Audio Engineering Society

Convention Paper 8813

Presented at the 134th Convention
2013 May 4–7 Rome, Italy

Loudness Measurement of Multitrack Audio Content using Modifications of ITU-R BS.1770

Pedro Pestana[1,3], Josh Reiss[2], and Álvaro Barbosa[1]

[1] *CITAR - Catholic University of Oporto, Rua Diogo Botelho, 1327, 4169-005 Oporto*

[2] *C4DM - Queen Mary University of London, Mile End Road, London E1 4NS*

[3] *also member of ILID - Lusíada University of Portugal and CEAUL, Centro de Estatística e Aplicacoes - UL*

Correspondence should be addressed to Pedro Duarte Pestana (pedro.duarte.pestana@gmail.com)

ABSTRACT

The recent loudness measurement recommendations by the ITU and the EBU have gained widespread recognition in the broadcast community. The material it deals with is usually full-range mastered audio content, and its applicability to multitrack material is not yet clear. In the present work we investigate how well the evaluated perception of single track loudness agrees with the measured value as defined by ITU-R BS.1770. We analyze the underlying features that may be the cause for this disparity and propose some parameter alterations that might yield better results for multitrack material with minimal modification to their rating of broadcast content. The best parameter sets are then evaluated by a panel of experts in terms of how well they produce an equal-loudness multitrack mix, and are shown to be significantly more successful.

1. LOUDNESS MEASUREMENT

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 efforts. This initiative led to recommendation ITU-R BS.1770 [1], later extended by EBU R128 recommendation [2].

Recent work [3], [4] has already treated loudness of multitrack materials according to BS.1770 / R128 loudness measurement recommendations with some level of success. It is not clear how well it can be applied to the task of individual sound source loudness judgment, since it was created for pre-mixed broadcast material. The authors have observed that this algorithm shows some consistent disagreements with perception through informal observations, and described initial results in [5]. Our observations in-

dicate that the loudness of percussive material with limited high-range spectral bandwidth (i.e: hi-hats, shakers, tambourines) is often underestimated by the algorithm.

The loudness measurement recommendation we are investigating, outlined in [1], is a straightforward single band, level-independent system. The signal is passed through two biquad filters, termed the pre-filter (a +4 dB high shelf at around 1681 Hz) and the RLB-filter (a hi-pass filter with a 38 Hz cutoff), before being squared and its level measured over a time-constant of 400 ms. For a more thorough explanation the interested reader is referred to the ITU and EBU documentation [1, 2]. A signal that is measured according to this recommendation has a value given in Loudness Units (LU), which are a logarithmic unit, similar to the decibel.

In Section 2 we summarize the findings of the subjective listening test in [5] that proposed to reveal whether the discrepancy that was noted is indeed true for a diverse panel of individuals. We further explore the results under Section 3, looking for underlying features that might explain why certain types of single tracks are misjudged by the algorithm. In Section 4, some algorithm parameter tweaks are proposed that might provide better loudness measurement to a more diversified range of material. The effectiveness gain of the modifications is analyzed and, in Section 5, the most promising solutions are then evaluated by a panel of expert listeners.

2. SUBJECTIVE TESTING

The tests described in [5] were performed at Lusíada University's AudioLab and at an audio classroom at the Restart Institute in Lisbon. 40 subjects[1] used professional studio-grade headphones, with full-range frequency specifications, through the exact same audio chain, calibrated so that it delivered 83 dB_{SPL} measured with a dummy-head, a value that conforms to mixing recommendations (e.g. [6]), that suggest the listener should be at a medium equal loudness contour level.

There was a previous 'calibration'-type test aiming to understand what would be a good measure of whether the EBU R128 recommendation resonated universally with human perception. This test used broadband material (full mixes) and the results showed a strong consistency within subject and between subject and algorithm at the reference level, allowing the authors to proceed with the main question.

The main test aimed at the evaluation of multi-track content. We had five songs split into 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 (as loud as the reference track) using a set of faders. All the tracks had previously been normalized to yield the same loudness, according to the algorithm, so if a subject set up all the faders at unity[2], it would mean perfect agreement between measurement and evaluation. If subjects change the level past unit or below unity, then there would be a loudness evaluation difference which we calculate and present as our main variable.

It was emphasized that this was a loudness-matching task, given that 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 equal-loudness. 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 Bech and Zacharov [7], and particularly by the great care with which similar tests in Skovenborg et al [8] were elaborated.

The test design did not allow the subject to use all tracks as reference, or else the test duration would become unwieldy. Our fixed references were the kick drum (results shown in Fig. 1) and the vocals (results shown in Fig. 2) 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

[1] Three professional sound engineers, fifteen final-year students in audio, and twenty two 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.

[2] Unity here does not imply that the faders were marked and scaled the same, it is merely our own hidden unity reference. Unity level was not always at the same fader position, and subjects were alerted of that fact, and told not to mix visually.

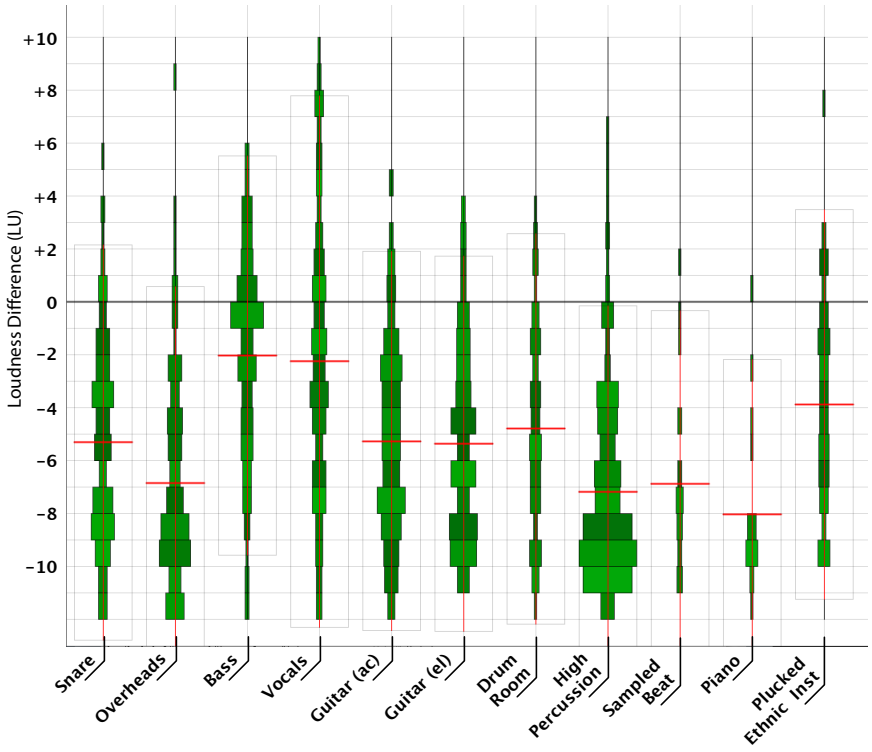

Fig. 1: Results referenced to kick drum. The continuous line establishes the kick drum LU value, and the vertical distributions are double-sided rotated histograms of how much the difference between evaluation and measurement for each track there is. The thick line around the middle of each histogram indicates the mean.

from song to song were fairly low. Reference choice is very critical and using an artificial test signal was discarded, as it was considered too ambiguous for comparison. Vocals were chosen as they are central in the listener's attention and the kick drum as it is often (in live mixing, for example) the first track that is dealt with, but these choices are far from equivalent. The results presented in figures 1 and 2 show immediately that there is a strong bias depending on which stimulus is presented as reference.

The centered histograms depicted for each instrument group give a clear concise picture of what subjects evaluated. The vertical spread shows the extreme variance of the task, while the thick horizontal lines indicate that the means were not at all in agreement with the algorithm. The vertical axis indicates by how many LU an actual subjective evaluation of the loudness of a track differs from the calculated loudness, that is:

$$D(t,s) = E(t,s) - M(t) \tag{1}$$

with $E(t,s)$ the calculated loudness of track t by subject s, and $M(t)$ the calculated loudness of the reference track. An indication of great agreement would be to have means around 0 LU. In reality they are off by more than 4 LU when referenced to the vocals and more than 8 LU when referenced to the kick drum. On average, subjects placed the vocals around -2 LU down from the kick drum in order to be judged as equally loud, and the piano around -8 LU down (see Fig. 1). This means that by balancing all the tracks in a song to be equally loud, with loudness defined according to the ITU recommendation, the vocal would be perceived as being 2 LU up from the kick drum and the piano 8 LU up.

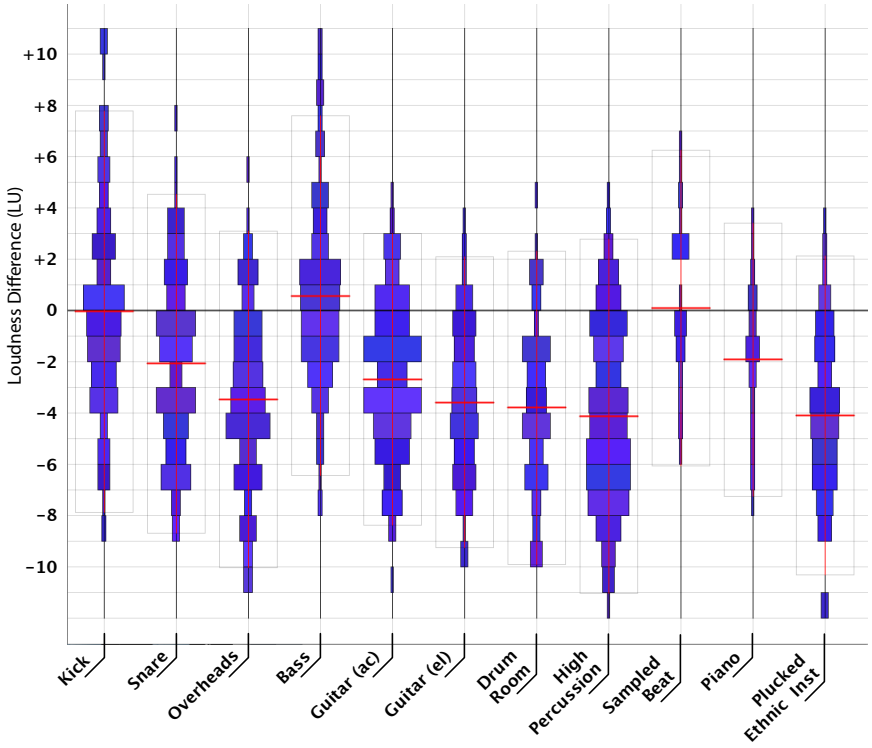

Fig. 2: Results referenced to vocals. Refer to the caption in Fig. 1 for explanation.

The high spreads shown could lead us to believe that even though the means show clear differences, the confidence intervals would overlap, leaving some room to ponder over whether the evaluation disagreements were due to random fluctuation. However, the Wilcoxon test (see [9]) showed that the null hypothesis (differences are insignificant) should be rejected, even though the 95% confidence intervals do in fact almost always contain zero. Inter-subject variability is greater than in the calibration test and the standard deviations were consistently in the 3–4 *LU* area when considering within-subject agreement or within example agreement, and in the 4–5 *LU* area when considering overall aspects.

Two potential problems are worth noting: as is clear by the width of the centered histograms, the piano and sampled beat only existed in one song, so their confidence intervals are much larger. Also the emphasis on bass and kick drum might be due to headphone frequency responses, and not to a deficiency in the algorithm, but we have run tests that seem to deny it.

3. 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 instrumentation is not meaningful. It is plausible that one 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 on [10]) was tested against the mean data, but the coefficient of determination (r^2) 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

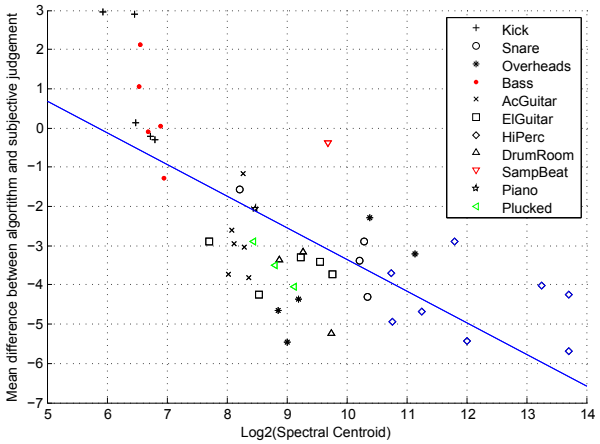

Fig. 3: The influence of the spectral centroid

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. It appears that the higher the log_2 centroid, the more the algorithm underestimates a track's true perceived loudness, suggesting either the algorithm's pre-filter or RLB-filter should show some additional high-end sensitivity[3].

We were surprised to find that a measure of spectral Q (centroid/bandwidth) and measures of temporal percussivity yielded no significant correlation, thus defeating our original observation that the high-Q transient elements were the most under-evaluated. Table 1 shows the features that were tested and their squared correlation coefficients. Whenever appropriate, log transformations were also evaluated, but are only listed if they were found to be a better fit.

4. ALGORITHM TWEAKS

In order to thoroughly explain user data we might need to move in the direction of more complex multiband models, but given ITU-R BS.1770's advantages, we tried to understand if some slight parameter modifications could lead to more consistent results. Two likely candidates were:

[3] The RLB-filter is a weighting filter, and as such is an inversion of the equal-loudness contours. To that extent it is missing a peak in the $1.5 - 4$ kHz region. The pre-filter already has a boosting high-shelf, and may be a more efficient candidate for change.

FEATURE	r^2
log2 Spectral Bandwidth	0.5502
log2 Spectral Centroid	0.5217
Spectral Kurtosis	0.3633
Spectral Crest	0.3512
Spectral Skewness	0.3393
Spectral Spread	0.2383
Spectral Tonal Power Ratio	0.2377
Spectral Flatness	0.2251
Spectral Q	0.3418
Peak-RMS Ratio	0.2045
Avg Event Peak-RMS level	0.2020
Crest Factor	0.1808
Best fit MFCC (4)	0.1601
Spectral Decrease	0.1254
Avg Event Crest Factor	0.1183
Zero Crossing Rate	0.1113
Autocorrelation Coefficient	0.1051
Standard Deviation	0.0811
Avg Event Attack Time	0.0771
Avg Event Length	0.0697
Loudness Range (EBU)	0.0665
Spectral Flux	0.009
Spectral Slope	0.0002

Table 1: Features tested against the evaluation and their squared correlation coefficient.

- The pre-filter's gain value. ITU's coefficients place it at $+4$ dB. This roughly simulates the big sensitivity boost we see in loudness contours, but it trades peak gain for a broader bandwidth (it is a high shelving filter). Expecting that its value could be a little bit higher, we varied it in the interval of -2 to 14 dB.
- The time constant used as windowing value to the gating block, defined in the specification as 400 ms. We experimented with time constants going from 20 to 600 ms in steps of 20 ms. This is a very sensitive parameter, and while it is not directly related to frequency content, it nevertheless influences how it affects measurement.

Let us consider a collection of vectors $\overline{M}(i,k,m)$ that hold the loudness of the vocal track of song i, considered in light of a modified ITU algorithm

with a pre-filter dB gain of $-2 \leq k \leq 14$ and a time constant of $m = \{20, 40, 60, ..., 600\}$. Let vectors $\overline{E}(i, j, k, m)$ represent the average loudnesses, as calculated using k and m of track j for song i, when set by the test subjects to be equally loud to the vocal track. The error in loudness measurement using k and m between perceived and measured loudness of track j for song i is then given by $\overline{D}(i, j, k, m) = \overline{E}(i, j, k, m) - \overline{M}(i, k, m)$. There are four different optimization schemes we will try:

- Scenario α: Calculate the matrix of discrepancies between measurement and evaluation with the vocal as baseline:

$$a_{\alpha(k,m)} = \sum_i \sum_j \left|\overline{D}(i, j, k, m)\right|. \qquad (2)$$

- Scenario β: Calculate the maximum absolute error:

$$a_{\beta(k,m)} = \max_{i,j} \left(\left|\overline{D}\left(i, j, k, m\right)\right|\right). \qquad (3)$$

- Scenario γ: Calculate the interval between maximum and minimum deviation from measurement (error spread):

$$a_{\gamma(k,m)} = \max_{i,j} \left(\bar{D}\left(i, j, k, m\right)\right) - \min_{i,j} \left(\bar{D}\left(i, j, k, m\right)\right). \qquad (4)$$

- Scenario δ: Noting that situation α is not a true error minimization process, as the errors we get taking the vocal as reference are almost unipolar, as seen in Fig. 2. To solve this, we can apply a similar process to $\overline{D}_a(i, j, k, m) =$

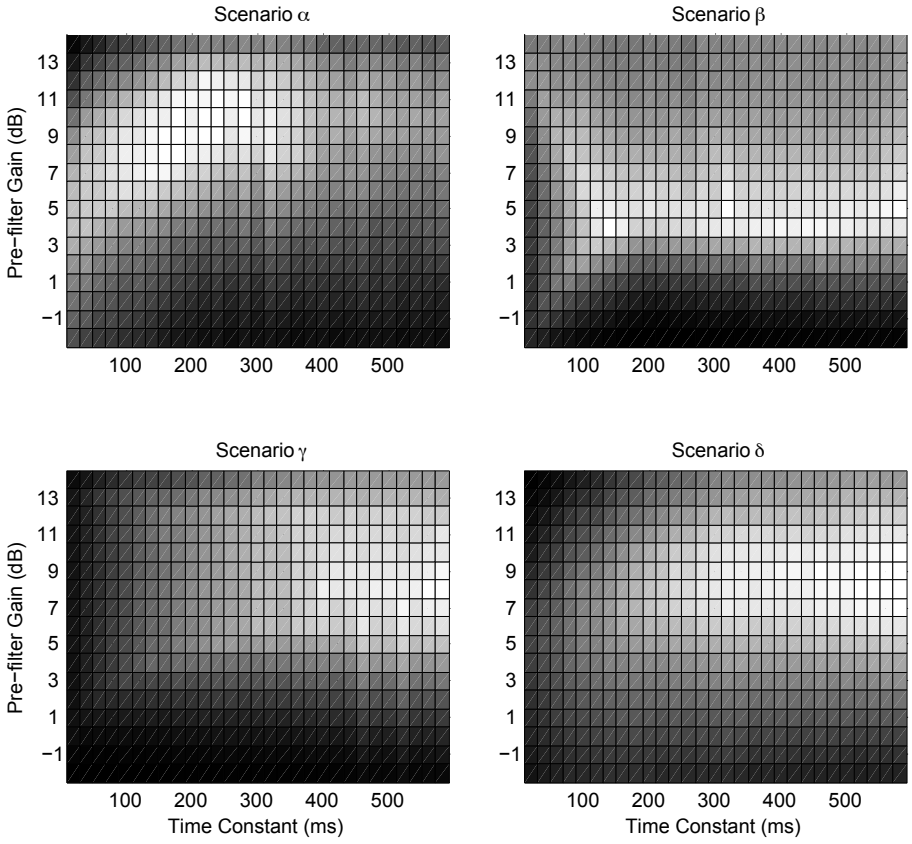

Fig. 4: Matrices that describe the performance of several parameter combinations (time-constant vs. pre-filter high-shelf gain) for the four different scenarios described in the text. Brighter colors indicate better results.

$\overline{E}(i,j,k,m) - \overline{E}_a(i,k,m)$, the evaluation referenced to the average value of the evaluation distribution:

$$a_{\delta(k,m)} = \sum_i \sum_j \left| \overline{D_a}(i,j,k,m) \right| . \quad (5)$$

Regardless of the chosen scheme, we now want to find the minimal matrix value:

$$\min_{k,m} \left(a_{s(k,m)} \right) , \quad (6)$$

with $s = \{\alpha, \beta, \gamma, \delta\}$, the collection of the four optimization options. In figure 4 we can see how the choice of parameters affects how well the specification would fit the user data. The four a_s values are plotted as intensity against the parameter choice, so the brighter a cell is, the better the fit, and the brightest cell will be the result of Equation 6. We see different clusterings of white regions for each of the hypothesis, indicating that they will sometimes contradict each other, even though it is safe to say that the most appropriate pre-filter gains are always larger than the one in the recommendation ($+4\ dB$). In the two bottom situations, a larger time-constant seems to be favored, whereas the first situation calls for a smaller one.

Let us consider a measure of improvement or accuracy as simply a percentage that reflects the ratio between the result of the best parameters (k_b, m_b) to the result of the original parameters, for each of the four situations outlined above:

$$acc(s) = \left(1 - \frac{a_{s(k_b,m_b)}}{a_{s(4,400)}} \right) \times 100. \quad (7)$$

Our calculations show us that if we are to consider scenario α, the best approach would be to have a time constant of 280 ms with a pre-filter gain of 10 dB. That would increase our accuracy by around 14%. Minimizing the matrix resulting from the calculation in β yields worse relative results (a 3% increase, which corresponds to the parameters 320/5). This possibly means that there is an isolated extreme case, whose measurement cannot be changed by the modification of our chosen parameters. Scenario γ is more promising, as it achieves a 29% improvement for all data (with parameters 600/8) and δ a 21% improvement at the position 580/9.

Figure 5 shows the new relationships between measurement and evaluation. The zero line is referenced to the vocal, and the bar plots show the new discrepancy of evaluation, if measurements are done with each of the new sets of potentially optimal parameters. It is interesting to see that the element for which there is a better overall consistency increase are the high percussion, a fact that possibly results from our raising of the pre-filter gain. The downside is that kick drum, bass guitar and acoustic guitar seem to generally be in a bigger disagreement between measurement and evaluation, regardless of the choice that is made. The overall improvement seems limited, but due to the very high variance of the evaluation, it may well be that some of the new measurements will resonate well with perception.

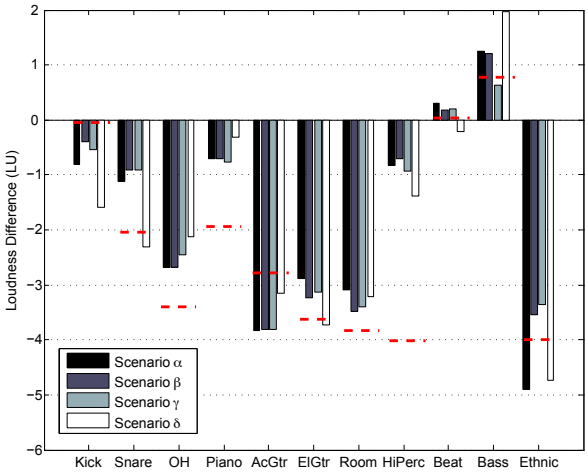

Fig. 5: New results referenced to vocals. The dashed horizontal lines represent the old results, and the four bars show how the scenarios presented in the text could improve the algorithm's agreement with the evaluation data.

5. NEW PARAMETER EVALUATION

It is yet unclear which optimization should be considered, and results seem to point into different directions. It is, however, quite straightforward to put each of the possible scenarios to test, and we performed a very informal follow-up subjective evaluation session with three professional mixing engineers. This was a multiple stimulus test, based on

MUSHRA [11], where the reference was a mix done by a sound engineer[4], and the usual anchor was the same excerpt, low-passed at 3.5 kHz. The remaining stimuli were produced by mixing the multitracks so that they would be equally loud following:

(a) The ITU/EBU recommendation.

(b) The alteration according to α, where the time constant is now 280 ms and the pre filter gain $+10$ dB.

(c) The alteration according to γ, where the time constant is now 600 ms and the pre-filter gain $+8$ dB.

(d) The alteration according to δ, where the time constant is now 580 ms and the pre filter gain $+9$ dB.

We took the best β parameters out of the test as they seemed to yield little improvement over (a) and would act as an additional distraction. The subjects were asked to score the stimuli according to how well the equal-loudness purpose was achieved for 12 runs of random tests drawn out of a pool of six songs[5]. Only four songs were used for the final analysis based on subject comments that they heard no difference between the examples of the remaining two songs. There is a strong reason for this as the two problematic songs were a song whose individual tracks are themselves stems (drum mix down, guitar mix down, etc.), which means there are no narrow bandwidth tracks, and an ethnic song with 6 instruments on the same frequency register. The inclusion of these songs would have flattened out the results as the subjects had scored all variations except the anchor similarly. The summary results for the remaining songs are presented in Figure 6.

While the reference and anchor were clearly identified, and following the recommendation seems unambiguously worse than modifying it, the overlapping confidence intervals of the tweaking options make it again difficult to establish one of them as the more sensible to chose. The parameter set that minimizes the total error (600/8) ended up as the lowest-scoring of all three. Its confidence interval does not even overlap with the best choice (280/10), and a subjective test with professional mixers should be considered a stronger answer, even if it is a very small-scale evaluation. It might seem that increasing the time constant simply reduces the spread of loudness values, but informal tests showed that this is not the case.

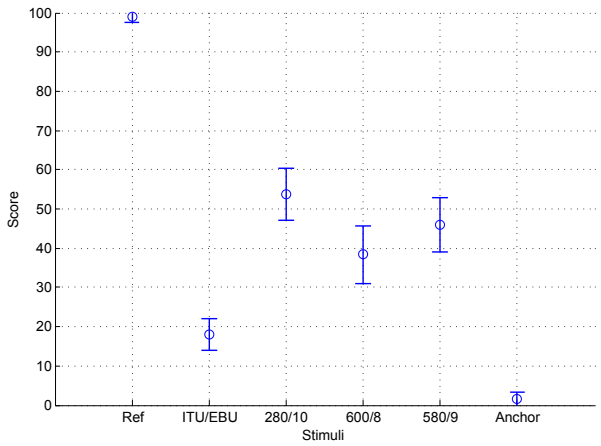

Fig. 6: Means and 95% confidence intervals for the six stimuli in the MUSHRA test. The reference and anchor are far left and far right, and the remaining show, from left to right: the results following the ITU/EBU recommendation, the results changing the parameters to 280/10, 600/8 and 580/9.

6. CONCLUSIONS

The ITU/EBU recommendation for loudness metering is both an effective and widespread method for the analysis of broadband mixes. We have seen that the evaluation of individual tracks in a multitrack context is a much more specific task, where the algorithm fails to agree with human subjects. This work presents some suggestions as to why this is so and what steps can be undertaken to make the algorithm applicable. Somewhat surprisingly, frequency content seems to be the only cause of discrepancy and we have proposed a different pair of fixed parameters that would optimize the recommendation if multitrack content is to be considered. This has been evaluated by an independent panel of expert listeners who have validated our hypothesized values.

Psychometric tests are very prone to bias and though

[4] Under the goal of equal-loudness.

[5] The songs were deliberately different from the ones made on the previously described test.

there was great care on the methodology, it cannot 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 that overwhelms the (expected) disagreement between individuals themselves.

We have shown that careful tweaking of selected parameters can improve the agreement between algorithm and the average user, and specifically suggested a time constant and pre filter gain alteration to a 280/10 combination. This was thought to be better by three professional sound engineers in an independent quality test, but further tests are needed to validate the proposed alternatives against a larger dataset, evaluated by a more diverse panel of listeners. Another promising approach is to understand how the evaluated data fits with psychoacoustic models such as Glasberg and Moore's [12].

7. ACKNOWLEDGEMENTS

The author P.D. Pestana was sponsored by national funds through the Fundação para a Ciência e a Tecnologia, Portugal, grant number SFRH/BD/65306/2009, and project (PEst-OE/MAT/UI2006/2011.

FCT Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA EDUCAÇÃO E CIÊNCIA

8. REFERENCES

[1] ITU, "ITU-R-BS.1770: Algorithms to Measure Audio Programme Loudness and True-peak Audio Level BS Series," tech. rep., International Telecommunications Union, 2006.

[2] EBU, "Tech Doc 3341 Loudness Metering: EBU Mode Metering to Supplement Loudness Normalisation in Accordance with EBU R 128," Tech. Rep. August, European Broadcast Union, Geneva, 2010.

[3] S. Mansbridge, S. Finn, and J. D. Reiss, "Implementation and Evaluation of Autonomous Multi-track Fader Control," in *Proceedings of the 132nd AES Convention*, (Budapest), Audio Engineering Society, 2012.

[4] J. A. Maddams, S. Finn, and J. D. Reiss, "An Autonomous Method for Multi-track Dynamic Range Compression," in *Proceedings of the 15th Int. Conference on Digital Audio Effects (DAFx-12)*, 2012.

[5] P. Pestana and A. Barbosa, "Accuracy of ITU-R BS.1770 Algorithm in Evaluating Multi-track Material," in *133rd AES Convention, Engineering Brief*, 2012.

[6] SMPTE, "RP 200:2012 "Relative and Absolute Sound Pressure Levels for Motion-Picture Multichannel Sound Systems — Applicable for Analog Photographic Film Audio, Digital Photographic Film Audio and D-Cinema"," tech. rep., Society of Motion Picture and Television Engineers, 2012.

[7] S. Bech and N. Zacharov, *Perceptual Audio Evaluation — Theory, Method and Application*. Chichester: John Wiley & Sons, 2006.

[8] E. Skovenborg, R. Quesnel, and S. H. Nielsen, "Loudness Assessment of Music and Speech," in *Proceedings of the 116th AES Convention*, (Berlin), Audio Engineering Society, 2004.

[9] P. Sprent and N. C. Smeeton, *Applied Nonparametric Statistical Methods*. New York: Chapman & Hall - CRC Texts in Statistical Science, fourth ed., 2007.

[10] ISO, "ISO/IEC 15938 Information technology – Multimedia content description interface – Part 4: Audio," tech. rep., International Organization for Standardization, Geneva, 2002.

[11] ITU, "ITU-R BS.1534-1: Method for the subjective assessment of intermediate quality levels of coding systems," tech. rep., International Telecommunications Union, Geneva, 2003.

[12] B. R. Glasberg and B. C. Moore, "A Model of Loudness Applicable to Time-Varying Sounds," *Journal of the Audio Engineering Society*, vol. 50, no. 5, pp. 331–342, 2002.