Audio Engineering Society

Convention Paper

Presented at the 129th Convention
2010 November 4–7 San Francisco, CA, USA

Measuring Dynamics: Comparing and Contrasting Algorithms for the Computation of Dynamic Range

Jon Boley[1], Michael Lester[2], and Christopher Danner[3]

[1] LSB Audio LLC, Lafayette, IN 47905, USA
jon@lsbaudio.com

[2] Shure Incorporated, Niles, IL 60714, USA
lester_michael@shure.com

[3] University of Miami, Coral Gables, FL 33124, USA
c.danner@umiami.edu

ABSTRACT

There is a consensus among many in the audio industry that recorded music has grown increasingly compressed over the past few decades. Some industry professionals are concerned that this compression often results in poor audio quality with little dynamic range. Although some algorithms have been proposed for calculating dynamic range, we have not been able to find any studies suggesting that any of these metrics accurately represent any perceptual dimension of the measured sound. In this paper, we review the various proposed algorithms and compare their results with the results of a listening test. We show that none of the tested metrics accurately predict the perceived dynamic range of a musical track, but we identify some potential directions for future work.

1. BACKGROUND

Dynamic range compression is often used for technical and artistic effect, but many audio professionals think that this family of technologies is often abused, thus degrading the quality of the final recording. In fact, many audio engineers argue that the dynamic range of music has steadily decreased (on average) over the past several decades - a trend known as "the loudness war" since the overall level can be increased. However, it is difficult to quantify this trend because there is no standard definition for measuring the dynamic range of a musical program. Several metrics have recently been proposed, yet there is little evidence to suggest that any of these quantitative techniques are accurate by any meaningful measure.

In this study, we examine a number of dynamics descriptors, the first of which is the recently published EBU definition of loudness range [1]. Loudness Range (LRA) is calculated by measuring the ITU-R BS.1770 [2] loudness within a limited time window (3sec) and building a histogram of these values. The LRA is defined as the difference between the 10th percentile and the 95th percentile on the histogram.

The Pleasurize Music Foundation [3] has a software package available for download which calculates a different form of dynamic range. This algorithm calculates the ratio of the peak to the RMS, but limits the RMS to those values which occur in the top 20% of the histogram. The authors argue that only using the top 20% allows them to compare a variety of program types (e.g., genres of music, speech, etc.), and is more likely to measure the effects of dynamic range compression since gain reduction is usually greatest for high-level signals.

Vickers [4] proposed a measurement he called *dynamic spread*, which is simply the p-norm of the signal. (Vickers recommends p=1, such that the dynamic spread is just the mean absolute deviation of the signal.) He points out that this could be estimated by a histogram, but does not recommend any particular method.

Tollerton [5] has proposed an algorithm he calls *pfpf*, which is somewhat similar to the EBU recommendation. However, this algorithm calculates BS.1770 loudness on 3 different time scales (10ms, 200ms, 3sec) and defines short, medium, and long-term dynamic range respectively as the range from the 50th percentile to the 97.7th percentile.

Unfortunately, none of the authors of these proposed metrics have offered any evidence to suggest that their chosen parameters (e.g., the time scales or the percentile ranges, or p in the *dynamic spread* algorithm) are the best choices. Our contention that a measure of dynamic range as a measurement of the listening experience (as opposed to a technical requirement) should be correlated to the actual perception of the dynamic range.

The following section outlines the details of a small listening test and the methods by which we compared the perceptual results to the various measurements. We then present the results of these analyses, followed by a short discussion and recommendations for future work.

2. METHODS

2.1. Stimuli

Ten music clips, each 10-20sec in length, were chosen from a variety of genres to represent small, medium, and large dynamic ranges. We specifically picked clips for which the objective metrics, with their default settings, produced widely varying results (that is, they did not agree with each other).

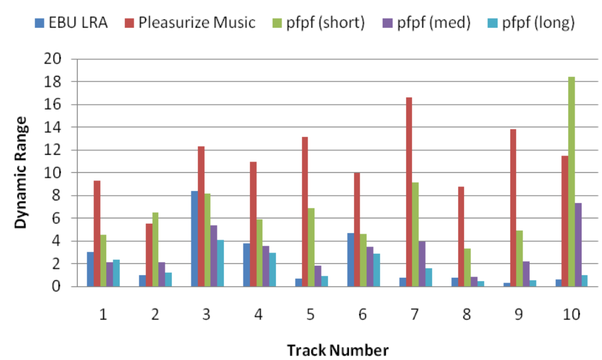

Figure 1. Dynamic / Loudness Range measurements for the 10 music clips

Figure 1 shows some of the measured values of loudness/dynamic range for each of the 10 music clips. The units are LU for all except the Pleasurize Music algorithm, which gives a value in dB.

2.2. Listening Test

A psychophysical experiment was performed to determine the relative perceived dynamic ranges of the musical tracks. A 2-alternative forced choice paradigm (using the method of constant stimuli) presented each listener with two musical clips and asked the question, "Which clip has the greater dynamic range?".

Sixteen university students studying music engineering served as the subjects for this experiment. For the ten music clips chosen, we presented each listener with all 45 pairs.

The responses are assumed to fall on a monotonically increasing psychometric function such that the number of "greater" responses is proportional to the actual dynamic range. In other words, as we increase the dynamic range of one track relative to another, we expect the listeners to identify the difference more often. This is illustrated in Figure 1.

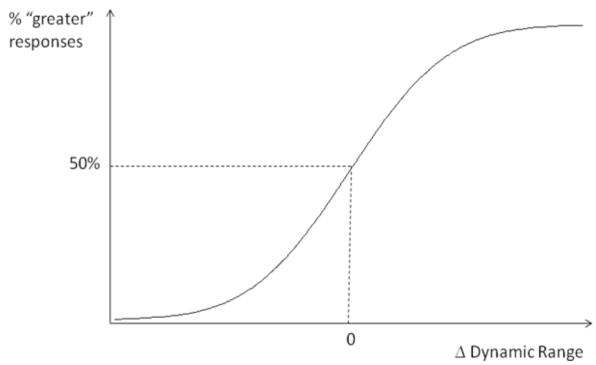

Figure 1. An illustration of the psychometric function assumed to underlie the data

2.3. Comparing Perception to Objective Metrics

For each algorithm, we explored the parameter space to determine the optimal settings for predicting the perceptual results. In particular, we evaluated the BS.1770-based metrics in fine detail. The BS.1770 loudness was calculated over windows of 100ms, 200ms, 400ms, 800ms, 1.5sec, and 3sec. The loudness range was then calculated for every interval on the histogram in steps of 5%.

Because the perceptual results are based on paired comparisons, we calculated the difference in dynamic range metrics for each pair of files. These difference values can then be compared to the results of the perceptual experiment.

3. RESULTS

3.1. Listening Test

The psychoacoustic results for all the paired comparisons are shown in Figure 2. Because we are not assuming any particular measure of dynamic range as the independent variable, they are plotted in sorted order, from lowest to greatest value. Note that these points are assumed to lie on the center portion of a psychometric function similar to the one illustrated in Figure 1.

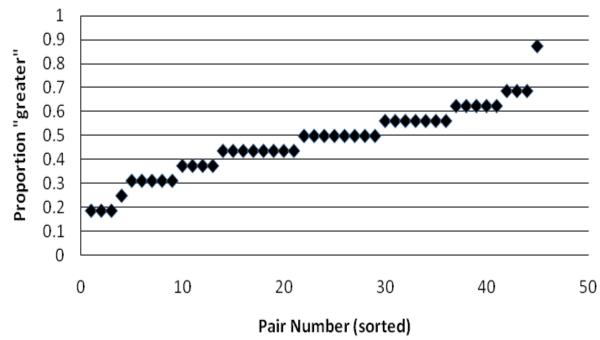

Figure 2. Sorted listening test results

3.2. Objective-Subjective Correlations

3.2.1. EBU Loudness Range

Using the recommended definition of loudness range (3sec window and 10%-95% range), we calculated ΔLRA for each pair of fibers. These values are plotted against the perceptual scores in Figure 3.

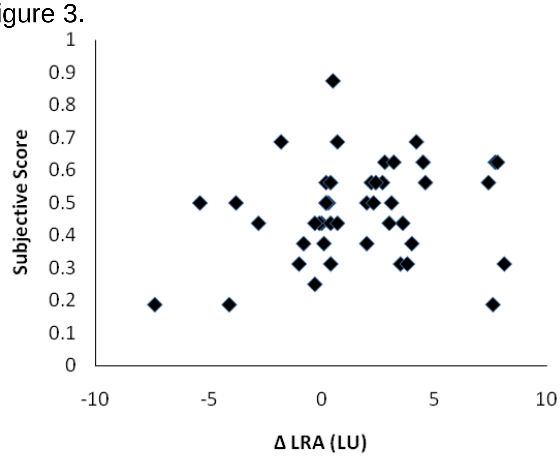

Figure 3. EBU Loudness Range difference values plotted against perceptual scores

The correlation of loudness range and the psychoacoustic data was found to be minimal (Pearson correlation, r = 0.21).

3.2.2. Pleasurize Music

The dynamic range measurement algorithm supported by the Pleasurize Music Foundation is compared to the listening test results in Figure 4.

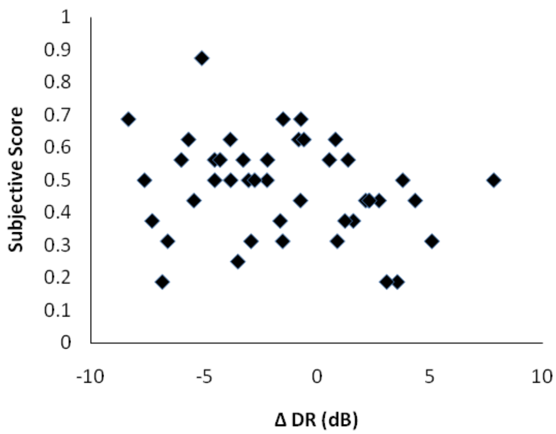

Figure 4. Pleasurize Music difference values plotted against perceptual scores

The correlation coefficient for the Pleasurize Music algorithm was also small (r = -0.18), suggesting that this measure does not accurately reflect the perceived dynamic range.

3.2.3. Dynamic Spread

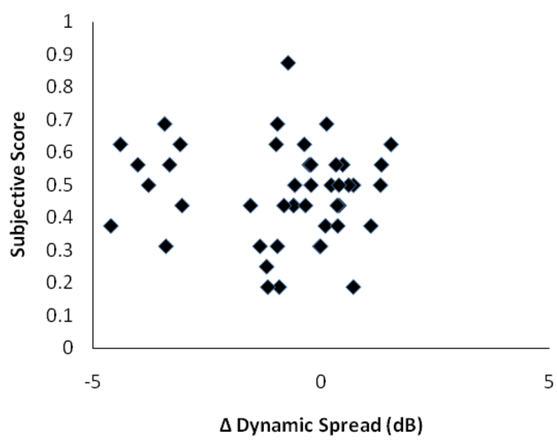

Figure 5. Dynamic Spread (mean absolute deviation) difference values plotted against perceptual scores

Results for the *dynamic spread* algorithm (simply the mean absolute deviation) are shown in Figure 5. The correlation coefficient for this algorithm was small (r = -0.20), as were the coefficients for higher order moments (2nd-order: r = 0.03; 3rd-order: r = 0.11; 4th-order: r = -0.03).

3.2.4. pfpf

The results for Tollerton's *pfpf* algorithm are shown in Figure 6. Differences in dynamic range values using short (10ms), medium (200ms), and long (3sec) time scales are plotted against the perceptual data.

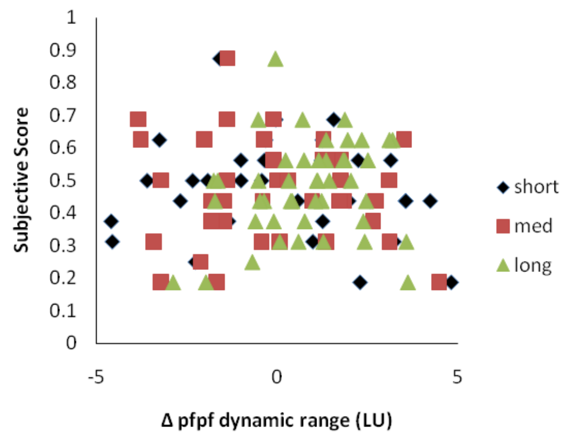

Figure 6. *pfpf* dynamic range difference values plotted against perceptual scores

The correlation coefficients for the *pfpf* algorithm were similarly small (short: r = -0.07; med: r = -0.23; long: r = -0.18). Although the larger correlation for the medium time scale suggests that this time scale (200ms) may reflect perception better than the other time scales, the correlation is still quite small and this metric only accounts for approximately 5% of the variance.

3.2.5. Other BS.1770 permutations

Based on BS.1770 loudness, we calculated the range at a number of different time scales and inter-percentile ranges. The top five correlations are shown in Table 1. All of these top five correlations are ranges of less than 1LU for all the files tested and can thus be considered irrelevant.

Rank	Correlation	Window	Percentile Range
1	0.409604	3000ms	30%-35%
2	0.399333	1500ms	35%-40%
3	0.341461	1500ms	35%-45%
4	0.307979	1500ms	35%-50%
5	0.305513	3000ms	30%-40%

Table 1. Top five correlations between loudness ranges and perceptual data

Table 2 shows the top five covariance values. Although the covariance numbers are small, it is interesting that they cover large ranges and correspond to the 400ms and 1.5sec window sizes.

Rank	Covariance	Window	Percentile Range
1	0.16575	400ms	0%-100%
2	0.155111	1500ms	0%-100%
3	0.154583	1500ms	0%-95%
4	0.154028	400ms	0%-95%
5	0.15275	1500ms	0%-90%

Table 2. Top five covariances between loudness ranges and perceptual data

4. DISCUSSION

Although we have not successfully found a sufficient algorithm for estimating perceived dynamic range, we have discovered a path that may warrant further exploration. The fact that the 400ms and 1500ms windows resulted in some of the highest covariance values indicates that these may be the time scales most directly related to dynamic range perception. Some combination of both a 400ms window and a 1500ms window may improve the correlation.

Future work should also include the collection of more psychoacoustic data. In particular, the dynamic ranges should be different enough that expert listeners can consistently differentiate between some of the files (expanding the "percent greater" to both 0 and 100%).

5. ACKNOWLEDGEMENTS

The gentlemen at the Communications Research Centre (CRC) Canada were kind enough to lend us their loudness meter for this research.

6. REFERENCES

[1] EBU Tech Doc 3342, "Loudness Range: A descriptor to supplement loudness normalisation in accordance with EBU R 128", Aug 2010.

[2] ITU-R BS.1770, "Algorithms to measure audio programme loudness and true-peak audio level", July 2006.

[3] http://www.pleasurizemusic.com

[4] Vickers, E. "Automatic Long-term Loudness and Dynamics Matching", in Proceedings of the 111th AES Convention, Nov. 2001.

[5] http://audiamorous.blogspot.com/2008/01/pfpf-experimental-estimator-of-dynamic_13.html