Audio Engineering Society
Convention e-Brief

Presented at the 133rd Convention
2012 October 26–29 San Francisco, CA, USA

Accuracy of ITU-R BS.1770 Algorithm in Evaluating Multi-track Material

Pestana, Pedro D.[1], Barbosa, Álvaro[1], and Reiss, Joshua D.[2]

[1] Research Center for Science and Technology of the Arts (CITAR). Portuguese Catholic University – School of the Arts. Rua Diogo Botelho 1327, 4169-005 Porto, Portugal
pedro.duarte.pestana@gmail.com abarbosa@porto.ucp.pt

[2] Centre for Digital Music, Queen Mary, University of London, London, E1 4NS, England
josh.reiss@elec.qmul.ac.uk

ABSTRACT

Loudness measurement that is computationally efficient and applicable on digital material disregarding listening level is a very important feature for automatic mixing. Recent work in broadcast specifications of loudness (ITU-R BS.1770) deserved broad acceptance and seems a likely candidate for extension to multi-track material, though the original design did not bear in mind this kind of development. Some empirical observations have suggested that certain types of individual source materials are not evaluated properly by the ITU's algorithm. In this paper a subjective test is presented that tries to shed some light on the subject.

1. LOUDNESS MEASUREMENT

Loudness is a subjective perceptual measure that is related to level, but not necessarily proportional to it, as there are many other factors upon which it depends. In an old, emblematic definition *"loudness is a psychological term used to describe the magnitude of an auditory sensation"* [1].

The physical factors influencing loudness – level [2], frequency content [1], transfer characteristics of equipment that will translate voltage to pressure, duration, direction, closeness [3], dynamic range compression [4] and reverberation [5] – are not the end of it, as *"the loudness produced will be the same for the same intensity only if the same or an equivalent ear is receiving the sound and also only if the listener is in the same psycho- logical and physiological conditions, with reference to fatigue, attention, alertness, etc."* [1]

Since Stevens [2] suggested the power relationship between loudness (L) and intensity (I) is compressive ($L=kI^{0.3}$, with k a constant), there have been numerous models trying to cope with the phenomenon. The most famous are those related to critical bands and specific loudness, after Zwicker [6], which have been perfected through the years (see [7] for a review) These are roughly based on the same basic sequential form and they are multi-band, computationally heavy models,

usually level-dependent, so that it is it difficult to use with pre-recorded material.

The American broadcaster CBS encouraged work at their laboratories in the 1960s that would, after a few improvements, develop into a model [8] that was a *de facto* reference for loudness measurement in broadcast well into the 21st century. Over the last decade there has been a significant amount of research on broadcast-related loudness perception and metering, a trend much inspired by the ITU-R efforts. The SRG-3 Special Rapporteur Group was created in the fall of 2000 to investigate into *"Audio metering characteristics suitable for use in digital sound production"* [9] This initiative eventually led to recommendation ITU-R BS.1770 [10], later extended on EBU R128 range of recommendations [11] [12].

Work on automatic mixing is fairly recent in its ADAFX [13] form, and a recent review [14] gives an updated state-of-the-art. Correct loudness measurement is an essential part of an automatic mixing process, and the existence of a recommendation that is becoming widespread is a blessing. Recent work [15] has already treated loudness of multi-track materials according to ITU-R BS.1770 / EBU R128 measurement recommendations with a good level of success. However, the authors have observed that this algorithm shows some consistent disagreements with perception through informal observations, and decided to create a test to understand how far it can be applied to the alternative task of individual sound source loudness judgment, since it was created for pre-mixed broadcast material. Our observations seem to indicate that it is the percussive material with limited high-range spectral bandwidth (i.e: hi-hats, shakers, tambourines) that is constantly under-evaluated by the algorithm, generating mixes in which these are too prominent.

2. SUBJECTIVE TESTING

2.1. Test Procedure

Tests were performed at Lusíada University's AudioLab and at an audio classroom at Restart Institute. 40 subjects used professional studio-grade headphones, with full-range frequency specifications, through the exact same audio chain, calibrated so that it delivered 78 dB SPL measured with a dummy-head. A very summary analysis will be given of the results.

2.2. 'Calibration' Test

A 'calibration'-type test was performed simultaneously to understand what would be a good measure of whether the EBU R128 recommendation resonated universally with human perception. Five mixed (unmastered) songs were chosen for loudness comparison. Subjects were presented with 10 five-fold matching comparisons, where the reference song was varied. The instructions were to match the loudness of the remaining four songs to that of the first song using a fader on screen and an x/or solo method. Unity level was not always at the same fader position, and subjects were alerted of that fact, and told not to mix visually. Some songs in some examples were duplicated, so that we could further test for consistency. We have been guided by the concerns and methodology suggested by [16] and particularly by the great care with which similar tests in [17] were elaborated.

Of the 40 subjects that took part on the test, 3 were professional sound engineers, 15 final-year students in audio, and 22 multimedia and music students with some (limited) exposure to audio engineering. The procedure was explained and the instructions given pre-test, and no one showed any doubt as to what was required. When averaged by each example instance (over the 40 subjects), means varied against the predictions of the algorithm by no more than ±2dB with standard deviations consistently under 1.5dB. When comparing the reference to a redundant copy, means were within ±0.15dB with standard deviations under 1dB, which seems to indicate that subject were accurate and attentive in performing the test. No separate group seemed to be more consistent than another, in that the standard deviation for pair-wise comparisons of the same two songs regardless of order were similar across subjects (0.8-1.7dB).

2.3. Main Test

The same type of test described above was modified to allow for evaluation of multi-track content. We now had the same five songs with individual tracks (each song had 9-11 different tracks). Subjects were given a fixed reference track and asked to alter the level of the remaining tracks until they sounded equally loud. It was emphasized that this was a loudness-matching task, since the subjects were used to performing to a different mindset in their profession/studies. Many subjects admitted after completing the test that it was very hard for them to keep their focus on the task.

Unlike the 'calibration' test, this one did not allow to use all tracks as reference — as it was, subjects already took on average 35 minutes to complete the test. Our fixed references were the kick drum and the vocals on alternate examples. Both elements were previously equalized so that they had similar spectral content across all five songs — this did not guarantee by itself that they would elicit equal loudness perception, but differences in answers from song to song were fairly low. Results are presented in figures 1 and 2, showing immediately that there is a strong bias in respect to which stimulus is presented as reference - when comparing against a vocal reference, subjects will on average agree with EBU R128 for the kick drum level needed for equal-loudness, whereas when evaluating the vocals against the kick drum reference, subjects will on average feel the vocals are 2 dB louder than what R128 predicts.

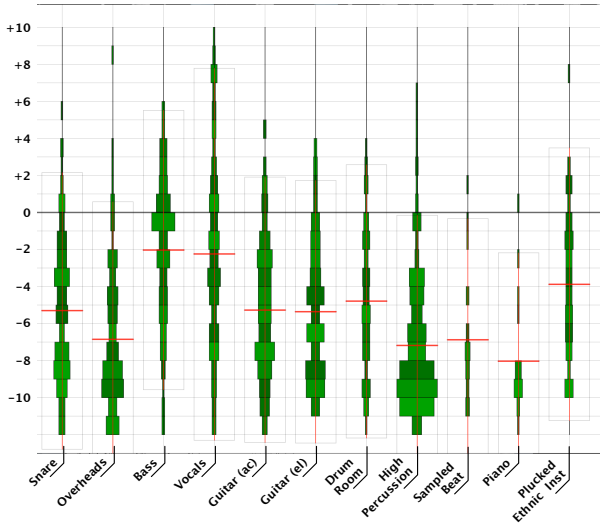

Figure 1. Results referenced to kick drum

We feel the centered histograms give a clear concise picture of what subjects evaluated. They show the extreme variance of the task, while the red horizontal lines indicate that the means were not at all in agreement with the algorithm. A measure of great agreement would be to have means around 0 dB, and in reality they are of by more than 4 dB when referenced to the vocals and more than 8 dB when referenced to the kick drum. The Wilcoxon test showed the null hypothesis (differences are insignificant) should be rejected, even though 95% confidence intervals almost always contain zero. Inter-subject variability is greater than in the calibration test. Standard deviations were consistently in the 3-4dB area when considering within-subject agreement or within example agreement, and in the 4-5dB area when considering overall aspects.

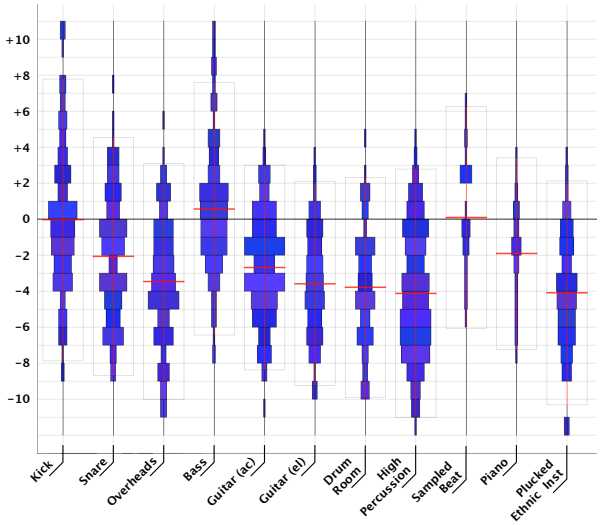

Figure 2. Results referenced to vocals

The figures show us a mirror of the algorithm's judgment, in that the lower an instrument group scores, the louder it would be overall, had a mix been done with equal-loudness purposes. We see that the vocals, kick and bass seem to be the elements that subjects need to have louder than the algorithm's prediction, while high percussions (hats, tambourines, shakers, etc.) and overheads are on the opposite side on the discrepancy spectrum. Some items show the great dependence on reference, such as the piano and the sampled beat, but are also the two categories with less observations. Two crucial aspects must be mentioned here: 1) the low end instruments (kick and bass) may not be rendered effectively by the monitoring apparatus, causing a misjudgment and 2) there might be a bias in evaluating vocals, as sound engineers are so used to placing them on top of the mix. Subjects may not be able to fulfill the intended equal loudness task.

If these are true, one notices that the apparently high variation of mean evaluation shown in the figures might actually not be greater than 4 dB (the ±2 dB indicated by the 'calibration' test, even though standard deviations are much higher).

2.4. Descriptor-based analysis

The fact that variation became larger when more data points were added may mean that the grouping together of tracks by their instrumental content cannot be done. It

is very plausible that we cannot lump together snare drums if their spectral and temporal profiles are different from song to song. This suggested it would be interesting to look for underlying features and see how they correlate to the mean choice of subjects for each isolated test that was performed.

A large array of low-level descriptors (loosely based in [18]) was tested against the mean data but the r^2 determination coefficient was only promising for the $\log_2$ Spectral Centroid and $\log_2$ Spectral Bandwidth. The r^2 values were 0.52 and 0.55 respectively, which does not suggest a strong variability explanation, but it does suggest a certain measure of dependency. The plot for the spectral centroid is shown in Figure 3. Note that both these features are redundant, in that there is an even higher correlation between both. The only interpretation we can get out of the figure is that the higher the log2 centroid, the more the algorithm under-evaluates a track's true perceived loudness, suggesting the RLB-filter should be low-passed or shelved.

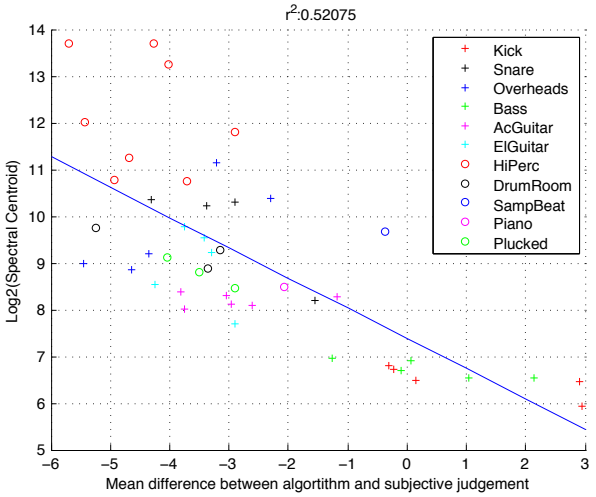

Figure 3. The influence of the spectral centroid.

We were surprised to find that a measure of spectral Q (centroid/bandwidth) and measures of temporal percussivity yielded no significative correlation, thus defeating our original observation that the high-Q transient elements were the most under-evaluated.

3. CONCLUSIONS

Psychometric tests are very prone to bias and though there was great care on the methodology, it can't be stated that the results are reliable (sample stratification is very localized, monitoring through headphones is complicated, reference tracks introduce a bias, etc.). It nevertheless seems fairly conclusive that there are trends in the degree of disagreement between subjective evaluation and algorithm, which seems a clear indication that some work should be pursued in extending the algorithm to multi-track material rather than just applying it.

4. ACKNOWLEDGEMENTS

This work was supported by Fundação da Ciência e Tecnologia, grant number SFRH/BD/65306/2009 and Fundação da FCUL.

5. REFERENCES

[1] Fletcher, H., & Munson, W. (1933). Loudness, Its Definition, Measurement and Calculation. *Journal of the ASA*, 5(October), 82–108.

[2] Stevens, S. (1957). On the Psychophysical Law. *Psychological Review*, 64, 153–181.

[3] Fastl, H., & Zwicker, E. (2007). *Psychoacoustics — Facts and Models*. Berlin: Springer, 3rd ed.

[4] Moore, B. C., Glasberg, B. R., & Stone, M. A. (2003). Why Are Commercials so Loud? Perception and Modeling of the Loudness of Amplitude-Compressed Speech. *Journal of the AES*, 51(12), 1123–1132.

[5] Skovenborg, E., & Nielsen, S. r. H. (2004). Evaluation of Different Loudness Models with Music and Speech Material. *Proceedings of the 117th AES Convention.*

[6] Zwicker, E. (1960). Ein Verfahren zur Berechnung der Lautstärke. *Acustica*, 10, 304–308.

[7] Moore, B. C. (2012). *An Introduction to the Psychology of Hearing*. Bingley, UK: Emerald Group, 6th ed.

[8] Jones, B. L., & Torick, E. L. (1982). A New Loudness Indicator for Use in Broadcasting. *Proceedings of the 71st AES Convention.*

[9] SRG-3 Status Report (2), September 2002 Document 6P/145-E.

[10] ITU-R-BS.1770: “Algorithms to Measure Audio Programme Loudness and True-peak Audio Level BS Series”.

[11] EBU (2010a). Tech Doc 3341 ”Loudness Metering: EBU Mode Metering to Supplement Loudness Normalisation in Accordance with EBU R 128”.

[12] EBU (2010b). Tech Doc 3342 ”Loudness Range: A Measure to Supplement Loudness Normalisation in Accordance with EBU R 128”.

[13] Verfaille, V., Arfib, D., Keiler, F., von dem Knesebeck, A., & Zölzer, U. (2011). Adaptive Digital Audio Effects. In U. Zölzer (Ed.) *DAFx*, (pp. 321–393). Chichester: John Wiley & Sons, 2nd ed.

[14] Reiss, J. D. (2011). Intelligent Systems for Mixing Multichannel Audio. *17th International Conference on Digital Signal Processing (DSP).*

[15] Mansbridge, S., Finn, S. and Reiss, J.D. (2012) Implementation and Evaluation of Autonomous Multi-track Fader Control. *Proceedings of the 132nd AES Convention.*

[16] Bech, S. and Zacharov, N. (2006). *Perceptual Audio Evaluation – Theory, Method and Application*. Chichester: John Wiley & Sons.

[17] Skovenborg, E., Quesnel, R. and Nielsen, H. (2004). Loudness Assessment of Music and Speech. *Proceedings of the 116th AES Convention.*

[18] ISO (2002). ISO/IEC 15938 Information technology – Multimedia content description interface – Part 4: Audio.