and S.H. Nielsen: Loudness assessment of music and speech. Audio Engineering Society Convention 116 (2004).

[7] S. Namba, S. Kuwano: Continuous multi-dimensional assessment of musical performance. Journal of the Acoustical Society of Japan (E) **11** (1990), 43–51.

[8] DIN 45631/A1: Berechnung des Lautstärkepegels und der Lautheit aus dem Geräuschspektrum - Verfahren nach E. Zwicker - Änderung 1: Berechnung der Lautheit zeitvarianter Geräusche (2010).

[9] S. Namba, T. Kato, S. Kuwano: Evaluation of loudness level of time-varying sounds. Proceedings of Inter-noise 2011.

[10] ANSI S3.4-2007: Procedure for the Computation of Loudness of Steady Sounds (2007).

[11] B.R. Glasberg, B.C.J. Moore: A model of loudness applicable to time-varying sounds. Journal of the Audio Engineering Society **50** (2002), 331–342.

[12] S. Namba, S. Kuwano: The relation between overall noisiness and instantaneous judgment of noise and the effect of background noise level on noisiness. Journal of the Acoustical Society of Japan (E) **1** (1980), 99–106.

[13] S. Kuwano: Continuous judgment of temporally varying sounds. - In: Recent Trends in Hearing research, Festschrift for Seiichiro Namba. H. Fastl, S. Kuwano, A. Schick (eds.). BIS - Verlag, Oldenburg, 1996, 193–214.

[14] S. Kuwano, S. Namba: Continuous judgment of level-fluctuating sounds and the relation between overall loudness and instantaneous loudness. Psychological Research **47** (1985), 27–37.