Audio Engineering Society

Convention Paper

Presented at the 127th Convention
2009 October 9–12 New York, NY, USA

Loudness Descriptors to Characterize Wide Loudness-Range Material

Esben Skovenborg[1] and Thomas Lund[2]

[1] TC Electronic A/S, Risskov, DK-8240, Denmark
EsbenS@tcelectronic.com

[2] TC Electronic A/S, Risskov, DK-8240, Denmark
ThomasL@tcelectronic.com

ABSTRACT

Previously we introduced the concept of loudness descriptors – key numbers to summarize loudness properties of a broadcast programme or music track. This paper presents the descriptors: Foreground Loudness, which measures the level of foreground sound, and Loudness Range, which quantifies the variation in loudness. Wide loudness-range material may benefit from level-alignment based on foreground loudness rather than overall loudness. We propose to use these descriptors for loudness profiling and alignment, especially when live, raw, and film material is combined with other broadcast programs, thereby minimizing level-jumps and also applying appropriate dynamics-processing. Finally, we introduce the "zap test" which can objectively characterize the quality of loudness-balancing schemes, based on statistics of the potential loudness-jumps. A study of *loudness-jump tolerance* is presented, considering loudness-jumps from a subjective point of view.

1. INTRODUCTION

The traditional way of controlling level in pro audio is responsible for unacceptable level jumps in digital television, for music CDs getting increasingly distorted, and for different audio formats and programme genres being incompatible. The most fundamental audio issue of all – control of loudness – makes people adjust the volume control over and over again. Level control procedures that focus solely on peak level or dialogue level have proven inadequate. There is a wish from broadcasters, as well as from film and music mastering professionals, for identifying a universal, efficient solution [1, 2, 3, 4, 5].

In the 'justification' section of the BS.1770 standard [6], the *ITU-R considers* "*…that modern digital sound transmission techniques offer an extremely wide dynamic range; … that listeners desire the subjective loudness of audio programmes to be uniform for different sources and programme types; that many methods are available for measurement of audio levels*

but that existing measurement methods employed in programme production do not provide indication of subjective loudness; ..."

This paper describes how loudness descriptors, based on this ITU-R standard, could be used to improve on the situation. Extending on a previous study [7], this paper focuses on the situation where wide loudness-range (WLR) programs, such as film or drama, are interspersed with narrow loudness-range programs, such as commercials or pop music.

1.1. Loudness Balancing of Programmes

Many different approaches have been followed in order to balance programs or tracks in loudness:

- **Listening**; the audio engineer simply listens through (major parts of) the programme; for this approach calibrated monitoring is recommended [8]; is time-consuming, hence impractical.

- **VU or PPM metering**; the VU meter has been used for adjusting and controlling the level of analog audio for decades, and the PPM meter might be considered its digital heir; the loudness of a programme could to some extent be inferred based on VU or PPM metering, although this would require an audio engineer experienced in the task, as well as prior knowledge of the genre of material; the display range of the VU meter is inadequate for measuring wide dynamic range material.

- **Loudness metering, spot check**; one or more representative "anchor points" are selected manually; the loudness of the anchors is then assessed using loudness metering [9] or simply listening; the anchors might consist of speech, but other types could be equally good or better, for example the 'outro' music of a film is supposed to be a quite good indicator of the films overall loudness.

- **Dialogue normalization (dialnorm metadata)**; storing the Leq(A) of 'normal' speech as metadata [10]; automatic gain-correction to align dialogue, at decoder/playback; however, not all material contains "dialogue"; requires either upstream manual setting of dialnorm, or automatic detection (Dolby patent-protected algorithm); metadata may be incorrect, even deliberately abused; may be lost due to infrastructure or format-conversion [11].

- **Peak-level normalization**; normalizing the sample peak-level of the track – often close to 0 dBFS – is a practice common for music CD mastering [12]; this approach, typically used in combination with limiting and/or clipping, is intended to maximize – *not* balance – the loudness; it is an effect of the ongoing "loudness war" and tends to yield poor audio quality [13, 14].

- **Dynamics processing**; although overdoing dynamics compression could make it easier to loudness-align different programs, the audio quality could suffer; over-compression is often used as a means of maximizing rather than balancing loudness; generally some amount of dynamics processing is relevant for WLR material, but the appropriate processing would depend on the programme content as well as on the intended media [8].

- **Loudness descriptors**; a few key numbers characterize a full programme, regardless of the programme's duration; automatic computation; compatible with 'live' (real-time) as well as server-based workflows; (details follow in the next section).

2. LOUDNESS DESCRIPTORS

Loudness descriptors are key numbers to summarize loudness properties of an audio segment, broadcast program or music track. In our previous paper [7], we introduced the Center of Gravity (CoG), which measures the overall loudness of an audio segment, and the Consistency descriptor, which measures the variation of the loudness. In this paper we present two variants of the two original loudness descriptors: Foreground Loudness, which measures the loudness of foreground sound, and Loudness Range, which is a replacement for Consistency. The original term, "Consistency" was considered by EBU P/LOUD to be overly positive, thereby guiding the mixing engineer in the direction of less loudness range. For completeness, we repeat the description of the CoG, below.

The computation of the loudness descriptors is **based entirely on the output of a loudness measurement algorithm**, such as the baseline method specified in BS.1770, which is defined for both mono, stereo, and 5.1-channel input signals [6]. The loudness descriptors make no assumptions about the *content* of the audio segments; they are accurate yet robust enough to be universally applicable. Thus, the loudness descriptors

are **independent of the sample rate, format, and content** of the input signal.

2.1. Center of Gravity

The Center of Gravity (CoG) descriptor measures the **overall loudness of the segment**. That is, if one segment should be aligned in loudness with another using only a gain offset, that offset would be the difference between the CoG values of the two segments.

The CoG is essentially an integrating loudness measurement – like the Leq-type measurement specified in BS.1770. However, rather than blindly integrating over all input-samples, the CoG employs an *adaptive gate*. The adaptive gate enables the CoG to be robust against 'silence' in the program without making any rigid assumptions about the *absolute levels* of the material to be measured. By employing a gate, the measurement essentially ignores regions of the material which are too quiet to be considered part of the programme – regions that would otherwise have *biased* the CoG measurement.

In our development of the descriptors, we found that it is impossible to find one single fixed gate threshold that would work well with different types and genres of program material. Instead, we are using an adaptive gate which employs a *relative* gating threshold. The adaptive gate is based on a feed-forward structure. Thus the adaptive gate can never "get stuck" at too high a threshold, even if the measurement begins with a region of very high loudness. We found that using a relative threshold of –20 dB, in the adaptive gate, yields good results with many types of material (see [7]).

The gate may furthermore employ a *look-ahead mechanism*, which allows the gate to close slightly before – and open slightly after – the input signal crosses the threshold level. The length of the look-ahead delay would correspond to the length of the analysis window or time constant used in the loudness measurement. Look-ahead would improve the accuracy when measuring short segments such as promos and commercials.

2.2. Foreground Loudness

Wide loudness-range (WLR) material may contain regions of relatively high loudness as well as regions which have considerably lower loudness but are still part of the programme (i.e. not background noise). We shall refer to these two components of WLR as foreground sound and background sound, respectively. WLR programmes may benefit from level-alignment based on the loudness of its foreground sound, rather than the overall loudness. The Foreground Loudness (FgL) descriptor measures this loudness level of a programme.

The FgL descriptor that we present here is analogous to the CoG, but with a considerably higher threshold for the adaptive measurement-gate. Thus, the **FgL measures the loudness level of the foreground sound** of the segment, while gating out passages of background sound. We have employed a relative threshold of –6 dB (or LU), in the adaptive gate, based on preliminary results of the EBU P/LOUD group [15]. That is, material with a loudness level lower than 6 dB below the overall (ungated) loudness level is ignored by the FgL. Note that whereas the CoG is based on the long-term loudness, the FgL is based on the short-term loudness, as defined in [9].

If loudness-aligning a segment having a high loudness range with another segment having a low range, using the CoG, passages of foreground sound in the former segment may appear unnaturally loud. Using the Foreground Loudness instead may lead to a more pleasing result. This corresponds to a strategy often used by audio engineers for aligning different programs in loudness: identify one or more characteristic anchor points in each programme, and then align those.

The 49 sound segments used as stimuli in the original ITU "SRG-3" loudness experiments [16] were reportedly all homogenous, i.e. edited such that each segment would consist only of – loudness-wise – similar material. The segments used in the loudness assessment experiments conducted by McGill and TC Electronic [17] were homogenous as well; these segments did have individual amounts of dynamics, but only little variation on the macroscopic time-scale. In other words, all these sounds were non-WLR – and for good reason: Due to the experimental methodology employed, it was imperative that the subjects could clearly identify a single loudness level for each sound segment. If a WLR segment had been included, some subjects might have assessed the loudness of the foreground sound, and others the overall loudness. It was the "common" component of the subjective loudness assessments that subsequently enabled Soulodre and us to use the loudness levels as reference data, against which different loudness models could be evaluated [18] [19].

As a consequence, it could be argued that the baseline loudness model, resulting from the SRG-3 tests (now recommended in the BS.1770) is only valid for homogenous sounds, i.e. non-WLR material. This background is one reason why there is a need for a CoG and/or FgL "on top of" the BS.1770 measurement. On the other hand, the measured CoG, the FgL, and the 'raw' integrated loudness will all be equal, for material with a sufficiently low loudness range and with no regions of silence.

2.3. Loudness Range

The Loudness Range (LR) is a descriptor which can quantify the variation in a time-varying loudness measurement. The **Loudness Range measures the variation of the loudness on a macroscopic timescale**, in units of LU (i.e. on a dB scale). The Consistency descriptor, which we introduced previously, is basically defined as minus half of the estimated Loudness Range – that is, the two are equivalent. However, as audio professionals seem to prefer the term Loudness Range over Consistency, we shall use the former henceforth (although LR could still be said to be a measure of loudness consistency).

In order to achieve a good compromise between precision and robustness, the measurement of Loudness Range is based on the *statistical distribution* of measured loudness. Thus, a short but very loud event would not affect the Loudness Range of a longer segment, and similarly the *fade-out* at the end of a music track would not increase the measured Loudness Range noticeably. Specifically, the *range* of the distribution of loudness levels is determined by estimating the difference between a low and a high percentile of the distribution. This method is similar to the Interquartile Range (IQR), used in the field of descriptive statistics to obtain a robust estimate of the spread of a data sample. Percentiles belong to *non-parametric statistics* and are employed in the computation of Loudness Range because the loudness levels can in general not be assumed to belong to a particular statistical distribution. This definition of Loudness Range turns out to work well for many different genres of material.

Like the CoG, the Loudness Range furthermore employs an **adaptive measurement-gate**. Certain types of material may have regions of very low loudness level (e.g. background noise), while otherwise being fairly even in loudness, i.e. overall consistent. If the Loudness Range had *not* used the adaptive gate, such material would (incorrectly) get a quite *high* Loudness Range measurement, due to the relatively large difference in loudness between the regions of background noise and those of normal (foreground) loudness.

The computation of LR is easily within the capabilities of today's digital processors. Nevertheless, let's consider an alternative (more naïve) definition of: *LR = highest_loudness – lowest_loudness*. This definition suffers from several problems: Suppose the *highest_loudness* corresponded to a single gunshot which was louder than anything else in a movie. Wouldn't it then be better if the *highest_loudness*, used to compute the overall *LR*, was a little lower in this case – that is, more "typical highest"? Is the *lowest_loudness* just the noise-floor of the material? Or wouldn't it be more useful if *lowest_loudness* corresponded to the lowest loudness *above* the noise floor? The real definition of LR, based on the distribution of loudness (as described above), tackles these issues.

Applying a loudness-correction (ALC) processor would typically *decrease* the Loudness Range of the material. Note that Loudness Range descriptor should not be confused with measures of dynamic range or crest factor etc. As an optional supplement to the Loudness Range, another loudness descriptor (being finalized) measures the variation of loudness on a *microscopic* timescale, corresponding to the amount of dynamic compression applied.

2.4. Scales and Ranges

The loudness descriptors presented here are based on the ITU-R **BS.1770** and **BS.1771** standards [6, 20]. Unfortunately, the use of terms and units in BS.1770-1 is not in agreement with common practice in acoustics such as specified by ISO and IEC international standards, and moreover the BS.1770-1 and the BS.1771 are not consistent with each other. This may cause confusion and misunderstandings. In [21] we outline the problems and propose a solution, in the hope that ITU-R will remedy these issues.

In this paper (being pragmatic) we use the **LKFS** unit to denote loudness level measurements based on the BS.1770 standard, and calibrated w.r.t. absolute level as specified. The unit **LU** (Loudness Units) is employed rather than LKFS for measurements of *relative* loudness level, such as a distance or range on the LKFS scale. In contrast to the phon- and sone-scales, used to measure

loudness (level) within psychoacoustics [22], the LKFS and LU units are measures of level on a dB-scale. This choice was presumably made by the ITU in order to make the LU-measurements more operational in the context of audio engineering. Hence, increasing the level of a signal by *X* dB will increase the corresponding loudness measurement by *X* LU.

Table 1 shows some properties of the Center of Gravity, Foreground Loudness, and Loudness Range, using the BS.1770 loudness measurement algorithm as pre-processor. The test-signals, referred to in the table, are composed as follows:

- Sine 1 kHz @ -20: 1 kHz sine wave, mono, -20 dBFS peak level.
- Sine 1 kHz @ -20, -40: 1 kHz sine wave, mono, -20 dBFS peak level, followed by another 1 kHz sine wave, mono, -40 dBFS peak level, of equal duration.

Note that the LR is unaffected by a constant gain change, whereas the CoG and FgL are "unity gain". Note also how the measured FgL is the very same for the two test tone signals: lowering 'half' of the signal by 20 dB (i.e. making it background sound) does not affect the FgL.

2.5. Standardization and EBU P/LOUD

While the BS.1770 standard specifies a baseline method for measuring an integrated loudness level [6], it does not address several of the challenges in the application of loudness-control, outlined above. However, it may be in the interest of everyone in the production chain to agree on a common and open definition of certain fundamental descriptors, for instance. TC Electronic is prepared to contribute to such an initiative.

For the purpose of recommending to its members an effective and general solution to the continuing loudness-problems, the EBU has formed the P/LOUD group [15]. While P/LOUD aims at basing its recommendation(s) on the BS.1770, its ongoing standardization efforts include a 'target' loudness level as well as the open definition of one or more loudness descriptors. TC Electronic is collaborating with P/LOUD to that effect.

Even though all parties are keen on getting a target loudness level standardized, it would be a mistake to recommend a specific number *before* it is clear what would be the best method to measure the loudness level to be aligned; for instance, if FgL was used, the target loudness level would be a different number, than if CoG had been used.

3. LOUDNESS DESCRIPTORS FOR WLR MATERIAL

Some programs contain regions of relatively high loudness as well as regions of relatively low loudness; we shall denote this wide loudness-range material (WLR). A programme with a LoudnessRange greater than, say, 15 LU might be called WLR, but there is no hard threshold. Using the terms from the previous section, in WLR there is a considerable difference in the loudness of background and foreground sound. In the following, we examine three characteristic examples having wide, medium, and narrow loudness range, respectively.

3.1. Three Characteristic Examples

The movie The Matrix (1999) is an extreme, yet authentic, case of WLR. The loudness of the movie's entire AC-3 audio track in 5.1 was analyzed. The raw

	CoG	***FgL***	***LR***
Maximum value	12 LKFS	12 LKFS	40 LU
Minimum value	−80 LKFS	−80 LKFS	0 LU
Very high value, in practice	−5 LKFS	−5 LKFS	25 LU
Very low value, in practice	−30 LKFS	−30 LKFS	1 LU
Sine 1 kHz @ -20	−23.01 LKFS	−23.01 LKFS	0 LU
Sine 1 kHz @ -20, -40	−25.98 LKFS	−23.01 LKFS	20 LU
+ X dB gain	+ X LKFS	+ X LKFS	unchanged
Segment is repeated	unchanged	unchanged	unchanged
Appending low-level signal	unchanged	unchanged	unchanged

Table 1. Properties of the loudness descriptors (some values are approximate).

AC-3 data not having left the digital domain was used, and DRC was disabled, so that the dynamic range of the audio was not reduced prior to our analysis. The LoudnessRange of The Matrix is 25.0 LU, which would probably be a challenge to the average consumer TV setup. In [7] we found that DVD movies generally have a LR (i.e. -2*Consistency) in the range 14 to 24 LU, and a CoG of around -26 to -21 LKFS.

The Matrix DVD's audio has a CoG=-21.0 LKFS and FgL=-17.2 LKFS. As this material is WLR the foreground sound is louder than the overall sound, which is confirmed by the FgL being around 4 LU louder than the CoG. This difference will have a consequence for the loudness alignment of the movie (see next section).

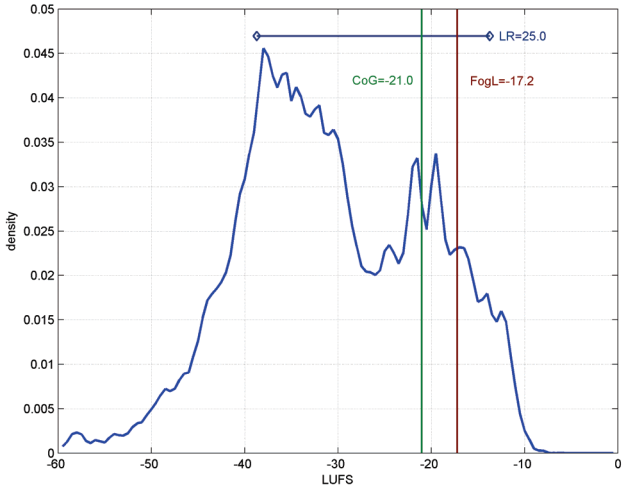

Figure 1. Movie "The Matrix", loudness-distribution with descriptors.

Figure 1 shows the statistical distribution of measured loudness, on which the descriptors are based, as well as indications of the descriptor values. Note that a relatively large proportion of the movie's sound has a loudness which is 10-20 LU below the Center of Gravity (which is unusual for an action movie). Regular dialogue occurs both in -44 to -30 LKFS interval and also between -25 to -18 LKFS (typically accompanied by some noisy action), so even though this is a movie, dialogue normalization would not seem to provide a simple answer to its loudness balancing.

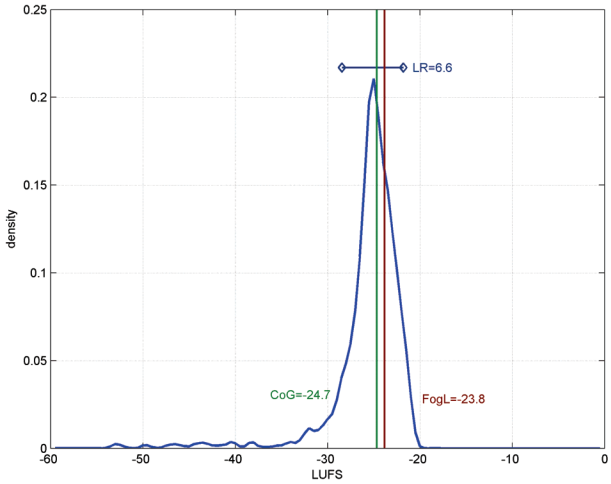

Figure 2. TV show "Friends", loudness-distribution with descriptors.

The TV show Friends (ep. 16), with a LoudnessRange of 6.6 LU, is shown in Figure 2. This is a typical example of a segment with a *medium* loudness-range. This DVD was a stereo production, and was transferred from the DVD player to digital using a peak-level of -6 dBFS, resulting in a CoG=-24.7 LKFS. Note that the LoudnessRange descriptor 'ignores' the brief parts with very low loudness at -30 to -50 LKFS), as we would expect (rather than allowing them to 'expand' the LR).

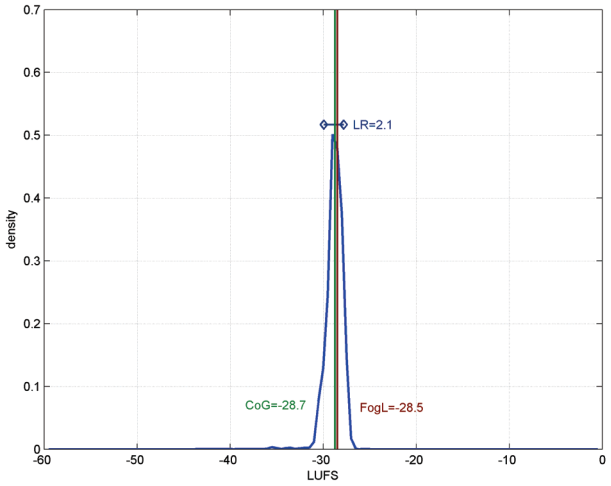

Figure 3. Even speech "Female Interviewer", loudness-distribution with descriptors.

At the other extreme, we have analyzed a segment consisting of female speech, very controlled and evenly spoken (Figure 3); the recording was used as the

reference segment in the original ITU "SRG-3" listening tests preceding the BS.1770. This segment has a narrow loudness range, LR = 2.1 LU. Note that, in this non-WLR case, the CoG and FgL are virtually equal, because the segment doesn't contain any regions which are more "foreground" than the rest.

The CoG of the 3 examples above were -21, -25, and -29 LKFS, respectively. In order to align the loudness of these programmes, simply apply gain to each programme, corresponding to the difference between the programme's CoG (or FgL) and the target loudness level. A recommended target loudness level has not yet been standardized – the jury is still out; however, the EBU P/LOUD group has been considering FgL = -21 LKFS as a candidate for a target loudness level, recommended for broadcast programmes. Assuming this target level, then the TV show analyzed above should be 2.8 dB louder, bringing the programme's peak level to -3.2 dBFS.

3.2. Towards an Objective Evaluation of Loudness Descriptors

Perhaps poor loudness alignment is most evident at the transition from one programme to another. Such transitions occur frequently while listening to one channel (inter-program transitions), or when zapping[1] (inter-channel transitions). In this context, and for either case, we propose a method which can quantify the quality of loudness-balancing schemes, based on statistics of the potential loudness-jumps: the "**zap test**".

Suppose we have two programmes, A and B, possibly of different duration, format, and genre. We could then randomly select some position in A and also some random position in B – let's call these positions cut-points. Now imagine an experiment where we would listen to segment A until the selected cut-point of A, at which time playback would immediately switch to segment B, starting at its cut-point. Most likely we would experience a jump in loudness at the transition. We could repeat the experiment (i.e. with new cut-points), which would probably yield a different loudness-jump. Now, instead of selecting just one cut-point in A and another in B, using computational statistics we can simulate this experiment with *all* possible pairs of cut-points in A and B, and then consider the statistics of all the resulting loudness-jumps, i.e. the loudness-jump distribution.

In this analysis we are using the long-term loudness, i.e. an integration time of 3 seconds, which we have found suitable for a time-varying yet robust loudness measurement [9]. If either one of the selected cut-points has a loudness level below the measurement-gate threshold (see the description of CoG), then that loudness-jump is excluded from the statistic; for instance, it doesn't make sense to compare the loudness of a silent region in a movie to that of another sound.

Note that we are only examining the *relative* loudness levels in this analysis, so the results are independent of what absolute target loudness level one might use. We simply analyze: if the CoG of programme A was aligned with the CoG of programme B, what loudness-jumps could we expect at transitions? (and similarly, aligning with FgL instead).

Table 2 provides some basic loudness-jump statistics, when transitioning between 3 programmes, having a wide loudness-range (The Matrix, movie), medium loudness-range (Friends, TV show), and narrow loudness-range (steady speech), respectively.

Let's first consider the transition between the narrow LR (Speech) and medium LR (Friends) programmes, corresponding to the top 4 rows of Table 2. The difference between the using CoG- and FgL-alignment is small, because no WLR material is involved; in both cases the median loudness-jump is less than 1 dB, i.e. barely noticeable. Changing channels from the TV show to (say) a news channel with even speech would 19 times out of 20 (=95%) result in a loudness-increase of less than 6 dB.

Zapping from the TV show to the WLR movie, as shown in the rows 7-8 of the table, yields an upwards jump in loudness of at most 8.6 dB (95% of the transitions) when using the CoG for level alignment; when using the FgL instead, the maximum upward jump is only 5.7 dB. This illustrates the difference between the CoG and the FgL in practice.

The transitions between the speech and The Matrix exhibits the same phenomena, only more exaggerated (and in the reverse direction). In particular, the upward loudness-jump, when zapping from The Matrix to a programme with even speech, is over 9-12 dB (50% of transitions), and over 19-21 dB (5% of transitions).

[1] *zap* [verb] To use a remote control to repeatedly change channel on a television

Such loudness-jumps could be problematic – especially if the viewer had turned the volume up, in order to properly hear the softer passages of The Matrix, before changing to a news channel. However, note that these large loudness-jumps are a consequence of changing between WLR and non-WLR material, regardless of the level-alignment strategy employed. In such cases, loudness-alignment of programmes needs to be combined with processing to match the loudness range with the media and playback situation, for instance to smoothen out the transition.

The zap test could – in a future study – be used to evaluate other loudness-alignment methods (than Cog or FgL); in particular, the method of manually identifying a representative "anchor" by listening to the WLR material (see sect. 1.1) could be compared with the automatic methods.

3.3. Listeners' Tolerance

The ratio between RMS level and peak level (often called "headroom") is an important property in live sound, music mixing, mastering and broadcast [8]. In [23] we have used the term **Dynamic Range Tolerance**, which quantifies the typical distance between RMS level and peak level that a consumer would tolerate inside a program or music track under different listening conditions, and may be used to define how much headroom is required for a given signal-path. It would be natural, in the context of loudness-balancing, to investigate how much shift in loudness a consumer would tolerate from one programme to another.

Truax [24] provides an interesting theory of radio programming, incorporating the listener's attention and the structure and dynamics of the programs and commercials. In [25](sect.3) Riedmiller et al. introduced the **Comfort Zone**, which is an empirically determined range in which TV programmes could vary, without viewers finding it unacceptable. Outside the Comfort Zone is a larger range, defined by when the viewers would feel the need to adjust the volume (loudness), and finally an even larger range outside of which the loudness would be rated as annoyingly soft/loud. It is assumed "*…that the non-speech elements of the programs have been appropriately balanced around the speech elements, [and therefore] listeners will not be annoyed by the natural changes in loudness that occur during programs if speech elements fall within their individual Comfort Zone.*", which seems to imply that the loudness ranges of the Comfort Zone only apply to speech material. Unfortunately, this dialogue normalization approach makes it difficult to compare with the loudness-jumps at programme-transitions, in our Table 2. Nevertheless, we note that all 3 ranges surrounding the Comfort Zone are asymmetrical, with the "too soft" limits generally being around twice the level of the "too loud" limits [25]. For instance, the "would turn volume down" is at +5.6 dB, whereas "would turn volume down" is at -10.2 dB.

The loudness-jumps occurring at the transition from one programme to another may often be what cause listeners/viewers to adjust the volume – regardless of whether the transition contains speech or not. Moreover, no evidence is provided in [25] to support its assumption that listeners will not be annoyed by

Programme A	*Programme B*	*Loudness alignment*	*Median loudness-jump (50%), A→B*	*Max loudness-increase (95%), A→B*	*Max loudness-decrease (5%), A→B*
Friends, TV show	Speech, even	FgL	0.8	6.3	-2.5
Friends, TV show	Speech, even	CoG	0.3	6.0	-3.1
Speech, even	Friends, TV show	FgL	-0.8	2.5	-6.3
Speech, even	Friends, TV show	CoG	-0.3	3.1	-6.0
The Matrix, 5.1	Friends, TV show	FgL	10.2	21.0	-5.7
The Matrix, 5.1	Friends, TV show	CoG	7.9	19.0	-8.6
Friends, TV show	The Matrix, 5.1	FgL	-10.2	5.7	-21.0
Friends, TV show	The Matrix, 5.1	CoG	-7.9	8.6	-19.0
The Matrix, 5.1	Speech, even	FgL	11.5	20.9	-4.0
The Matrix, 5.1	Speech, even	CoG	8.8	18.7	-7.4
Speech, even	The Matrix, 5.1	FgL	-11.5	4.0	-20.9
Speech, even	The Matrix, 5.1	CoG	-8.8	7.4	-18.7

Table 2. Loudness-jump statistics (in LU); programmes A and B are a pair of the 3 example sound segments analyzed above; the percentages refer to the distribution of potential loudness-jumps from programme A to B.

changes in non-speech loudness (cf. the quotation, above).

It is therefore relevant to study the amount of such loudness-jumps typically tolerated by a listener: the **loudness-jump tolerance (LJT)**. To investigate the loudness-jump tolerance an informal pilot study was carried out. In this test, 9 subjects submitted in total 220 ratings of inter-program loudness jumps between homogenous, non-WLR segments, under different listening conditions. Representing broadcast audio, news, hyper-compressed music, commercials and films, 8 sound segments were chosen, all with a LoudnessRange of less than 1.5 LU (i.e. very little loudness-variation). Using the BS.1770-based CoG descriptor, the segments were loudness-normalized, and the relative loudness of each segment could then be controlled using individual playback gain.

A subject was first asked to adjust the relative loudness of a Calibration segment to a comfortable listening level. Based on 'playlists' containing the segments with individual playback gain, the subject was then required to rate each programme-transition by a mark in one of five fields:

- No level objections,
- The level increased, but not enough that I would get the remote,
- The level increased so much that I would get the remote,
- The level decreased, but not enough that I would get the remote,
- The level decreased so much that I would get the remote.

Overall, subjects would want to turn *down* the level in more than 95% of the cases where an inter-program loudness jump created an *increase* of 5 LU or more (Table 3). Subjects would turn *up* the level in more than 95% of the cases where an inter-program loudness *decrease* of 8 LU or more occurred. In other words, subjects seemed more sensitive to a loudness increase than to a loudness decrease at programme transitions. This **asymmetrical loudness-jump tolerance**, also supported by [25], appears to be a fact which needs to be taken into consideration when applying loudness measurement and loudness descriptors in production and distribution.

	50% of ratings would adjust volume	*95% of ratings would adjust volume*
Loudness *increase* between programs	3 LU	5 LU
Loudness *decrease* between programs	6 LU	8 LU

Table 3. Inclination to reach for the volume control because of inter-programme level jumps.

Table 2 (previous section) showed that the expected loudness-increases at programme transitions would be less than 6 dB, for non-WLR programmes, using either CoG or FgL for loudness alignment. For WLR material, however, there exists no level alignment, based on a constant gain offset, that can prevent loudness-jumps which would almost certainly be un-tolerable – regardless of normalization strategy.

3.4. Robustness

To find out how much loudness descriptors change when audio undergoes bit-rate reduction, we conducted an investigation where the loudness descriptors were computed for audio segments in both linear PCM and perceptually coded (bit-rate reduced) versions: AAC@256 kbps, AAC@128 kbps, AAC@128 kbps (2 codec generations), MP3@128 kbps, and MP3@64 kbps. Both music tracks, movies, and typical broadcast programmes were included the test.

We found the **Loudness Range to be robust against the codecs, in all cases, changing less than +/-0.2 LU** (if at all). The Center of Gravity was generally quite robust as well, with deviations less than +/-0.2 LU, with a few exceptions: for hyper-compressed (pop/rock) music, for instance AC/DC's Rock'n Roll Train, the CoG decreased by up to 1.0 LU with cascaded codecs; that is, the audio would loose 1 LU of loudness through the codecs. The considerable amounts of distortion on hyper-compressed programs, in combination with high frequency roll-off when low bit-rate codecs are employed, is believed to play a role in this trend, though no further analysis has been performed.

This evaluation demonstrates that the loudness descriptors are robust against all the codecs tested, even at bit-rates lower than commonly employed in broadcast audio.

4. APPLICATIONS

Do consumers actually prefer the uniform sound of pop music, or the surprisingly loud scenes in a movie? Nobody knows for sure, but consumers dislike level jumps between regular programming and commercials so much that the issue has made it to the political agenda in several countries.

With BS.1770 based measurements and universal descriptors, pro audio is provided with a common way of looking at loudness and peak level that can be used across genres and across formats to improve on today's level chaos.

When universal descriptors and compatible metering are applied already in music, post or film **production**, the engineer/musician/journalist not only has the means to get the target loudness correct, she also is provided with a standardized way of evaluating the Loudness Range of a program. Where dynamics processing is required because the LR exceeds a certain number (e.g. 8 LU), the most surgical place to apply it is in production rather than downstream of the studio.

Universal descriptors are applicable in **delivery specifications** for all genres. When the studio knows the precise downstream expectations, the signal-path becomes transparent, and less time needs to be spent on subsequent correction of the two fundamental parameters, normalization and loudness range.

At a **broadcast station**, loudness descriptors may be used for optimizing programs automatically for various platforms, for instance to bring up loudness range processing on Mobile TV when the LR descriptor exceeds a certain number, e.g. 8 LU. Applied during ingest or inside a station server, optimum level offsets may be determined, and potential loudness-related problems diagnosed. In HDTV, when one or both programs at a transition point are WLR, transition processing may be invoked to prevent potentially annoying switchovers.

For platforms employing the **AC-3** codec, CoG or FgL could be used to set the "dialnorm" metadata parameter. The efficient solution, by which to easily cover multiple broadcast platforms, is to normalize audio at the station. With typical broadcast material and commercials, CoG or FgL is used to do so. With WLR material (identified objectively by LR being above a certain limit), spot checking a relevant anchor segment may possibly yield better results. Normalizing audio at the station has another advantage than helping multiple platforms: Dialnorm may be left statically at the same CoG/FgL value, thereby minimizing station workload and things that could go wrong.

If a TV station has chosen to rely on the AC-3 decoder for dynamics range processing in HDTV, the LR descriptor may be used to objective set a sensible DRC profile. It should be noted however, that AC-3 DRC processing is rudimentary and not based on Leq(K) weighting. In that sense DRC is not predictable if production uses BS.1770 based metering, and consequently cannot be part of a transparent signal-path spanning from production to delivery.

Post-transmission, results may be systematically logged, possible listener complaints diagnosed, and adjustments in procedures or target numbers carried out based on an informed foundation.

The Loudness Range descriptor may furthermore be used for **validating** a signal-path, because a measurement taken at any point downstream of production will yield the same result as a measurement taken at the mixing stage – provided the entire signal-path preserves the relative loudness levels and hence the Loudness Range. This aspect is important, as it allows a broadcaster to verify that metadata or other unpredictable factors don't distort the audio delivered to various brands of consumer TVs or STBs.

The CoG and FgL descriptors can be very effective in loudness-aligning programs. However, if a programme is WLR – which can be determined by its LoudnessRange – and that programme is to be balanced against another which is not WLR, then care must be taken. Depending on the context, dynamics processing may be the answer; reducing the LoudnessRange of the WLR programme will lead to a better loudness balancing. In other situations, such as certain types of classical music, an artistic judgment by an audio engineer may be a better alternative. In any case, the automatically measured LoudnessRange would reveal the situation.

5. CONCLUSION

Wide loudness-range material (WLR) poses a special challenge to incorporate seamlessly with other programming at a broadcast station and at the consumer. In this paper we have presented the **Foreground**

Loudness (FgL) descriptor which measures the loudness level of a programme's foreground sound. Intended for loudness-balancing of WLR material, as an alternative to the Center of Gravity (CoG) descriptor, the Foreground Loudness can reduce loudness-increases at transitions from typical broadcast programming to wide loudness-range programmes. The downside is a corresponding increase of loudness-jumps at transitions from WLR programmes to non-WLR material such as promos or commercials.

We also presented the **Loudness Range (LR)** descriptor (replacing the earlier Consistency descriptor), which quantifies the variation in loudness. A narrow, a medium, and a wide-loudness range programme were examined. These examples had very different loudness-distributions, and demonstrated the Loudness Range descriptor in action as an objective measure capable of identifying programmes around which annoying loudness jumps would be prone to occur. We furthermore introduced the "**zap test**", which is a novel method to objectively characterize the quality of loudness-balancing schemes, based on statistics of the potential loudness-jumps. The difference between loudness-alignment using the Center of Gravity and using Foreground Loudness was illustrated using the zap test. To provide further insight into **listeners' tolerance** of loudness-jumps at programme-transitions, a preliminary study indicated that the listener is more likely to reach for the level control when a subsequent programme is louder, than if it is softer.

Based on these analyses, we conclude that a mere gain offset at the broadcast station or at the consumer for each programme can *not* guarantee against annoying level jumps, if at least one of the programmes pre or post transition is wide loudness-range. We recommend using the **Loudness Range** descriptor, as a supplement to **CoG** or **FgL** as an effective means to detect WLR-related problems – *one number is not enough*.

6. ACKNOWLEDGEMENTS

The authors would like to thank Søren H. Nielsen for providing useful feedback to this manuscript.

7. REFERENCES

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

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

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

[4] 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.

[5] AES Staff Writer (2008) "If it's loud does that mean it's bad? - Two workshops on broadcast issues from the 123rd Convention", Journal of the Audio Engineering Society, vol.56:6, pp.493-498.

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

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

[8] Katz, B. (2002) "Mastering Audio: The Art and the Science", Oxford: Focal Press.

[9] Skovenborg, E. & Nielsen, S.H. (2007) "Real-Time Visualisation of Loudness Along Different Time Scales", in Proc. of the 10th Int. Conference on Digital Audio Effects (DAFx-07), Bordeaux, France.

[10] ATSC (2007) "A/53: ATSC Digital Television Standard, Part 5 – AC-3 Audio System Characteristics", ATSC Standard.

[11] Jacobs, B. (2008) "Dialnorm: A good idea gone bad?", Internet URL: http://broadcastengineering.com/audio/dialnorm_good_idea/, Broadcast Engineering.

[12] Lund, T. (2006) "Stop Counting Samples", in Proc. of the AES 121st Conv.

[13] Jones, S. (2005) "The Big Squeeze -- Mastering engineers debate music's loudness wars", Internet web page: http://mixonline.com/mag/audio_big_squeeze/, Mix.

[14] Nielsen, S.H. & Lund, T. (2003) "Overload in Signal Conversion", in Proc. of the 23rd AES Intl. Conf.

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

[16] Soulodre, G.A., Lavoie, M.C. & Norcross, S.G. (2003) "The Subjective Loudness of Typical Program Material", in Proc. of the AES 115th Convention.

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

[18] Soulodre, G.A. & Norcross, S.G. (2003) "Objective Measures of Loudness", in Proc. of the AES 115th Convention.

[19] Skovenborg, E. & Nielsen, S.H. (2004) "Evaluation of Different Loudness Models with Music and Speech Material", in Proc. of the AES 117th Convention, San Francisco.

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

[21] Nielsen, S.H. (2009) "Note on measurement units for loudness", Internet URL: http://www.tcelectronic.com/media/nielsen_loudness_units.pdf.

[22] Zwicker, E. & Fastl, H. (1999) "Psychacoustics: Facts and Models" (2nd. ed.), Springer Series in Information Sciences, 22, Berlin: Springer-Verlag.

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

[24] Truax, B. (2000) "The Electroacoustic Media: Audio Mediation", chap. 11 in Acoustic Communication, Ablex Publishing Corp.

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