User Evaluation on Loudness Harmonisation on the Web

Gerhard Spikofski, Peter Altendorf and Christian Hartmann

Institut für Rundfunktechnik GmbH, Munich, Germany

In December 2011, we already published preliminary results of the User Evaluation on Loudness Harmonisation in the NoTube blog. At the same time, we called for more participation to increase the data set for the final analysis. By the end of the year, we gathered the results of over 90 participants. In this paper, we would like to give an update of the preliminary results of December 1st 2011 and an outlook on the entire analysis results which will be presented at the 132nd AES convention (Budapest, 26-29April 2012).

1. DESCRIPTIVE RESULTS

The final results presented in the following are based on the evaluation period from September 19th to December 31th 2011. Besides partners from the NoTube consortium, we invited participants from other communities such as the the EBU audio expert group "FAR-PLOUD" or the "Surround-sound-Forum" within the VDT (Verband Deutscher Tonmeister). The results are based on a database of 94 participants. The first part of the presentation can be considered as purely descriptive. The results with respect to the listening condition parameter under test and assessment of loudness adaptation and loudness range adaptation are presented in terms of calculated percentage collected for each attribute.

In the second part of this interpretation we give an outlook on the interdependence between listening parameters and loudness/loudness range adaptation assessments. The corresponding analysis is based on non-parametric statistical techniques like Spearman Rank Correlation or Kruscal-Wallis H-Test to test the significance of observed differences.

The distribution of the listening level in Figure 1 shows that the majority of participants chose rather low listening levels. The preferred level was Level 3 which was selected by over 38% of the test persons. Because Level 7 is only representative for one single participant, the percentage of Level 7 was combined with that of Level 6.

The distribution of the speaker type in Figure 2 shows a nearly equal representation of headphones (both in-ear and on-ear), built-in speakers and external speakers (Stereo, PC&Stereo with subwoofer) with a predominance of headphones. In the following analysis both in-ear and on-ear headphones are treated as one category.

The distribution of the indicated background noise as presented in Figure 3 shows a clear dominance of weak background noises. Only four users declared to have strong background noise.

The results of the evaluation of both loudness and loudness range adaptation are presented in Figure 4 and Figure 5.

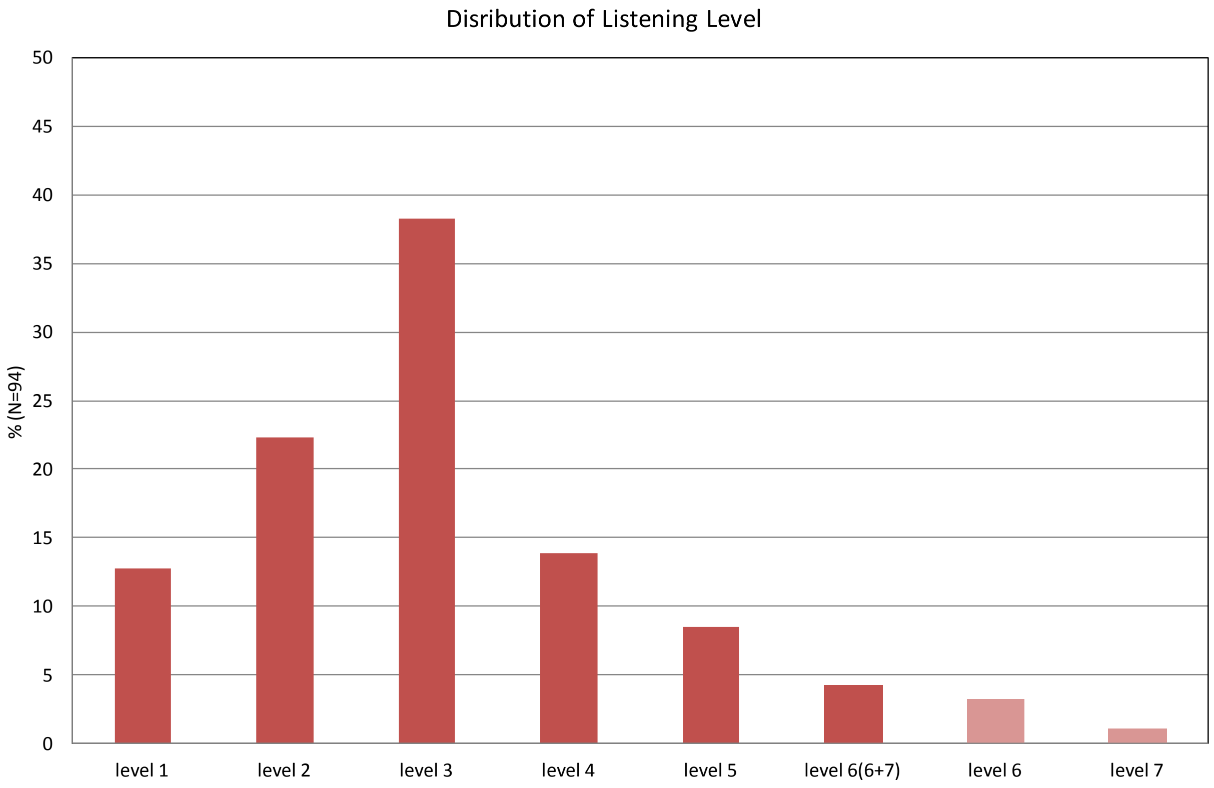

Figure 1: Distribution of listening level

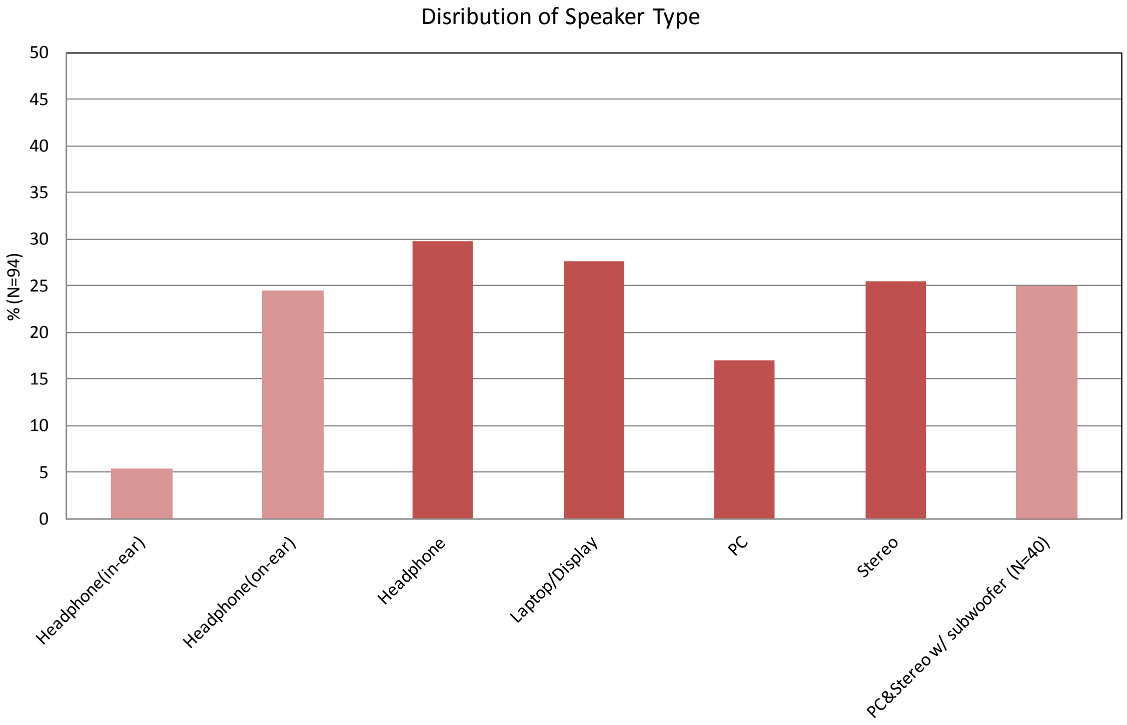

Figure 2: Distribution of speaker type

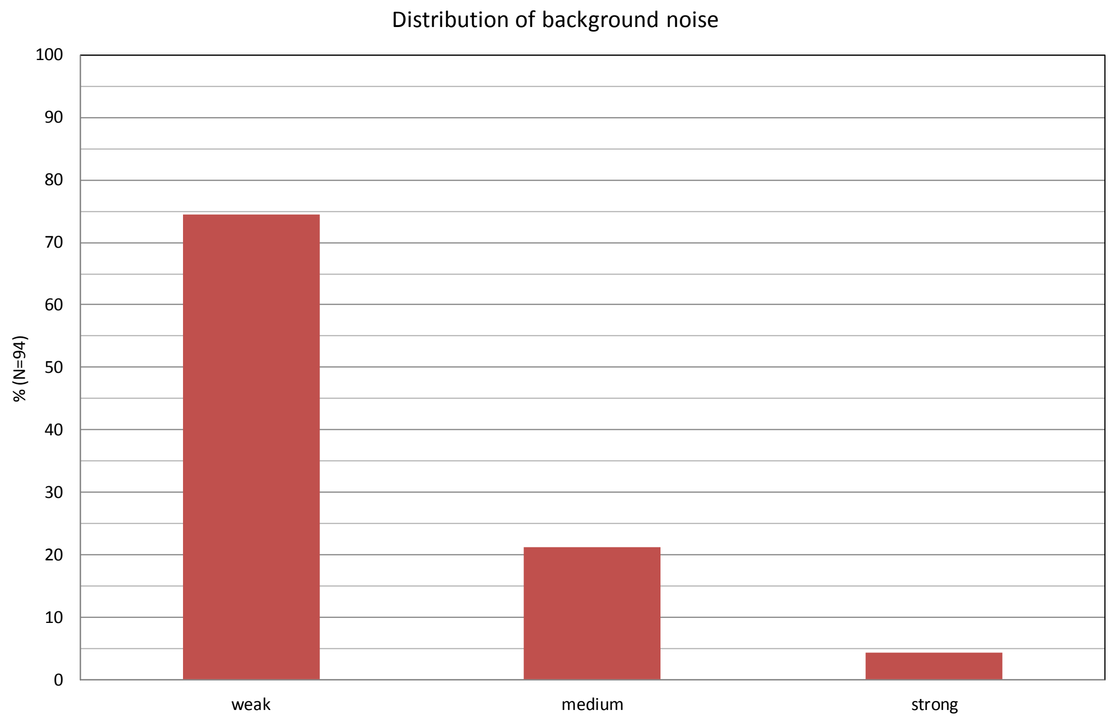

Figure 3: Background noise

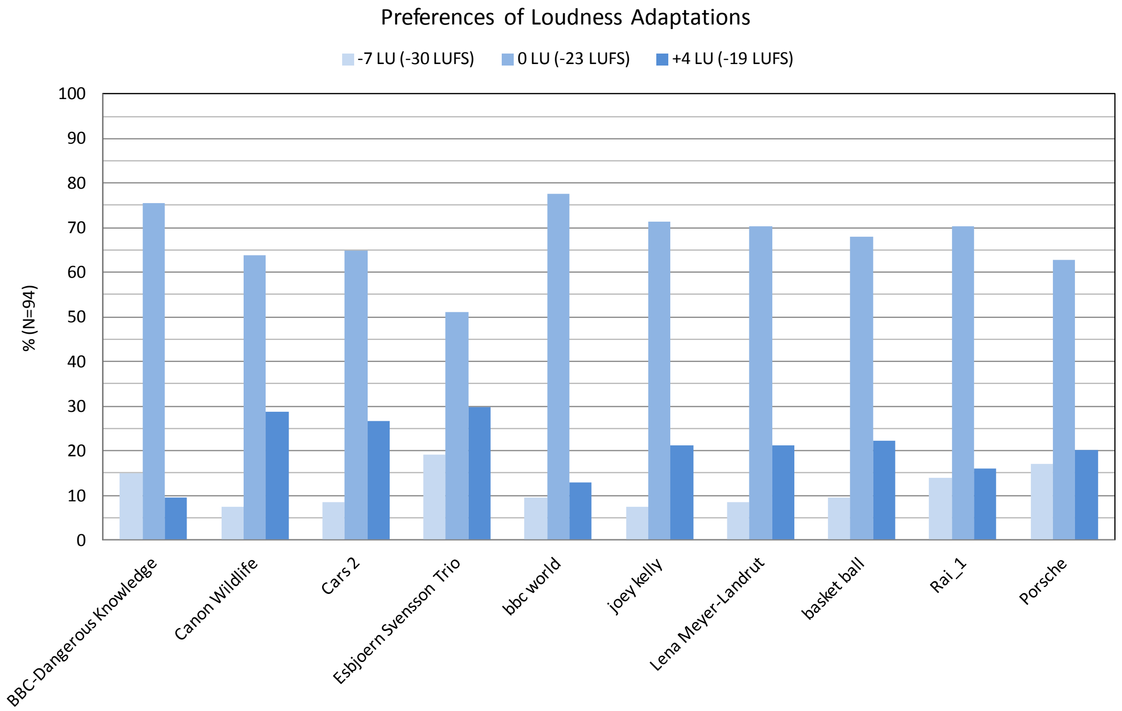

Figure 4: Evaluation of loudness adaptation

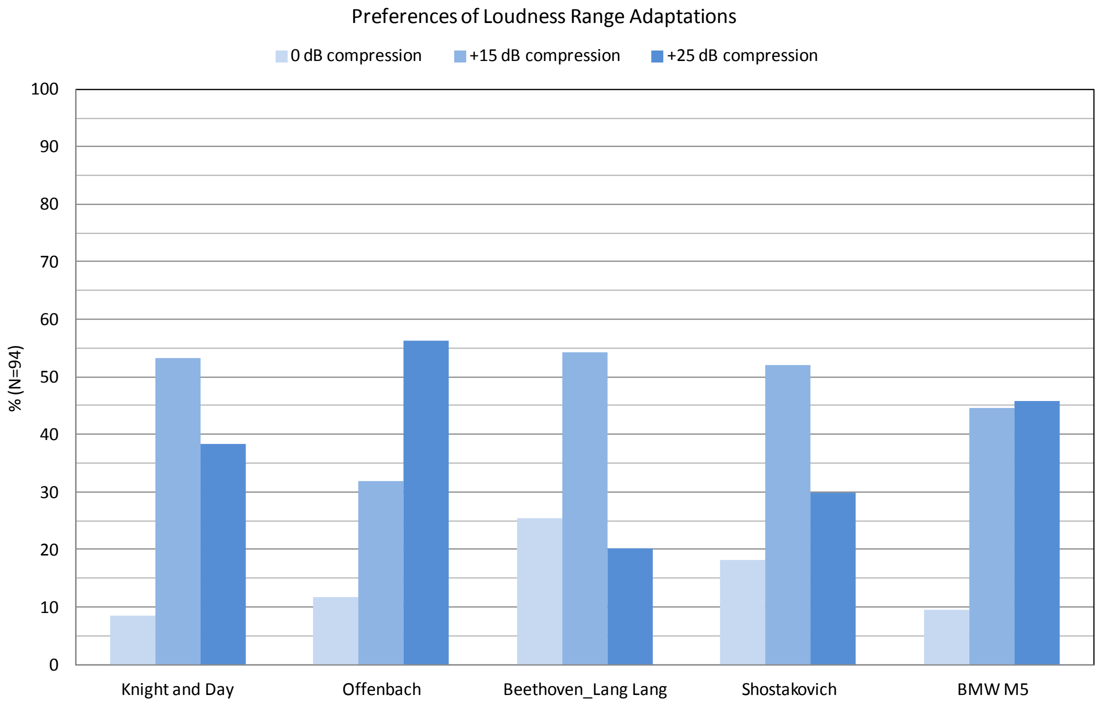

Figure 5: Evaluation of loudness range adaptation

Except for the item “Esbjoern Svensson Trio”, the results of the assessment of loudness adaptation look quite homogeneously. In each case, and by trend also at the excluded item, a significantly clear preference of the loudness normalization to the target loudness of -23 LUFS can be observed (>50% preference for all items). Thus, with respect to the evaluation of loudness adaptation, a first conclusion can be drawn from the descriptive data presented above: As in the previous evaluation, it shows again the excellent performance of the EBU-R 128 loudness normalisation, which verifies its sufficiently accurate approximation of the human loudness perception for broadcast audio signals.

It also answers the open question whether the loudness harmonisation following EBU-R 128 is among others depending on genres or individual listening conditions. Considering the listening conditions covered in this loudness web evaluation, there seems to be no influence observable. Anyhow the aspect shall be investigated in more detail in the second part of this report.

Compared to the results in Figure 4 (loudness adaptation) the presented results of the assessment of the loudness range adaptation in Figure 5 do not seem to appear just as homogeneously. Nevertheless there is an identifiable tendency. The participants seem to prefer rather medium or even strong loudness range compression than uncompressed audio with high loudness range. These results seem to be predestined to investigate correlations between loudness range adaptation aspects and parameters under test like “type of speaker”, “listening level” or “background noise”. An outlook on these correlations is reported in the following section. The full analysis will be presented at the 132nd AES convention in April 2012.

2. ANALYSIS RESULTS

In the second part of the interpretation of the loudness web evaluation results, the interdependence between listening condition parameters and loudness/loudness range assessment are analyzed. The initial point of this analysis is to look at the differences between items under test both for loudness and loudness range adaptation experiments. The basic statistical tools that are used for this purpose are non-parametric methods like Spearman Rank Correlation, Kruskal-Wallis H-Test or Wilcoxon-Test.

Spearman's rank correlation coefficient or Spearman's rho is a non-parametric measure of statistical dependence between two variables. It assesses how well the relationship between two variables can be described using a monotonic function. If there are no repeated data values, a perfect Spearman correlation of +1 or −1 occurs when each of the variables is a perfect monotone function of the other. Besides general information about the distribution under test, the result of the Spearman's rank correlation test delivers simply the correlation coefficient.

The Kruskal-Wallis one-way analysis of variance by ranks is a non-parametric method for testing whether samples originate from the same distribution. It is used for comparing more than two samples that are independent, or not related. The factual null hypothesis is that the populations, from which the samples originate, have the same median. When the Kruskal-Wallis test leads to significant results, then at least one of the samples is different from the other samples. The test does not identify where the differences occur or how many differences actually occur.

The result of the Kruskal–Wallis H-Test is, besides information about the distributions under test, “rank average”, “degree of freedom”, “test value” and the corresponding “probability of decision P”. The Wilcoxon signed-rank test is a non-parametric statistical hypothesis test that may be used when comparing two related samples or repeated measurements on a single sample to assess whether their population mean ranking differs (i. e. the Wilcoxon signed-rank test is a test of paired differences). The results of the Wilcoxon test is, besides information about distributions under test, “rank sums”, “test value” and the corresponding “probability of decision P”.

Assuming a probability of decision of 95 %, i. e. an error probability of 5 %, the actual P-value gives the information about holding the factual null hypothesis or discarding it. If $P \leq 0.05$ the null hypothesis has to be discarded, contrariwise if $P > 0.05$ the alternative hypothesis (no significant differences between samples under test) has to be accepted. In a first global analysis of similarities between all parameters and sequences under test, the Spearman Rank Correlation was applied to all pair combinations.

Parameters	**Items for loudness adaptation**	**Items for loudness range adaptation**
Age	BBC-Dangerous Knowledge	Knight and Day
type of speaker	Canon Wildlife	Offenbach
size of speaker	Cars 2	Beethoven_Lang Lang
distance to speaker	Esbjoern Svensson Trio (e.s.t.)	Shostakovich
background noise	BBC World	BMW M5
	Joey Kelly	
	Lena Meyer-Landrut	
	Basketball	
	Rai_1	
	Porsche	

Table 1: Parameters and items (video clips) under test

The Spearman Rank Test indicated that from all 90 pair comparisons under test only the following pair combinations showed reasonable correlations (R > 70 %).

Spearman Rank Correlation	**Speaker type**	**Distance**	**Size**
Speaker type	1.00	0.88	0.87
Distance	0.87	1.00	0.97

Table 2: Spearman Rank Correlation > 0.70

The resulting interrelations can easily be comprehended. Large speakers need large distance and reversely.

In the next step of the analysis of interdependence, the assessments of loudness and loudness range adaptation are analyzed with respect to significant differences between test sequences. Depending on the results of Kruskal-Wallis, groups of test sequences showing no significant differences were established. The representative results of the extracted sub-groups are than tested with respect to the dependency on the listening parameters under test.

The initial point is again to look at the differences and to identify significant differentiations. If significant differentiations are identified, comprehensive information is obtained about individual listening parameters that influence the global results. Consequently, the further analysis of detailed individual results focuses on the identification of significant differences. Results based on non-significant differences do reflect the overall results and are therefore presented only exemplarily. The individual results for both categories loudness and loudness range adaptation are presented in the following Sections 2.1 and 2.2. Each of these sections starts with the provision of a table that indicates all parameters with respect to the analysis of differences.

2.1. Loudness adaptation

For the test items used, the analysis according to Kruskall Wallis reveals that there are (formally) two groups of test sequences which differ significantly from each other, see the following Table (note that group ldn-b is represented by only one test item).

Group ldn-a P = 0.48 (sorted by rank averages)	**Group ldn-b**
BBC World	Esbjoern Svensson Trio (e.s.t.)
BBC-Dangerous Knowledge	
Joey Kelly	
Lena Meyer-Landrut	
Rai_1	
Basketball	
Cars 2	
Canon Wildlife	
Porsche	

Table 3: Two significantly different groups of test items in loudness adaptation

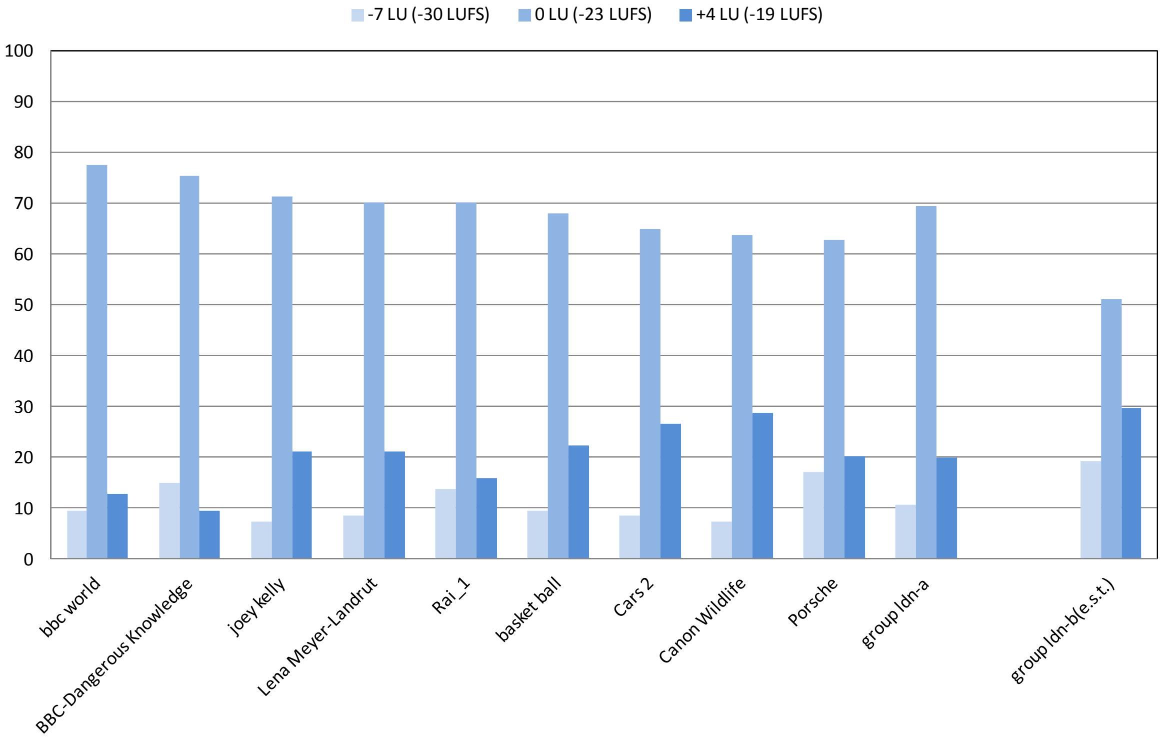

Figure 6: Loudness adaptation – Two significantly different groups of test items

The results in Figure 6 look quite homogenously. The loudness normalization after EBU-R 128 is significantly preferred in both groups of test sequences. In the case of "Esbjoern Svensson Trio (group ldn-b)" the relationship "normalized/attenuated" or rather "normalized/boosted" was altered compared to group ldn-a. A detailed analysis of the interdependence between loudness adaptation and listening parameters will be presented at the 132nd AES convention in April 2012.

2.2. Loudness range adaptation

For the test items used, the analysis according to Kruskall-Wallis reveals that there are four groups of sequences which differ significantly from each other, see the following Table.

Group lra-a P = 0.06 (ordered by rank average)	**Group lra-b P = 0.07 (ordered by rank average)**	**Group lra-c P = 0.16 (ordered by rank average)**	**Group lra-d P = 0.27 (ordered by rank average)**
Beethoven_Lang Lang	Shostakovich	Knight and Day	BMW M5
Shostakovich	Knight and Day	BMW M5	Offenbach
		Offenbach	

Table 4: Four significantly different groups of test items in loudness range adaptation

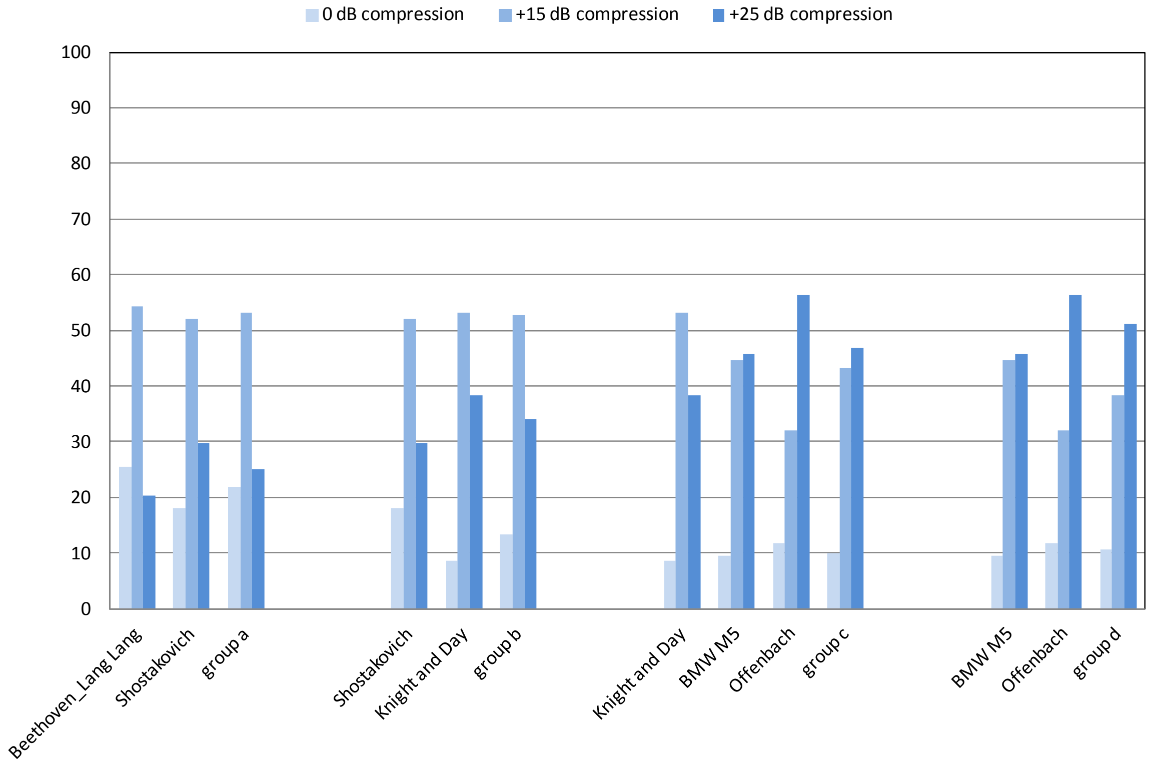

Figure 9: Loudness range adaptation – Four significantly different groups of test items

The already observed inhomogeneity of results considering the loudness range items (see Section 3 Descriptive results) can be documented by the analysis presented here. Four significantly different groups of sequences were analyzed. Following the trend of the results from group lra-a to group lra-d, an upward trend of boosted preference of compression can be observed. From this point of view, the investigation of independence between loudness range preference and listening condition parameters are especially interesting. A detailed analysis of the interdependence between loudness range adaptation and listening parameters will be presented at the 132nd AES convention in April 2012.

3. CONCLUSIONS

Looking at the presented results with respect to the assessment of loudness adaptation, the overall results, covering the collectivity of all listening parameters under test, appear quite homogeneously. In case the individual listening level was adjusted with a speech signal to a target loudness of -23 LUFS, a significantly clear preference for the same loudness value can be observed for every other test item. This again verifies the sufficiently accurate approximation of the human loudness perception by the algorithm defined in EBU R128 and ITU-R BS.1770-2 not only for broadcast audio signal but also for the audio part of video clips presented over the Internet.

Beyond that, the results indicate no dependency between the preferred loudness level and the genre of the test signals. A further statistical breakdown shows a large independence of the required loudness adaption from the listening parameters such as “age”, “speaker type” and “listening level”. Solely a strong background noise seems to increase the need for higher loudness levels. This uncomplex behavior and handling also predestines EBU R128 as a useful tool for loudness harmonization on web based audio content.

Compared to the presented descriptive results of the assessment of the loudness range adaptation, the overall results do not seem to appear as homogeneously. Nevertheless there is an identifiable tendency. The participants seem to prefer rather medium or even strong loudness range compression to uncompressed audio with high loudness range.
In particular with respect to the assessment of loudness range adaptation, a remarkable variation can be observed between the individual test items and for the interdependence between assessment and individual listening parameter.