Audio Engineering Society

Convention Paper

Presented at the 128th Convention
2010 May 22–25 London, UK

Determining an Optimal Gated Loudness Measurement for TV Sound Normalization

Eelco Grimm[1], Esben Skovenborg[2], and Gerhard Spikofski[3]

[1] Utrecht School of Music Technology, The Netherlands
eelco.grimm@kmt.hku.nl

[2] TC Electronic A/S, Risskov, Denmark
EsbenS@TCElectronic.com

[3] Institut für Rundfunktechnik GmbH, Germany
spikofski@irt.de

ABSTRACT

Undesirable loudness jumps are a notorious problem in television broadcast. The solution consists in switching to loudness-based metering and program normalization. In Europe this development has been led by the EBU P/LOUD group, working towards a single target level for loudness normalization applying to all genres of programs. P/LOUD found that loudness normalization as specified by ITU-R BS.1770-1 works fairly well for the majority of broadcast programs. However, it was realized that wide loudness-range programs were not well-aligned with other programs when using ITU-R BS.1770-1 directly, but that adding a measurement-gate provided a simple yet effective solution. P/LOUD therefore conducted a formal listening experiment to perform a subjective evaluation of different gate parameters. This paper specifies the method of the subjective evaluation and presents the results in term of preferred gating parameters.

1. INTRODUCTION

Loudness control in television broadcast has posed a problem for broadcasters and viewers alike for many years. Program normalisation practices based on measuring the level – rather than the loudness – of the program has resulted in loudness jumps occurring both between programs and between channels. The extent of this problem is now so large that TV viewers *expect* to have to adjust the volume to compensate. Moreover, the lack of proper loudness control has enabled a "loudness war" between commercials, which have consequently become hyper-compressed in order to be as loud as possible thus sacrificing good sound quality.

This challenge has received much attention in recent years, reflected in numerous papers on this subject, for instance [1, 2, 3, 4, 5, 6, 7]. When recommendation ITU-R BS.1770 in 2006 standardised a simple loudness measurement method for broadcast [8], an important step towards a practical solution was taken. Various broadcast organizations started to develop solutions based upon the BS.1770. In Europe this development was led by the *EBU P/LOUD* working group [9]. The authors are active members of this group. P/LOUD realized that broadcasting is becoming a more and more fully-automated process, with few occasions of human quality control. The group therefore decided to work towards a single target level for loudness normalization that would apply to all genres of programs, and that could be checked and enforced by automatic procedures within media asset management systems at the moment of ingest and at other points in the broadcast chain.

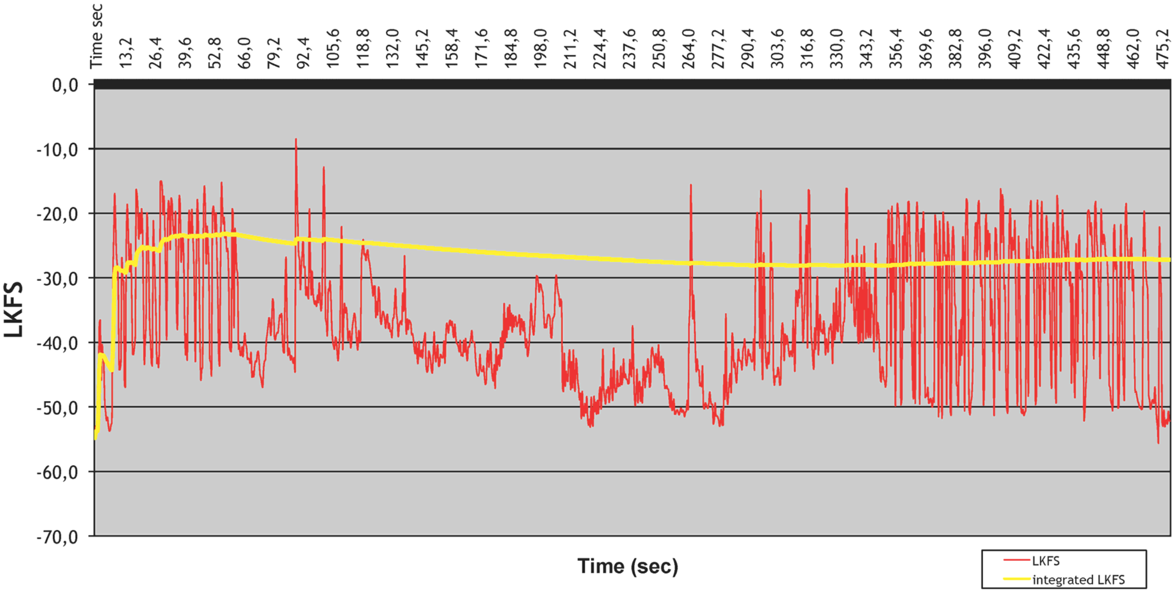

Figure 1. Momentary loudness (red) and integrated loudness (yellow) for 'No Country for Old Men' excerpt.

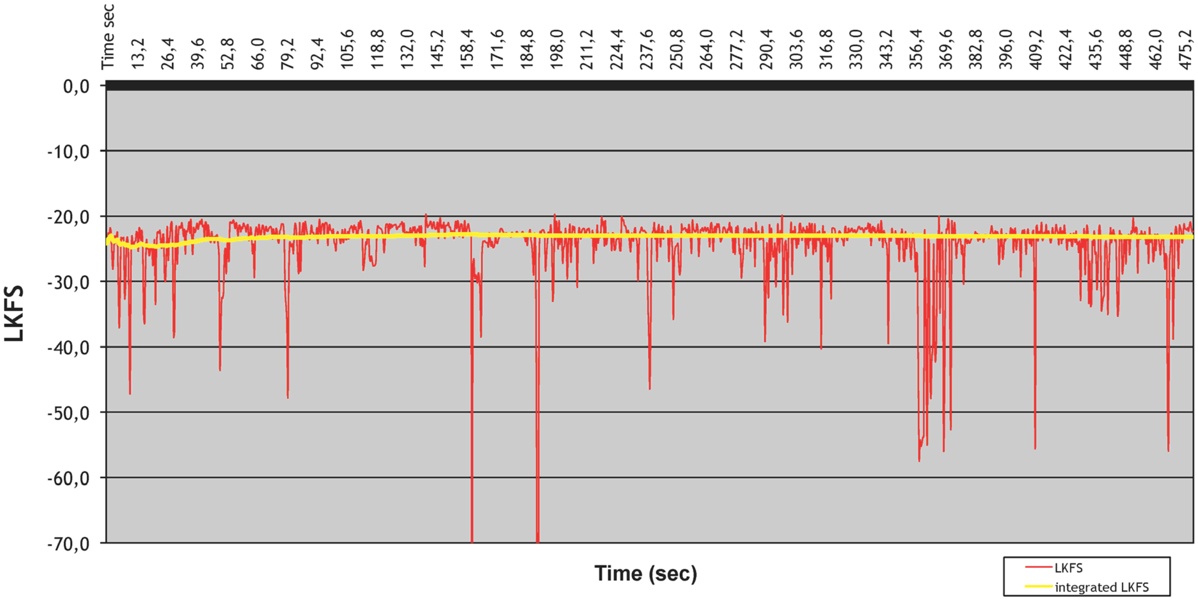

Figure 2. Momentary loudness (red) and integrated loudness (yellow) for a commercial block.

The *Leq*-type of loudness measurement, specified by ITU-R BS.1770, was developed based on loudness-matching experiments involving a set of short, homogeneous sound segments [10]. P/LOUD found that loudness normalization based on this method works fairly well for the majority of broadcast programs, of many different genres. Therefore this open standard would be a good foundation for P/LOUD's intended EBU recommendation specifying loudness normalization in production, distribution and transmission of broadcast programs.

However, during the P/LOUD research process to define the appropriate target level it was realized that the integrated loudness, specified by ITU-R BS.1770, does not ensure a balanced loudness in all cases. An integrated loudness measurement of programs with longer passages of (near-)silence, like the movie 'No Country for Old Men' [11], turns out relatively low. The result is that after loudness normalisation the movie dialog will be considerably louder than that of strongly compressed programs like typically commercials. P/LOUD realised that the risk of commercial blocks being played out at lower loudness levels than some movies would hardly be accepted by broadcasting stations. In other words, the investigations indicated a problem adapting ITU-R BS.1770 loudness normalization to material like 'No Country for Old Men'.

Let us consider this example in more detail. Figure 1 shows the measured momentary loudness for an excerpt of 'No Country for Old Men', together with the integrated loudness level over time. This plot shows that the excerpt contains relatively long periods of near-silence, where just prairie wind and distant noises are heard. Therefore the measured integrated loudness of the movie is relatively low, compared to a typical commercial block, aligned to the same subjective loudness (figure 2).

Several options to solve this problem were considered in the P/LOUD group.

In order to automatically determine "dialnorm" (dialog normalization) meta-data, Dolby had developed (and patented [12]) an algorithm named *Dialog Intelligence* [13]. This algorithm is capable of recognizing regions of speech and only measure loudness of these regions of the program. P/LOUD experiments, however, revealed that in more than a few cases the algorithm was not capable of recognizing speech in the presence of music, as commonly used in commercials. Moreover, the different use of speech in different genres can generate unwanted variation in normalization, even if an algorithm could reliably distinguish speech. Since the goal of P/LOUD was to define loudness normalization of broadcast-ready material – regardless of genre – that would not require human intervention, it was decided to pursue a different method working independently of the signal type (like music, speech, or sound effects).

Another proposal considered was to introduce a "permitted interval of program loudness" instead of a single target loudness level [14]. However, P/LOUD noted that this proposal would not force sound engineers to adjust the loudness of a movie correctly. So the risk would remain that some movie might sound considerably louder than the adjacent commercial block, even when both programs would be inside the permitted interval. This proposal was therefore also dismissed.

2. A GATED, INTEGRATED LOUDNESS MEASUREMENT

The sound of many movies can be considered as consisting of *foreground sound* such as speech or the sound of "the action", and *background sound* which could be environmental noise (e.g. rain or wind) or the sound of something happening "in the background". Although these concepts are difficult to define rigorously, the preliminary investigations of P/LOUD pointed to the need for a measurement of program loudness which would exclude some of the background sound. It was therefore concluded that a measurement-gate might be needed in addition to the integrated loudness level of ITU-R BS.1770-1 [15], in order to pause the measurement when the loudness level drops below a certain threshold. Informal tests indicated that this strategy could indeed offer a solution.

Previously, TC Electronic had introduced a loudness descriptor, *Center of Gravity* (CoG), which measures the overall loudness of the program of any genre, based on BS.1770. The CoG employs an *adaptive gate*, which enables the CoG to be robust against regions of (near-) silence in the program without making any rigid assumptions about the absolute levels of the material to be measured [7]. By employing this gate, the CoG essentially ignores regions which are too quiet to be considered part of the program. The qualitative investigations of P/LOUD showed that although the principle of the CoG's measurement gate was beneficial, a somewhat more 'aggressive' gating would be desirable for the measurement of movies such as 'No Country for Old Men'. In other words, not just the silence had to be excluded but also (some of) the background sound of the program, in order to measure the so-called *Foreground Loudness* [16].

A commercial typically has a Narrow Loudness Range (NLR), while movies are typically Wide Loudness Range (WLR) programs. As an objective measure for this concept, the *Loudness Range* (LRA) was introduced by TC Electronic [7, 16]. LRA is based upon the cumulative loudness level statistics, in combination with an adaptive measurement-gate, defining the range between 10% and 95% of level distribution as *Loudness Range*. Thus, WLR programs may have a Loudness Range of for example 20 LU, whereas the Loudness Range of NLR programs is typically less than 5 LU.[1]

P/LOUD's qualitative investigation indicated that there seemed to be three types of WLR material:

- **WLR(1)** type has lots of loud parts and far fewer silent parts. This is the familiar ‘blockbuster’ action movie like “The Terminator”. The predominant loud parts will lead to a relatively high integrated loudness level which after normalization means the anchor part of the movie may be lower than that of a commercial. This seems acceptable because the major loudness impression of the movie is primarily defined by the action sequences rather than the speech. In any way, adding a gate to the measurement would not help in aligning anchors of both types better because one can only gate out the low level parts from the measurement, not the loud parts.

- **WLR(2)** type is where the proportion of the louder parts and the softer parts are in balance. There is a proper alignment after loudness normalization – even without a measurement gate. We should be careful that a gate does not change this balance, the gate should therefore not have a too high threshold.

- **WLR(3)** type has considerably more silent parts than loud parts. To give an impression: since ITU-R BS.1770-1 basically specifies an Leq measure, one needs to add 50% of silence to a fragment to lower the integrated loudness level by 3 dB. WLR(3) is the scarcest type of WLR programs. Although members of the P/LOUD group have put large effort into searching for appropriate WLR(3) testing material, in less than 25% of the submitted fragments a 'medium' gate had more than 2 dB influence on the resulting measurement. However, as the 'No Country for Old Men' example showed, the WLR(3) programs do confront us with a serious alignment problem. Subjective alignment of such movie against NLR material showed the appropriate alignment-gain of the movie to be 4 to 5 LU lower than determined by an (ungated) integrated loudness measurement.

P/LOUD concluded that adding a qualified measurement-gate to the integrated loudness seemed to be a simple yet effective solution for measuring the program loudness of WLR as well as NLR programs.

3. DIFFERENT GATING METHODS AND PARAMETERS

In *gated, integrated loudness measurements* the following gating parameters have to be considered and defined:

Gate threshold: This parameter has to be considered as the most important one. If the threshold is too high a fairly large proportion of the WLR program being measured may be gated out, resulting in a too low alignment-gain. On the other hand, if the threshold is too low the measurement-gate may not fully have the desired effect.

Gate type: We distinguish between three types of gates, denoted *fixed*, *relative*, and *recursive*, respectively.

- With a **fixed-threshold gate**, the gate threshold has a fixed, absolute loudness level (i.e. LKFS value). Since the target level is also fixed, there is a fixed distance between target level and gate threshold. This is the simplest of the gate types. Its main disadvantage is that a measurement with fixed gate almost never provides the correct normalization gain figure because the proportion of a program which is gated out varies when the absolute level changes.

- In contrast, the gate threshold in the **relative-threshold gate** (a.k.a. dynamic gate or adaptive gate) is relative to the ungated, integrated loudness level. In effect, the distance between target level and gate threshold level is program dependent. A

[1] LU = 'Loudness Unit', [17]. The LU is equivalent to a dB.

measurement with relative gate is independent of the absolute level of the program and therefore always provides a correct normalization gain figure, regardless of the playback gain.

- The relative-threshold, **recursive gate** (so-called iterative gate) is the last type considered. The gate threshold is now relative to the *gated* integrated loudness level. Because the threshold is specified relative to the output of the gated measurement itself, this is a recursive definition of a gate. An iterative process during the loudness calculation makes sure that the distance between gated loudness level and gate threshold is kept constant. Since the gated loudness measurement will be used for determining the normalization gain, after processing there is also a fixed distance between target level and gate threshold. A measurement with recursive gate is independent of the absolute level of the audio, similarly to the relative gate type. The disadvantage of a recursive gate is a higher complexity – both conceptually and implementation-wise – compared to the other two types.

Gating block length: The third gate-parameter is the gating block length. How long should a silence last before it is gated out of the measurement? The block length should not be too short because then every short breathing pause would be excluded. It also should not be too long because that would reduce the effectiveness of the gate, which could only be compensated by raising the threshold of the gate to quite high levels. One of the results from a preliminary experiment on a WLR data set showed that a -7 LU / 250 ms relative gate gave similar integrated loudness levels as a -6 LU / 500 ms and a -4 LU / 1000 ms relative gate. It was decided by P/LOUD to select a 400 ms block length for the gating types to be investigated in further experiments because it seemed to be a good compromise. Moreover, 400 ms matches the block length chosen for the “momentary loudness” in the EBU Mode loudness meter [18].

Mid 2009 P/LOUD decided to conduct a listening experiment to perform a subjective evaluation of different gate parameters, in order to determine a measurement gate that would be preferable from subjective point of view, yet would not distort the loudness measurements of NLR segments for which an ungated measurement works well. It was evident the *fixed* gate type had serious, principal disadvantage, but preliminary investigations indicated good results with both the *relative* and *recursive* gate types. Hence P/LOUD concluded to ignore the fixed gate type, but to include both the relative and recursive types in the test.

4. DESIGN OF THE (NEW) SUBJECTIVE EVALUATION EXPERIMENT

P/LOUD concluded that subjective testing was required to determine the most preferable gate threshold. It was realized that neither standardized subjective test methods like ITU-R BS.1116 and ITU-R BS.1534 [19] nor basic loudness-matching experiments would work. The loudness of WLR(3) type material with near-silence parts is practically impossible to assess with fragments of a short duration. The result would depend entirely on the particular selected, edited fragment, rather than on the WLR program as such.

In this context the fundamental question had to be answered: how to make the fragment under test representative for the full program? P/LOUD performed a preliminary experiment in which a number of WLR programs were submitted by P/LOUD members. Each submitted test material was accompanied by a short fragment that was considered as its *anchor*. However, it turned out to be quite difficult for the participants to select the anchor parts. Should it be speech, speech with background music, dominant foreground sound...? What about pure music material? Everyone who participated in this experiment seemed to have a different view on what the characteristics of a matching anchor were.

Because P/LOUD could not fully answer the question of finding representative anchor parts of the WLR(3) type material, it was concluded that loudness-matching experiments as used in previous subjective loudness research [10, 20] would not be suitable for this purpose. In the P/LOUD experiment to test the gating parameters, the WLR fragments would need to be long enough to ‘naturally’ contain both foreground and background sound.

4.1. Experimental procedure

P/LOUD concluded to design the following new subjective evaluation test. In this test subjects do not have to adjust loudness themselves, but instead compare a few longer sequences and then indicate which one offered the most consistent loudness experience. It was decided to emulate a “home television evening”. In the end that should be the goal of the designed experiment: to obtain a loudness normalization that can provide a nicely levelled television session at home. One consequence is that the listening level would accordingly have to be relatively low (pink noise -18 dBFS RMS according to EBU Tech. 3276 [21] aligned to 70 dB(C), two speakers). After some discussion it was decided to carry out the listening experiments in studio environments referring to EBU Tech. 3276, mainly to ensure controlled, uniform listening conditions. Moreover it was assumed that the selected subjects would consist of broadcasting engineers, experienced in assessing loudness differences and used to attentive listening for longer periods of time.

The test procedure was designed as *double-blind* experiment – the gating types under test were designated anonymously: A, B, C, D. Neither test instructor nor test subjects were aware of the assignment. The procedure allowed subjects to listen solo or in a group. They were asked to assess sequences of alternating long (up to 6 minutes) WLR(3) material and short (less than 30 seconds) NLR material, all normalized using one type of gating. The sequences had to be experienced without pauses or fast forward actions. Considering for instance 9 test items, this means continuously listening for about 20 minutes. Afterwards playback was changed to another sequence of the same fragments in the same order, but normalized using another gate type. This was repeated with the other two sequences.

After the listening session the participants had to rank the sequences in their preference order. The primary question to the subjects was: "*Which of these sequences provides you best with an experience in which you find no need to adjust the volume?*". It was stressed that the subject should not evaluate the loudness fluctuations *within* individual fragments, but only of the fragment's average loudness in the context of the sequence as a whole. As a secondary response, subjects were asked to rate all four sequences as either ‘poor', 'appropriate', 'good', or 'excellent'. If subjects listened in a group, they were not allowed to discuss their experiences before ranking the sequences individually. In the questionnaire space was left to note qualitative statements individually and after discussion in the group.

Next, a second and a third series with other audio fragments were run, with breaks in between. Naturally the key of gating types was unknown to everyone and the order of types was randomized in each session. After a run the sequences were first ranked and afterwards the sequence key (A, B, C and D) was revealed in order to fill out the questionnaire. Since each series of sequences lasted for about one and a half hour, the total length of the test, including breaks and discussions, was about six hours. Although this type of listening test is uncommon, it does relate closely to the way professional broadcast engineers perform their jobs and we were confident that they could perform this task well.

4.2. Selection of Stimuli and Conditions

The first thing we had to do was to gather enough WLR(3) material as test stimuli. It was decided to use a pre-screening procedure by comparing the ungated integrated loudness level of a candidate with the -6 LU / 400 ms relative gated measurement of the same file. As 0.5 LU was found to be an average difference between these two measurements on medium loudness range material, and 2 dB was considered as the smallest relevant loudness difference, a minimum difference between both measurements of 2.5 LU was used as *pre-screening criteria* for the stimuli. The pre-screening would ensure that the ‘contrast’ in the experiment would not be too small, due to the stimuli selection. In total a test set of 24 pre-screened WLR(3) fragments were compiled. We added a pool of 27 NLR fragments to that. This set of fragments was selected to be representative in genre, as summarized in table 1.

	Genres / sources of fragments
Wide Loudness Range (WLR) fragments	drama, documentary, classical music, sports
Narrow Loudness Range (NLR) fragments	commercial, promo, announcement

Table 1. Genre and source of the fragments used as stimuli in the main experiment.

The next step was to establish the conditions under test, that is, the gate types and thresholds. Analysis of the selected WLR fragments showed that the average difference between ungated measurement and -20 LU relative gate measurement was smaller than 1 LU. It was therefore decided to exclude the ungated measurement as a condition, and select -20 LU instead because P/LOUD felt this type of 'silence' gate would be the minimum specification with respect to the intended recommendation. Further analysis showed that the difference between -10 LU relative gate and -10 LU recursive gate was smaller than 1.5 LU on 85% of the fragments, with approximately 2.6 LU difference for the remaining 4 items. It was considered as a too small difference to include the -10 LU recursive gate in the extensive listening test. Since the 'relative' gate algorithm is more efficient than the 'recursive' one, -10 LU relative was selected for the test. The other candidates were -6 LU relative gating and -6 LU recursive gating. These two types of measurements did show differences larger than 1.5 LU in 50% of the fragments, in some cases reaching as much as 9 LU! Thus both were included in the test. Higher gate-thresholds, such as -3 LU, were dismissed as being too close to the target level and therefore potentially gating out too large amounts of the program. Intermediate steps had a too small difference to be noticeable. Thus the resulting four gate types to be tested were: *-20 relative*, *-10 relative*, *-6 relative* and *-6 recursive*.

4.3. Gating types and stimuli

The table 2 shows the overall effect of the 4 different gated measurements, relative to the corresponding ungated measurement. The average differences (in LU) between gated and ungated loudness levels are shown for the NLR and WLR fragments under test, as well as the maximum and minimum differences for the WLR fragments. Recall that these WLR fragments are *not* representative for WLR material in general, because of the pre-screening of the WLR fragments under test. Keeping in mind that the -20 LU relative gate on average only differs from the corresponding ungated measurement by less than 1 LU, we shall assume that the results of the -20 rel. gate represent the *ungated* condition.

5. QUALITATIVE ANALYSIS OF THE EXPERIMENT

The main assessment for the subjects to make was to ranking of the 4 conditions. Statistical analysis of the ranking results is given in section 6. As a secondary response we asked for quality-scale ratings to facilitate further analysis of the results. Since the subjects were performing the test in their own environment, the test instructions could only be supplied in written form. Although this had the benefit of a larger group of people participating in the test, the disadvantage was that not all subjects were verbally instructed by the same person. As a consequence of this, we found discrepancies between subjects in their interpretation of the 4-level rating scale to be used for the qualitative ratings. We therefore decided to base our main results on the *rankings* only.

Nevertheless the submitted ratings also provided some insight. Most interesting was to view the distribution of the rating 'poor'. This rating was given to the -6 relative gate by 8% of the subjects and to the -10 relative gate also by 8% of the subjects. In contrast the -20 relative gate was judged 'poor' by 36% of the subjects and the -6 recursive gate by 41% of the subjects. The -10 relative and -6 relative were rated as ‘good’ or ‘excellent’ by 69% and 62%, respectively.

5.1. Comments from test subjects (a posteriori)

Subjects were requested to give general comment on the test. This turned out to be very valuable. Also discussion that took place during the P/LOUD meeting shortly after the test provided valuable additional information about how the subjects performed and experienced the test. Encouraging was that no participant judged the test as too difficult. Surprisingly the long duration of the test was not judged as being problematic. The contrary, we received notes like "Good test in general", "It was a fascinating experience" and "Excellent test!". One subject said this type of listening test forced him into a “holistic listening” attitude that helped focussing on the perspective of the test (“a television evening”). Another comment we received – more than a few times – was that all four sequences were much better aligned in loudness than the current reality of television broadcast. This confirms the importance of loudness normalisation in general.

Gating method	*Avg. diff. ungated NLR fragments (LU)*	*Avg. diff. ungated WLR fragments (LU)*	*Max. diff. ungated WLR fragments (LU)*	*Min. diff. ungated WLR fragments (LU)*
-6 rec.	0.2	5.6	11.7	3.2
-6 rel.	0.2	3.7	7.5	2.6
-10 rel.	0.0	2.5	5.8	1.1
-20 rel.	0.0	0.7	2.4	0.0

Table 2. Average and maximum effect of the 4 measurement-gate types on WLR and NLR fragments.

Two categories of qualitative feedback gave very interesting extra information about the test results. The first was that many people mentioned the *classical music* fragments were relative low in loudness compared to speech in WLR or NLR fragments. Many subjects that ranked higher gate thresholds like -6 relative as best overall performer noted an exception for classical music that seemed better treated with the -20 relative gate. Some people marked that classical music is different from other genres because the prime element can vary from a solo violin to a 'tutti' brass section. Therefore the loudness of the so-called 'foreground sound' that attracts the viewer's main attention varies in classical music much more than for instance in a movie, where the foreground sound is mainly speech. Because classical music covers only a small percentage of television broadcasts most participants were prepared to compromise on this point, but the feeling that classical music was best served with the lowest gate threshold was shared unanimously.

The second category of interest was people appeared to be divided in two almost equal groups, one tending a bit more towards higher gate thresholds, the other a bit more towards lower gate thresholds. During the P/LOUD meeting this was discussed and we found that people who were concerned about loudness jumps at the moment of transition and/or focussed more on the foreground sound tended to select a higher threshold (i.e. -6 rel, sometimes even -6 rec) where people that focussed more on *overall loudness consistency* tended to prefer the lower threshold (-10 rel, sometimes even -20 rel). This indicates that subjective taste does have an influence on the outcome.

Two other interesting remarks were given. One is the selected acoustical level was judged as a little low – some parts of the classical music fragments were almost silent with certain sequences. Others, however, commented it was the right choice, given the non-optimal listening conditions in a typical living room with limited dynamic range playback. The other remark was a few people missed the corresponding video with the sound. We had omitted the video for practical reasons (download size and authoring time). Some perspective effects however were experienced as disturbing, and would probably not have been that with the related scene on screen.

5.2. Inconsistent subjects

The change of the alignment-gain that a gate would cause, relative to the ungated measurement of integrated loudness, indicates an ordering of the 4 gate types. This ordering is determined by the row-order of the 4 gating types in table 2: -20 rel. < -10 rel. < -6 rel. < -6 rec., corresponding to the how large an effect the gate has. This ordering can be used to construct a test of whether the response of a given subject is consistent (with itself). For instance, if some subject would select the -20 rel. gate as first preference, selecting -6 rec. for second preference would not be cogently consistent. Based upon this criterion, we constructed a list of 'consistent' preference rankings, where the 2nd and 3rd preference were closer to the 1st than the 4th. We added orders where two adjacent options were swapped, especially for ranking 3 and 4. What was remaining were preference rankings that showed strong jumps in alignment gains (table 3). These were classified as *inconsistent responses*.

Inconsistent responses were found for 9 subjects, all from the same 2 test sites. Comments recorded via the test instructor at one of these sites, indicated that the subjects in question might have misunderstood the task and/or the instructions given prior to the experiment. Thus, the consistency test provides an objective means for checking whether the test subject has understood the task at hand – as an addition to the comments provided by the test subjects.

1st	***2nd***	***3rd***	***4th***
-20 rel.	-6 rec.	-6 rel.	-10 rel.
-6 rec.	-20 rel.	-6 rel.	-10 rel
-10 rel.	-6 rec.	-20 rel.	-6 rel.
-10 rel.	-6 rec.	-6 rel.	-20 rel.
-6 rec.	-10 rel.	-20 rel.	-6 rel.
-6 rel.	-20 rel.	-6 rec.	-10 rel.

Table 3. The six inconsistent ranking orders.

6. STATISTICAL ANALYSIS OF THE RESULTS

The result of the experiment consisted of the responses of a total of 54 subjects from 14 participating broadcast stations and labs in Europe. The primary response from each subject was a ranking of the 4 conditions in terms of subjective preference, as guided by the provided instructions. Suppose we name the 4 conditions (i.e., the 4 gate types) A, B, C, and D. Let us furthermore assume that the conditions were sufficiently different such that the subject could in fact form a preference without "ties" – an assumption which seem to be confirmed by the qualitative feedback, see section 5.1. Now suppose one subject's response was B, D, A, C, meaning that condition B was his first preference, D the second best, etc. We could now use "integer scoring" to map this response into a sample (3,1,4,2) corresponding to the rank of each condition. However, we could *not* assume that the distance - in terms of the subject's preference - from 1 to 2 is equal to the distance from 2 to 3, and so on. It would therefore not be valid to apply parametric statistics such as ANOVA-analysis or any other type of normal-distribution based statistical test. Instead, we shall consider the responses as belonging to an *ordinal scale*, and thereby make no assumptions of the relative distances between the preference-levels [22]. This type of data can be analysed using methods of *non-parametric statistics* [22].

Responses from 9 subjects do not fulfil the criteria of the “consistency test”, described in section 5.2. The consistency test was formulated *a priori*, that is, before the subject responses were analysed. Moreover, the consistency criteria are based on objectively measurable properties of the individual gate-types. Thus the consistency criteria are independent of the statistical analysis of the results, presented below. We therefore find it reasonable to consider the responses of these 9 subjects as *inconsistent*, and we shall regard the analysis of the data *without* the inconsistent responses as our main results.

First we need to test whether the 4 conditions were in fact rated differently by the subjects. That is, we want to test whether or not the 4 samples could have been drawn from the same population. The Friedman test and the Kruskal-Wallis test are both non-parametric procedures analogous to the analysis of variance [23]. Both these tests are based on ranks of the responses. However, the Friedman test is a test for the significance of the difference among the distributions of *dependent* samples, whereas the Kruskal-Wallis assumes *independent* samples. In our experiment, the 4 samples in each observation are clearly not independent because they consist of a *ranking* of the 4 conditions. Therefore we shall use the Friedman test.

The results of the Friedman two-way analysis of variance are shown in table 4. The null-hypothesis in this test is that there are no significant differences between the mean ranks of the 4 conditions, in other words, that the variation in the responses for the 4 types of gating might as well be coincidental. In the tests we shall use 95% confidence interval (alpha=0.05). The results show that we can reject the null-hypothesis, both for the data set including all responses (p=0.018), and for the data set excluding the inconsistent responses (p=0.0015) (table 4). That is, we can conclude that the different gating types were *not* all rated as equally preferable. Furthermore, note that the p-value is considerably lower for the data set excluding inconsistent responses, even though 9 observations are removed which would normally increase the p-value if these 9 responses had been of the same quality as the others.

Source	***SS***	***df***	***MS***	***Chi-sq***	***Prob>Chi-sq***
Columns	25.62	3	8.541	15.37	0.001524
Error	199.4	132	1.510		
Total	225	179			

Table 4. Friedman's ANOVA Table (n=45, excluding inconsistent responses)

Now that the Friedman test has shown that the tested gate types are not all the same, the next task is to perform a test to determine which *pairs* of gates are significantly different, and which are not. For this purpose we can use the multiple comparisons procedure [23].

The two figures illustrate the multiple comparisons test of the 4 conditions (alpha = 0.05), for all 6 pairs of conditions. For each gate type, the dot indicates the mean-rank. If the lines for 2 gate types overlap in the figure, one is not significantly better than the other. Tukey's HSD criterion (a.k.a. Tukey's range test) was used in the computation of the critical value.

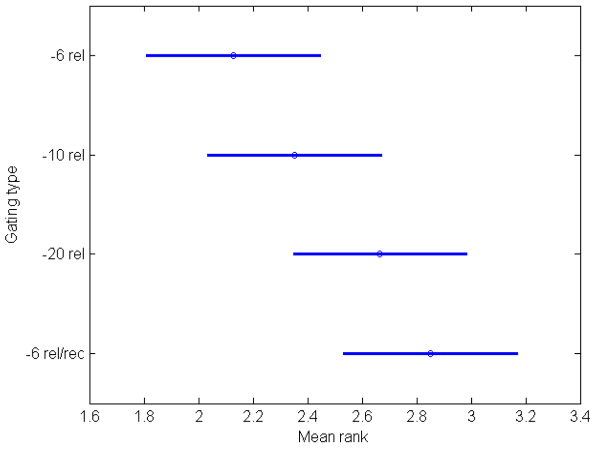

Figure 3: Multiple comparison (n=54, all responses included)

Table 5 shows that the -6 and -10 LU relative-threshold gates received the best preference ratings, and that the difference between these two is *not* significant. Furthermore, the -6 relative-threshold gate is better then the worst 2 gates, with statistical significance.

When considering the *statistical significance*, in the analysis of experiments, it is important to remember that the significance – at a given confidence level – depends not only on the magnitude of the actual/true differences between test conditions, but also on the discriminative power of the experiment as such, and on the number of subjects participating.

Gating method	***Mean rank***	***Significantly better than***
-6 rel.	2.04	-6 rec., -20 rel.
-10 rel.	2.22	-6 rec.
-20 rel.	2.77	(none)
-6 rec.	2.95	(none)

Table 5. The preference of the 4 gating types. Lower mean rank means 'better' preference rating (n = 45, alpha = 0.05)

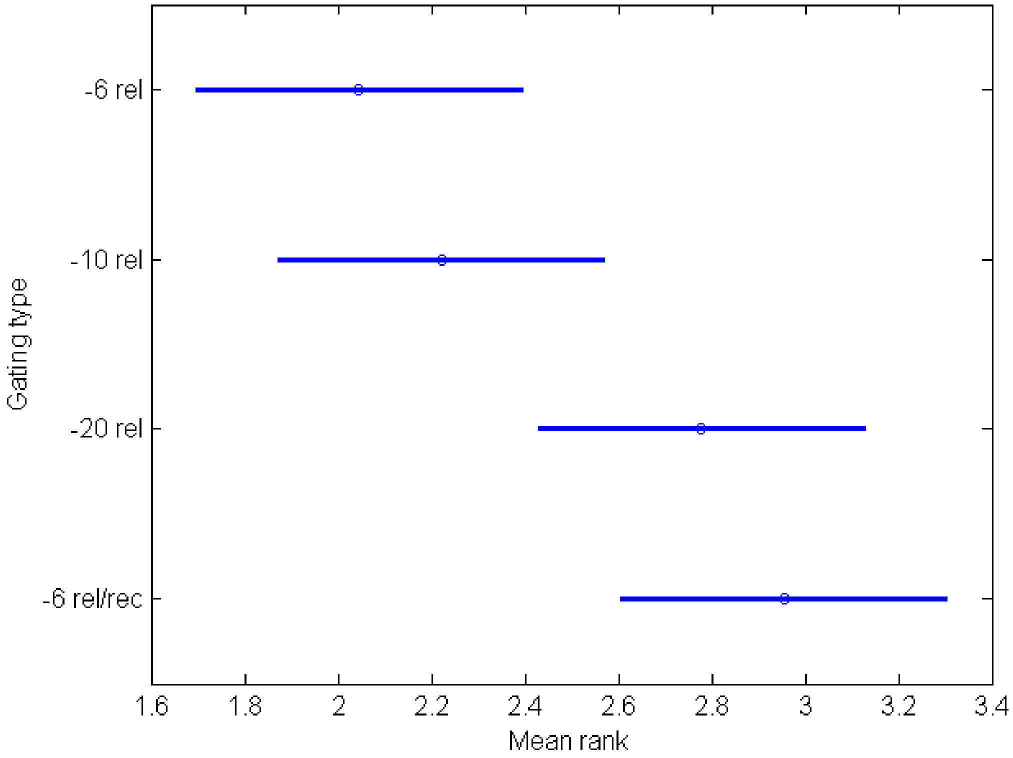

Figure 4: Multiple comparison of the 4 gating types (n=45, excluding inconsistent responses)

7. CONCLUSION

Investigations by EBU P/LOUD group indicated: An ungated integrated measurement of loudness level, as specified by ITU-R BS.1770-1, works well for loudness normalisation and alignment for the majority of program types, and is an effective solution to the problem of too loud commercials and promos. Thus, the integrated loudness measure would provide a solid basis for the broadcasters' transition from program alignment based on level to one based on loudness. However, P/LOUD found that ITU-R BS.1770-1 – when used directly – is insufficient for loudness normalisation of certain types of Wide-Loudness Range (WLR) programs, such as movies with a considerable proportion of silent background sound. Introducing a measurement-gate can solve this problem. Proper choice of gating method and parameters is critical for the gate to be effective while not ruining "backward compatibility" (referring to BS.1770-1) of measurements of Narrow Loudness-Range (NLR) material.

A listening experiment was then performed within P/LOUD: A loudness-matching experiment of the type underlying the ITU-R BS.1770 could *not* be used because the new test needed to include WLR segments as stimuli. Therefore we designed a new experimental method, involving relatively long "holistic listening". Each subject would rank 4 gate types, pre-selected as being promising yet noticeably different. Stimuli were pre-screened to "exercise" the gates, and to represent different genres of WLR material. The test question was: "*which of these sequences best provides you with an experience in which you find no need to adjust the volume?*"

Results from the experiment: 54 subjects from 14 different broadcast stations or labs in Europe; responses of 9 subjects were inconsistent and hence excluded. As stimuli, random sequences of interleaved subsets of 24 WLR fragments and 27 NLR fragments were used. Gates with -6 LU and -10 LU relative-threshold, 400 ms blocks, were rated as best; the difference between these two was small, and *not* statistically significant. The -6 LU relative-threshold gate rated better than the worst two gates, with statistical significance. Preliminary reports from related experiments conducted at NHK and ARIB in Japan indicate similar results.

These findings can now provide the basis for P/LOUD’s upcoming EBU recommendation R128.

8. ACKNOWLEDGEMENTS

Numerous members of the P/LOUD group contributed to the experiment described in this paper with helpful comments and/or as test subjects. Special thanks to Richard van Everdingen for his assistance in preparing graphs and measurements for this paper.

9. REFERENCES

[1] Klar, S. & Spikofski, G. (2002) "On levelling and loudness problems at television and radio broadcast studios", in Proc. of the AES 112th Convention, Munich.

[2] Emmett, J. (2003) "Audio levels - in the new world of digital systems", EBU Technical Review, vol.2003:January.

[3] Spikofski, G. & Klar, S. (2004) "Levelling and Loudness - in radio and television broadcasting", EBU Technical Review, vol.2004:Jan.

[4] Moerman, J.P. (2004) "Loudness in TV Sound", in Proc. of the AES 116th Conv.

[5] Lund, T. (2006) "Control of Loudness in Digital TV", in Proc. of the NAB-2006 Convention.

[6] Spikofski, G. (2006) "Who can do louder? Loudness jumps in broadcasting and how to avoid them?", in Proc.of the 24th Tonmeistertagung International Convention, Leipzig..

[7] Skovenborg, E. & Lund, T. (2008) "Loudness Descriptors to Characterize Programs and Music Tracks", in Proc. of the AES 125th Convention, San Francisco.

[8] ITU-R (2006) "Rec. ITU-R BS.1770, Algorithms to measure audio programme loudness and true-peak audio level.", International Telecommunications Union.

[9] EBU (2009) "EBU Project Group P/LOUD (Loudness for broadcasting)", Internet web site: http://wiki.ebu.ch/loud/Main_Page.

[10] Soulodre, G.A., Lavoie, M.C. & Norcross, S.G. (2003) "The Subjective Loudness of Typical

Program Material", in Proc. of the AES 115th Convention.

[11] Coen, J. & Coen, E. (2007) "No Country for Old Men", Miramax Films.

[12] Vinton, M.S. et al. (2004) "Controlling loudness of speech in signals that contain speech and other types of audio material", Patent number US 7,454,331, Dolby Laboratories Licensing Corporation.

[13] Riedmiller, J.C., Lyman, S. & Robinson, C. (2003) "Intelligent Program Loudness Measurement and Control: What Satisfies Listeners?", in Proc. AES 115th Conv., New York.

[14] Grimm, E.M., van Everdingen, R. & Schöpping, M.J.L.C. (2008) "Towards a Recommendation for a European Standard of Peak and LKFS Loudness Levels", in Proc. of the IBC 2008.

[15] ITU-R (2007) "Rec. ITU-R BS.1770-1, Algorithms to measure audio programme loudness and true-peak audio level.", International Telecommunications Union.

[16] Skovenborg, E. & Lund, T. (2009) "Loudness Descriptors to Characterize Wide Loudness-Range Material", in Proc. of the AES 127th Convention, New York.

[17] ITU-R (2006) "Rec. ITU-R BS.1771, Requirements for loudness and true-peak indicating meters", International Telecommunications Union.

[18] Camerer, F. (2010) "Loudness - an Audio Levelling Revolution. Report of the scope and work of the EBU-group P/LOUD", EBU.

[19] Bech, S. & Zacharov, N. (2006) "Perceptual Audio Evaluation - Theory, Method and Application", Wiley.

[20] Skovenborg, E., Quesnel, R. & Nielsen, S.H. (2004) "Loudness Assessment of Music and Speech", in Proc. of the AES 116th Convention, Berlin.

[21] EBU (1998) "Listening conditions for the assessment of sound programme material: monophonic and two–channel stereophonic", EBU Tech. 3276 - 2nd ed., EBU.

[22] Siegel, S. & Castellan, N.J. (1988) "Nonparametric statistics for the behavioral sciences" (2nd ed.), McGraw-Hill.

[23] Conover, W.J. (1999) "Practical nonparametric statistics" (3rd ed.), Wiley.