Audio Engineering Society

Convention Paper

Presented at the 143rd Convention
2017 October 18–21, New York, NY, USA

Alternative weighting filters for multitrack programme loudness measurement

Steven Fenton and Hyunkook Lee

Applied Psychoacoustics Lab (APL), The University Of Huddersfield, Queensgate, Huddersfield, UK

Correspondence should be addressed to s.m.fenton@hud.ac.uk

ABSTRACT

The ITU-Recommendation BS.1770 is now established and recognised throughout most of the broadcast industry. Programme loudness measurement is undertaken through the summation of K-weighted energy and this summation typically involves material that is broadband in nature. In this work, we undertook controlled listening tests to investigate the performance of the K-weighting filter in relation to perceived loudness of narrower band stimuli, namely octave-band pink noise and individual tracks of a multi-track session. We propose two alternative filters based on the discrepancies found and evaluate their performance. The new filters yield better performance accuracy than the K-weighting filter in the lower octave bands when measuring the pink noise bands. In addition, the measurement window size is also investigated and is shown to yield significant variance in the prediction of loudness for certain types of stem. Finally, we propose an informed set of parameters that may improve loudness prediction used in auto mixing systems.

1 Introduction

In broadcast and mastering, the ITU-R BS.1770 [14] algorithm for objective loudness measurement is in widespread use, its primary purpose being to normalise audio levels in the broadcast chain. Previous studies, which evaluated the algorithm based on correlation to listener perceived loudness, found that whilst the algorithm can predict programme loudness, the correlation accuracy largely depends on the use of broadband stimuli [1].

It has been shown that if the loudness is measured whilst only considering the lower octave bands, or if it is based upon the measurement of individual stems within a multitrack session (perhaps for use in the control of an auto mixing system), the variance in loudness prediction is greatly increased and in some cases under estimated [2,3,4].

Despite this, the algorithm offers an elegant solution with respect to reduced implementation complexity when compared to other models such as the Moore-Glasberg model [5]. Research has also shown that the algorithm, when used in loudness feature extraction to automatically control levels of stems in a multitrack, can outperform more complex loudness models [6].

In view with these findings, further research into the effectiveness of the ITU-R loudness algorithm when used to predict the loudness of narrow band and individual stem based stimuli (commonly found in a song multitrack) could be beneficial to the wider community. Indeed, empirical research by the authors has observed some disagreements with loudness perception and the algorithm output when measuring vocal stems, narrow band and other transient sources. Similar findings have also been reported [2-4,7].

The aims of this paper are therefore outlined as follows:

- To investigate the frequency-weighting curves used in the ITU-R BS.1770 algorithm with a view to their effectiveness in predicting the loudness of octave band width noise stimuli.
- Investigate modifications to the frequency-weighting curves and measurement algorithm to establish if a more optimal parameter set is possible.
- Repeat the investigation using individual stems commonly found in a song multitrack.
- Perform an informal auto-mixing exercise based on the baseline and optimal parameter set.

2 Octave-Band Noise Listening Test

In-order to investigate the frequency-weighting curves used in the ITU-R algorithm, a loudness matching experiment was undertaken. For this test, octave-band pink noise stimuli were chosen.

The reason for this type of stimuli was that it would allow a frequency specific response to be obtained, it would allow others to repeat the test with the same readily available test stimuli and it would also allow a uniform and logarithmic range of frequencies to be tested. As noise is a temporally static signal, no bias with respect to transient onsets would be introduced into the loudness judgements. All stimuli were presented in a continuous and click-free loop.
8 octave bands were tested, with the lowest band having a centre frequency of 63 Hz and the upper band a centre frequency of 8 kHz. Testing of the 32 Hz octave band was deemed impractical due to the available playback system.

The octave band filters used to create the stimuli were 8th order. Prior to playback, all stimuli were convolved with an inverse filter response obtained at the listening position using a sine-sweep. This enabled all the stimuli to be presented without any room bias effects.

2.1 Listening test method

All listening tests took place within the ITU-R BS.1116 compliant listening room at the Applied Psychoacoustics Laboratory at the University of Huddersfield. The listening room dimensions are 6.2 (W) x 5.2 (D) x 3.8 (H) in metres, RT = 0.25s, NR 12.

Each listener was seated 2 metres from a centrally positioned Genelec 8040A powered speaker. The reference stimuli playback level was fixed at 75dB(Z). All stimuli were mono WAV format, 24bit, with a sample rate of 48 kHz. All listeners were asked to face the speaker when making judgements.

13 listeners were asked to equally match the loudness of octave band filtered pink noise stimuli to a known 1 kHz octave band pink noise reference.

Stimuli were presented in random order as REF-X pairs. The reference was always fixed as the 1 kHz octave band. Each listener was allowed half an hour for the test with an optional break in the middle.
The responses were automatically recorded on a laptop computer. The test interface is shown in Figure 1.

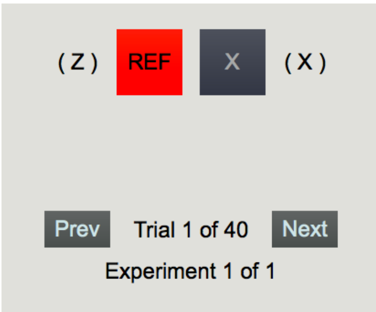

Figure 1. Test Interface.

The interface was designed to allow the listener to select either the reference or stimuli under test using Z and X on a keyboard respectively. Clicking Prev and Next allowed the listeners to step through the stimuli and make any changes as required. Audio level adjustment was achieved using a rotation knob device (Griffin Powermate) which the listener would turn either clockwise or anticlockwise to increment or decrement the level in 0.1dB steps. There was no indication of the level on screen or on the rotation knob itself thus reducing possible bias effects.

Each listener repeated the experiment with a total 80 comparisons being made across the 8 octave bands. All stimuli were randomised in order. In total, 1040 responses were recorded. A high level of inter-subject consistency was observed across all the octave bands with a mean standard deviation of 1.49dB observed when taking all the bands into consideration. The largest standard deviation was observed in the 125 Hz band with a value of 2.016dB.

The octave band noise stimuli were also measured using the ITU-R algorithm. LU relative to the 1 kHz band was calculated for each band and the results compared against the listener data. Measured loudness and the loudness matching adjustments, both have different SI. However, the unit of adjustment in dB can be plotted on the same LU axis due to its equivalence to this scale.

All measurements taken using the BS.1770 algorithm are of the 'integrated programme' loudness type employing a gating function as specified in the ITU recommendations.

2.2 Listening test results

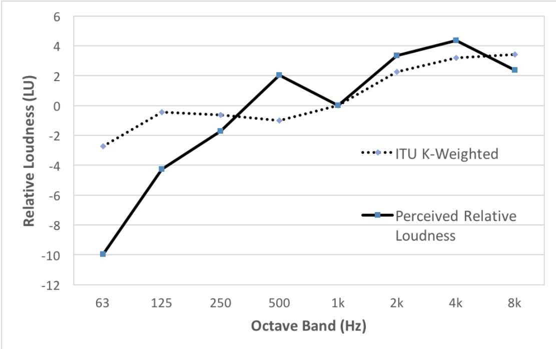

Figure 2. Octave band pink noise equal-loudness contour compared to BS.1770 LU measurements.

Observing Figure 2, the solid line shows the median value of perceived loudness relative to the 1 kHz octave-band. Therefore, the equivalent level of gain adjustment needed to achieve equal loudness would be the inverse of this curve. As can be seen, there are some differences between the listener perceived loudness and the objective measurement (shown as the dotted line).

These results correspond well to previous studies involving loudness perception of octave-band pink noise stimuli [2,3]. In Stevens [8] method for prediction of loudness, a -12dB roll off per octave is incorporated at 9 kHz, whilst we didn’t test beyond the 8 kHz octave band, a roll off was observed beyond the 4 kHz band.

Taking the 63 Hz octave-band as an example, there is an overestimation of energy made by the algorithm when measuring this band. The listeners in this case would apply approximately 9.975dB of gain to attain equal loudness to the 1 kHz band. The LU measurement would suggest that only a boost of 2.73dB was required. Our observations suggest a greater roll off and/or change of cutoff is required in the filter to match the perceived response.

The objective measurements obtained clearly show the K-Weighting filter response implemented in the ITU-R algorithm. There is however a 1dB dip evident in the 500 Hz octave band. This could be a result of integration window size employed in the algorithm. This will be explored later.

If the R.M.S error as a function of gain is calculated across the 8 bands, a value of 3.184dB is achieved. One would expect this value to equal 0dB for a perfect match between the objective and perceptual responses.

3 Alternative Filter Weightings

To attempt to model the perceptual results more accurately, three alternative filter arrangements were employed into the BS.1770 algorithm and the pink noise measurements were repeated. Two of the filter arrangements were of similar complexity to the existing K-weighting filter whilst another of higher complexity (namely in order) was tested for comparison.

The first arrangement employed was one suggested by Dash et al. [2]. This filter employs two 2nd order IIR filters and a gain stage. The first IIR filter is configured as a high pass filter whilst the second is a dip filter. The gain stage is utilised to compensate for the loss in level introduced by the dip.

The second arrangement employed is similar to the K-weighting filter albeit for modification of both the hi-shelf gain and the hi-pass cut off frequency. These were chosen to be 5dB and 130 Hz respectively. In addition, a 2nd order peak filter was added with a centre frequency of 500 Hz.

Finally, a third arrangement was tested which replaced the 2nd order peak filter with a higher order equivalent. This was to mimic the exact bandwidth and bell shape of the 500 Hz octave-band filter.

All the filter responses are shown in Figure 3.

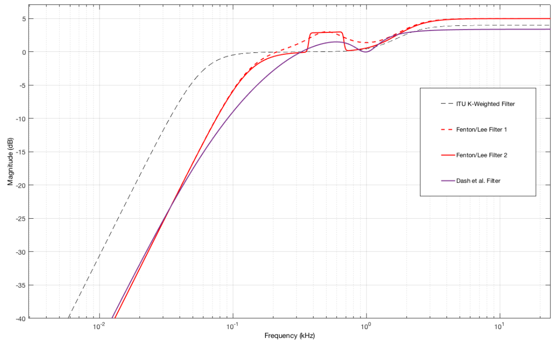

Figure 3. Alternative Weighting Filters.

3.1 Error analysis of weighting filters

If the relative LU measures of each band are examined relative to the listener equal-loudness offsets obtained in the octave-band noise listening test, R.M.S error as a function of gain is possible. The following table shows the calculations made for each filter type. To make the measurements comparable, the LU measurements are referenced in each case to its 1 kHz band. All values shown are in dB unless otherwise stated.

Table 1. R.M.S errors of the filters compared to listener perceived loudness.

Octave-band Centre (Hz)	**ITU K-Weighted Filter**	**Fenton/ Lee Filter 1**	**Fenton/ Lee Filter 2**	**Dash et al. Filter**
63	-7.244	2.262	1.490	3.646
125	-3.794	-0.756	-1.357	1.711
250	-1.114	-1.010	-0.983	0.095
500	3.050	1.364	0.547	1.930
1k	0	0	0	0
2k	1.087	1.428	0.769	1.280
4k	1.182	1.104	0.312	1.813
8k	-1.040	-1.186	-2.011	-0.397
R.M.S Error	**3.184**	**1.286**	**1.121**	**1.769**

As can be seen, the highest R.M.S error is evident with the K-Weighting filter, with the largest errors being due to the 63 Hz, 125 Hz and 500 Hz bands. The filter proposed by Dash et al. [2] shows an overall R.M.S error of 1.769. This reduction is error is due to the improved correlation to listener perception in the 63 Hz, 125 Hz and 500 Hz bands.

The authors' proposed filters perform with R.M.S errors of 1.286 and 1.121 respectively. The latter being the higher order implementation, Filter 2. Given the error improvement of only 0.165dB evident, the use of more complex filter arrangements for this purpose is perhaps not warranted.

Evidenced by the LU measures taken using all the filters, there is still some discrepancy evident in the 8 kHz band when compared to the listener perceived loudness. Referring to Figure 2, the 8 kHz band shows a dip of approximately 2dB compare to the 4 kHz band. Further reductions in error due to this band could be achieved if this dip was modelled.

Taking both filter complexity and the R.M.S errors of the arrangements into account, the best performing filter is the second arrangement (Filter 1). This has a hi-shelf gain of 5dB and hi-pass cut-off of 130 Hz. In addition, a 2nd order peak filter was added with a centre frequency of 500 Hz.

Previous test results [9] have shown that whilst the ITU-R algorithm is effective when evaluating broadcast material, typically containing speech and broadband spectra, larger variances in prediction are seen when measuring signals that exhibit a narrower frequency spectrum, such as those found in 'effects' or 'single instrument' type sounds [1]. For the objective measurement of temporally static and narrower band signals, there is therefore some improvement that can be achieved in loudness prediction through the modification of the existing filters employed in the ITU-R algorithm.

4 Multitrack Stem Listening Test

Individual stems of a multitrack, commonly found in a music production session could be considered 'atypical' of broadcast material typically measured using the BS.1770 algorithm. With that said, previous research by Wichern et al.[6] demonstrated that the algorithm, when used for loudness feature extraction to automatically control levels of stems in a multitrack, can outperform more complex loudness models. In this work, the authors suggested possible modifications to the Moore-Glasberg models to refine them towards music material rather than laboratory test tones. However, without consideration to 'partial' loudness, the use of a simpler mechanism of loudness extraction for use in auto-mixing applications, at least to attain a suggested starting rough mix point, is preferable.

Based on the results outlined in Section 2, an additional loudness matching experiment was performed to establish the effectiveness of the new filters in predicting the loudness of individual stems typically found in a multitrack session.

The stimuli were multitrack stems which were mono WAV format, 24bit and 48 kHz.

The stems were kick, snare, bass (electric), guitar 1 (clean electric), guitar 2 (distorted electric), vocal, overheads and tambourine. Each stem was 14 seconds in length and were presented in a continuous loop.

14 listeners were asked to equally match the loudness of each stem to the reference using the same method of adjustment outlined in Section 2. Stems were presented in random order as REF-X pairs. The reference was always fixed as vocal stem. The reference playback level was fixed at 75dB(Z). Each listener was allowed half an hour for the test with an optional break in the middle.

Prior to playback, all stems were convolved with an inverse filter response obtained at the listening position using a sine-sweep. This reduced any room bias effects. The responses were automatically recorded on a laptop. The test interface is shown in Figure 1.

Each listener repeated the experiment with a total 80 comparisons being made. All stimuli were randomised in order. In total, 1120 responses were recorded. A high level of inter-subject consistency was observed across the stems however, some variance in the ranges was observed, this can be observed in Figure 4.

LUFS measurements were taken for every filter instance on all the stems using the ITU-R algorithm. Relative LU levels were then calculated with respect to the vocal stem. These levels were then compared to the perceived equal-loudness offsets derived from the listeners.

4.1 Perceived loudness of stems

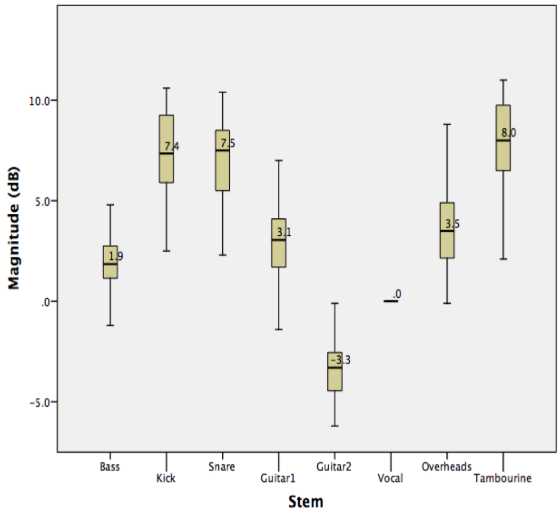

Figure 4. Relative dB adjustment for each stem.

The level of dB adjustment required for each stem to achieve equal loudness to the vocal stem is shown in Figure 4. For clarity, the median adjustment levels are highlighted.

This data can be directly compared to the objective measures made using the different filter types. This comparison is shown in Figure 5. The listener perceived adjustment is shown as the inverse of the median levels derived from Figure 4.

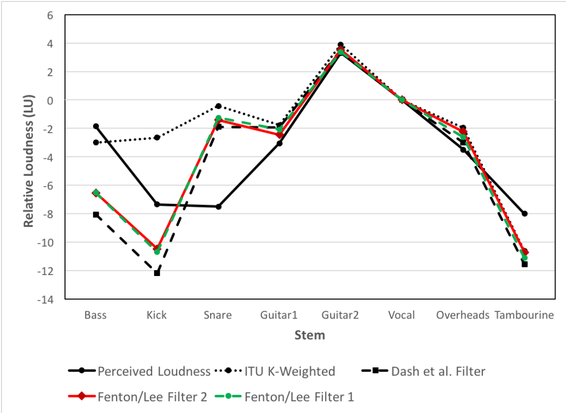

Figure 5. Relative perceived loudness compared to measured loudness for differing filter types.

All filter types gave similar results when measuring the guitar 1, guitar 2 and overhead stems. These results correlate well with the listener perceived loudness. Table 2 shows the relative dB error of each measured stem to the perceived listener level per filter, also shown are the overall R.M.S errors when taking all the stems into account. For example, the kick stem, when measured by the K-Weighting filter, is overestimated (-ve offset) by 4.703dB when compared to the listener perceived equal loudness judgement.

Table 2. R.M.S errors of the perceived loudness with the objective measures.

Stem	ITU K-Weighted Filter error (dB)	Fenton/Lee Filter 1 error (dB)	Fenton/Lee Filter 2 error (dB)	Dash et al. Filter error (dB)
Bass	1.144	4.672	4.683	6.198
Kick	-4.703	3.342	3.120	4.835
Snare	-7.074	-6.242	-6.098	-5.611
Guitar1	-1.286	-0.969	-0.611	-1.123
Guitar2	-0.592	-0.077	-0.205	-0.204
Vocal	0	0	0	0
Overheads	-1.541	-0.875	-1.266	-0.501
Tambourine	2.631	3.1195	2.713	3.567
R.M.S Error	3.255	3.229	3.127	3.667

Firstly, the bass stem, gave the most surprising result with the K-Weighting filter having the smallest error of only 1.144dB compared to the perceived listener loudness. A possible explanation for this could be that the bass guitar loudness was not being entirely judged by its low frequency content but rather its transient content, perhaps in the form of its perceived 'punch'. The revised filters, having less weighting in the lower octaves that the K-weighting filter would give a lower loudness estimate, thus in these cases contrary to the weightings suggested in the pink-noise test. The Dash et al. filter performed the worst with respect to prediction of the bass stem loudness.

The kick stem loudness was somewhat overestimated by the K-Weighting filter with a relative loudness of -2.647dB compared to a perceived relative loudness of -7.35dB.

Filters 1 & 2 gave the smallest errors when predicting guitar 2 with errors of -0.077 and -0.205 respectively. The largest error observed is the K-Weighted filter prediction of the snare stem with an error of -7.074dB.

Taking all the stems into account and calculating the overall R.M.S error with respect to the listener perceived loudness, all four filters perform at a similar level. There is very little difference between the K-Weighting filters and the two new filters proposed although Filter 2 has the lowest overall R.M.S error with a value of 3.127dB.

4.2 Effect of measurement window size

Thus far, the measurements presented have all been taken with the standard 400ms, 75% overlap window size specified in the ITU recommendation. As one might expect, the loudness range (LRA) of the stems will vary dependent upon the nature of the audio. Stems that are largely transient in nature will have a larger LRA than those that tend towards steady state. This can be seen in Figure 6 which shows the LRA measurements for each stem. The measurements were repeated using the different filter arrangements and the general trend of LRA remained the same.

Given these differences, and the nature of the LRA statistic [10], the stems exhibiting higher LRA can be assumed to have larger dynamic range, as such modification of the measurement window size could yield smaller errors between the listener perceived loudness and the objective measures being made. Pestana et al. [4] made similar observations but spectrally, they only modified the shelving gain factor of the filter employed. The filters employed in this paper are modelled on the perceptual response to octave band pink noise stimuli.

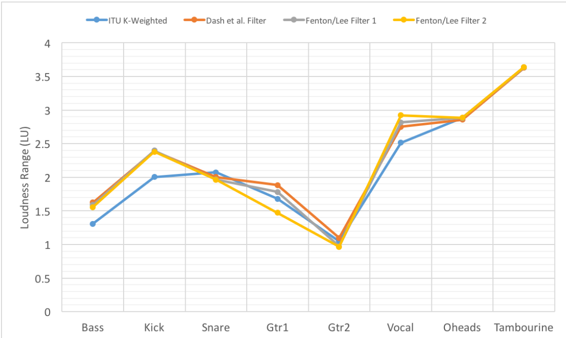

Figure 6. LRA Measurements for each stem, per filter.

Figure 7 shows the individual stem errors based on the modification of the measurement window size for each filter. The window sizes tested were 140ms, 280ms, 400ms, 500ms and 800ms. The data has been shown as a heat map, with the minimum error (0dB) shown as white.

In all filter cases, the bass, guitar 1, guitar 2 and overheads appear to be largely immune to variation in window size with respect to the calculated error. Apart from the overheads, these stems exhibit the smallest loudness ranges (Figure 6). The overheads stem is primarily formed by the drum kit, incorporating ride, hi-hat and cymbals. The loudness, despite having variation could be considered more steady state than, for example, the close miked snare stem. As such, all the 'steady state' sounds could be considered largely immune to the window size variation.

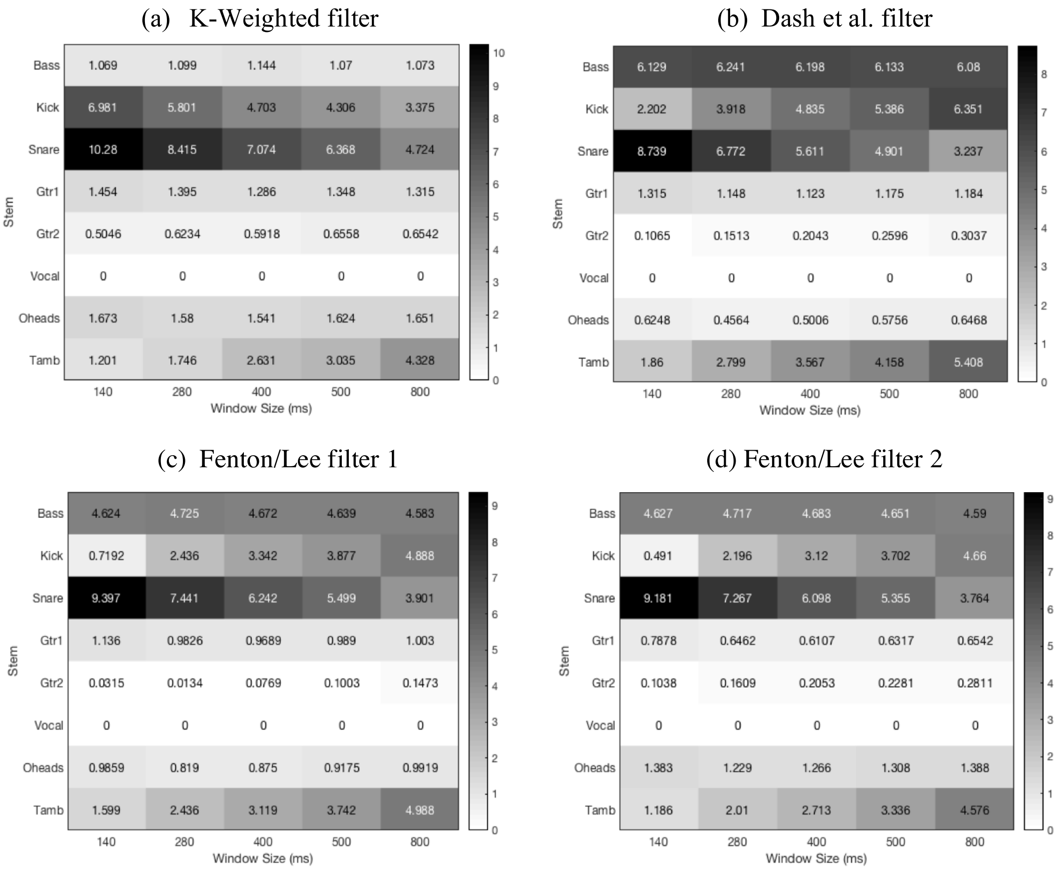

Figure 7. LU errors of different filters based on window size.

With that in mind, the accuracy of prediction would be down primarily to the filter weighting curves. In the case of the overheads, a heavier weighting applied to the shelving filter above the 5dB tested could yield better prediction accuracy. Pestana et al. [4] utilised shelving gains in their study of 8dB, 9dB and 10dB, all giving an approximate error decrease of between 1-1.5dB.

For the Fenton/Lee Filter 1, Filter 2 and the Dash et al. Filter, the best window size for prediction of the kick stem loudness was 100ms, with Filter 2 performing the best. In these cases, the error can be seen to increase as the window size is increased. The K-Weighting filter displayed the opposite, with the 800ms window giving the best prediction and the 100ms the worst.

The snare stem, in all filter cases, was the stem exhibiting the greatest error. The largest being 10.28dB in K-Weighting filter case measured with a 100ms window size. These errors decreased in every case as the window size was increased.

Comparing these results with a previous study, Pestana et al. [4] also found an increase in kick error

magnitude as the window size was increase however, we saw the opposite in the K-Weighted filter case. The snare error magnitude in all studies showed a decrease as the window size was increased.
Finally, the tambourine stem errors increased in all filter cases as the window size was increased.

Interestingly, if the total R.M.S error of every filter / window combination is compared the best performing is the K-Weighting filter with a window size of 800ms. Its error magnitude is 2.704dB. The worst performing combination is the Dash et al. filter with a window size of 140ms, its error magnitude is 3.943dB.

The Fenton/Lee Filter 1 achieves a minimum error of 3.214dB whilst Filter 2 achieves a minimum error of 3.108dB, both with window sizes of 500ms.

5 Optimised Parameters for Stem Measurement

In cases where narrow band or instrument based stimuli are being measured for loudness, it's been shown that both filter modification and measurement window size optimisation in the ITU-R algorithm would yield better loudness prediction.

It has been shown that for temporally static octave band pink noise stimuli, Filter 1 would offer the best performance if implementation complexity is a factor.

An auto mixing system can be implemented by measuring each stem loudness within a multitrack session and balancing them either by equal loudness [11,12] or using a system of relative loudness made towards a reference stem such as the vocal [6]. The latter is the preferred as previous studies have shown that engineers do not mix towards the equal loudness goal [13,14].

Should the method of loudness extraction utilise the ITU-R algorithm, an optimised parameter set outlined in Table 3 could yield a more accurate representation of the desired final mix.

Stem	Filter Type	Window Size (ms)
Bass	K-Weighted	400
Kick	Filter 1	140
Snare	K-Weighted	800
Guitar (Clean Electric)	Filter 2	800
Guitar (Distorted Electric)	Filter 1	140
Vocal	*	*
Overheads	Dash et al. Filter	280
Tambourine	Filter 2	140

Table 3. Optimised parameters.

For the reference vocal stem, the variation in window size (across all the filters tested) resulted in very little LU variation. However, due to variations in loudness measurement as a result of different weighting filters, it's important to reference to the vocal stem measured with the same filter type as the stem being adjusted. This will ensure that the appropriate dB gains are applied in the automixing process.

6 Conclusions

Whilst the ITU-R algorithm has been proven to be successful in the loudness prediction of broadband material, the algorithm doesn't agree with listener perceived loudness involving some single instrument and narrow band stimuli. Whilst this may not be of concern due to the stimuli being somewhat atypical, one should be aware of the possibility of under or over estimation of loudness in these cases. This work presents both alternative filters and window size optimisations that can reduce the error magnitude with these atypical stimuli.

Further work is planned with a view to implementation of an auto mixing system based around the optimised parameters proposed. This will also incorporate a wider stem set and a controlled listening test to evaluate the performance of the final mixing stage.

References

[1] Soulodre, G.A. & Lavoie, M.C., “Stereo and Multichannel Loudness Perception and Metering”, 119th Convention of the Audio Engineering Society, NewYork, October 2005, Preprint 6618

[2] Dash, I., Miranda, L. & Cabrera, D., “Multichannel Loudness Listening Test”, 124th Convention of the Audio Engineering Society, Amsterdam, May 2008, Preprint 7451

[3] Dash, I., “Octave-band Analysis on ITU-R Loudness Listening Test Data”, 126th Convention of the Audio Engineering Society, Munich, May 2009, Preprint 7811

[4] Pestana, P., Reiss, J. & Barbosa, A., “Loudness Measurement of Multitrack Audio Content using Modifications of ITU-R BS.1770”, 134th Convention of the Audio Engineering Society, Rome, May 2013, Preprint 8813

[5] Moore B. C. J., "Development and Current Status of the 'Cambridge' Loudness Models,", *Trends in Hearing* , vol. 18, pp. 1-29, 2014.

[6] Wichern, G., Wishnick, A.S., Lukin, A., & Robertson, H., “Comparison of Loudness Features for Automatic Level Adjustment in Mixing”, 139th Convention of the Audio Engineering Society, New York, October 2015, Preprint 9370

[7] Pestana, P., & Barbosa, A., "Accuracy of ITUR BS.1770 Algorithm in Evaluating Multi-track Material", 133rd Convention of the Audio Engineering Society, San Francisco, Engineering Brief , 2012.

[8] Stevens, S.S., "Procedure for calculating loudness:Mark VI", J. Acoust Soc. Am. Vol 33 pp 1577 .1585, November 1961.

[9] Soulodre, G.A. & Lavoie, M.C., “Investigations in real-time Loudness Metering”, 110th Convention of the Audio Engineering Society, Paris, May 2006, Preprint 6822

[10] Skovenborg, E., "Loudness Range (LRA) - Design and Evaluation", 132nd Convention of the Audio Engineering Society, Budapest, April 2012, Preprint 8616.

[11] Gonzalez E. P. & Reiss, J., 'Automatic gain and fader control for live mixing.", IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, October 2009.

[12] Mansbridge S., Finn S., & Reiss J., "Implementation and evaluation of autonomous multitrack fader control", 132nd Convention of the Audio Engineering Society, April 2012, Preprint 8588.

[13] Pestana P., "Automatic Mixing Systems Using Adaptive Digital Audio Effects" PhD, Universidade Catolica Portuguesa, 2013.

[14] ITU-R BS.1770-4, Algorithms to measure audio programme loudness and true-peak audio level, International Telecommunications Union, Geneva, Switzerland, 2012