REAL-TIME EXCITATION BASED BINAURAL LOUDNESS METERS

Dominic Ward, Sean Enderby, Cham Athwal

DMT Lab
Birmingham City University
Birmingham, UK
dominic.ward@bcu.ac.uk

Joshua D. Reiss

C4DM
Queen Mary University of London
London, UK
josh.reiss@eecs.qmul.ac.uk

ABSTRACT

The measurement of perceived loudness is a difficult yet important task with a multitude of applications such as loudness alignment of complex stimuli and loudness restoration for the hearing impaired. Although computational hearing models exist, few are able to accurately predict the binaural loudness of everyday sounds. Such models demand excessive processing power making real-time loudness metering problematic. In this work, the dynamic auditory loudness models of Glasberg and Moore (J. Audio Eng. Soc., 2002) and Chen and Hu (IEEE ICASSP, 2012) are presented, extended and realised as binaural loudness meters. The performance bottlenecks are identified and alleviated by reducing the complexity of the excitation transformation stages. The effects of three parameters (hop size, spectral compression and filter spacing) on model predictions are analysed and discussed within the context of features used by scientists and engineers to quantify and monitor the perceived loudness of music and speech. Parameter values are presented and perceptual implications are described.

1. INTRODUCTION

The need to measure perceived loudness is imperative within fields such as psychoacoustics and audio engineering. In particular, accurate loudness alignment of complex stimuli is crucial when conducting controlled listening experiments and configuring multichannel systems [1]. In broadcasting, the loudness of a wide range of program material must be consistent yet natural to maintain a comfortable listening experience. In recent years, a number of researchers in the automatic mixing community [2, 3] have developed systems to automatically balance the loudness of multi-track content according to perceptual loudness features extracted from the audio.

Although there is a clear need for models of loudness, developers of real-time applications are often forced to sacrifice prediction accuracy. The purpose of this paper is to demonstrate how the dynamic loudness models proposed by Glasberg and Moore [4] and Chen and Hu [5] can be modified to obtain fast and efficient estimates of perceived loudness whilst maintaining agreement with empirical data. Before describing the models in detail, background information on single and multiband approaches to loudness prediction is given.

1.1. Loudness Models

Following the ITU broadcast standard [6] first published in 2006, subsequent EBU recommendation [7] and metering specifications [8] and related standards [9], a number of commercial loudness meters have appeared over the past few years. These meters employ single band loudness models which consist of a frequency weighting function, simulating the transmission response of the outer and middle ear, followed by an energy integrator. Such efficient estimators have proven to work well on broadcast material [10] and have been extended to capture the perception of single instruments [11] making them attractive candidates for real-time meters [12].

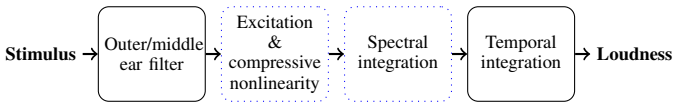

Figure 1: Block diagram of a simplified loudness model. Dotted boxes indicate stages used by multiband models.

In contrast to single band loudness models, sophisticated multiband models founded on empirical measurements of auditory filters and excitation patterns [13, 14, 15] have received little interest outside of psychoacoustics. As shown in Figure 1 multiband models also correct for the response of the outer and middle ear, but go a step further by accounting for the frequency selectivity of the cochlea. The signal is decomposed into frequency bands by means of an auditory filter bank. The output of the filter bank is called an excitation pattern and approximates the distribution of energy on the basilar membrane. When plotted as a function of filter centre frequency, the frequency scale is transformed to a perceptual scale with units Cams [16]. A compressive nonlinearity, modelled either as part of the filter bank or separately, describes the active mechanism of the auditory system. The specific loudness (SL) pattern represents compressed intensity as a function of filter centre frequency with units sones/Cam, where sones is the unit of loudness [17].

The spectral integration stage calculates the area under the SL pattern which estimates the instantaneous loudness (IL). This is the key aspect of multiband models that differentiate them from their single band counterpart. For narrow-band noise of constant intensity, loudness increases once its spectral bandwidth exceeds a critical bandwidth [13]. Single band models cannot account for this phenomenon known as spectral loudness summation. For time-varying sounds, the IL is smoothed by either a low-pass filter or empirically derived sliding window. Dynamic models output a loudness time series, from which various features can be computed, such as average or peak loudness. In short, multiband models have been shown to explain the observed variations in loudness caused by experimental factors such as sound intensity, frequency, spectral bandwidth and masking [18, 19].

The primary drawback of multiband models is their computational demand which consequently limits their application. This work explores this problem by optimising the excitation based loudness models of Glasberg and Moore [4] and Chen and Hu [5],

referred to as the GM02 and the CH12 respectively. Both models are refinements of Zwicker's stationary-sound loudness model [18] and have been extended to deal with time-varying sounds (see [20] for a detailed historical review). The key difference between Zwicker's and Glasberg and Moore's procedure is the way in which excitation patterns are computed and the equations defining the critical bandwidth as a function of frequency. The procedure used by the GM02 is based on more recent measures of auditory filters [14] and excitation patterns are directly derived from the output of the filter bank [21]. The revised model for steady-state sounds accurately predicts absolute thresholds, equal loudness contours and binaural loudness and became the basis for the 2007 ANSI standard [22].

The model of Chen and Hu is important because unlike its predecessors, the active process within the cochlea is modelled as an integral part of the filter bank [23] - there is no transformation from excitation to specific loudness. This is consistent with the idea that frequency selectivity and cochlear compression are the result of a single active process [24]. The authors of this recent loudness model also made improvements to the spectral decomposition procedure specified by Glasberg and Moore to correct for inflated loudness estimates.

1.2. Faster Multiband Models

Efficient implementations based on the core excitation model used by the GM02 have been established. In [25] a real-time loudness meter was developed with optimisations at the excitation transformation stage for increased efficiency. Although this model accounted for the loudness of short duration sounds it was less accurate in predicting the loudness of amplitude-modulated sounds compared to the GM02 as discussed in [26]. A binaural loudness device was established by [27] but the temporal integration and binaural summation stages were simple approximations when compared to more recent techniques [4, 28]. Other authors have concentrated on fast calculation of excitation patterns by using pruning techniques [29, 30] or making use of nonuniform spectral sampling [31, 32]. These proposals were specific to the GM02 and it is not clear how flexible they are, e.g. when a specified error limit is to be achieved. Finally, Burdiel et al. [33] discussed the computation savings of the GM02 obtained through parameterisation. Importantly, they demonstrated that reducing the number of analysis bands gave minimal impact on loudness estimates of musical sounds when compared to the effects of other parameters. Although real-time performance was obtained, that implementation was limited to monaural sounds and again specific to the GM02.

A full binaural dynamic loudness meter incorporating more recent loudness theory has not yet been established. It is the purpose of this study to present efficient parameterised implementations of both the GM02 and CH12 to facilitate real-time binaural loudness metering, with a focus on analysing the errors introduced. Additionally, the models have been realised as software plugins for use in digital audio workstations[1]

An overview of both models and implementational details of the meter is given in Section 2. In Section 3, the procedure used to analyse the performance of the models is provided and results are outlined in Section 4. Section 5 discusses key findings and draws comparisons to previous approaches, before concluding remarks and suggestions for future work are given in Section 6.

[1]Code available at https://github.com/deeuu/LoudnessMeters.

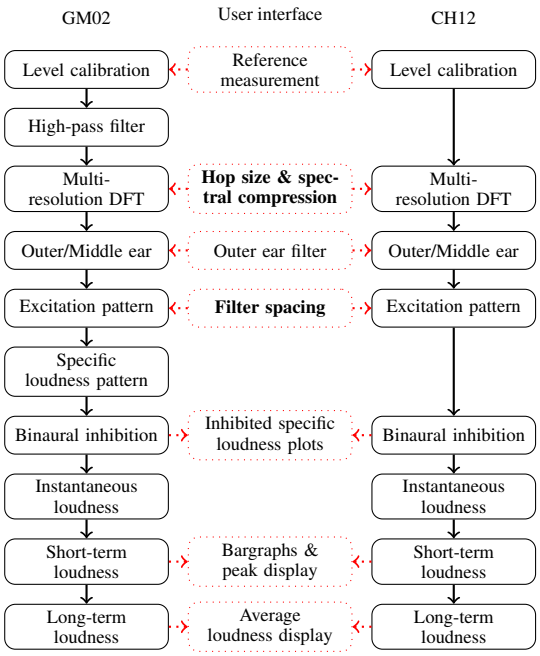

Figure 2: Block diagram of the loudness models and meter interface. Bold text indicates parameters investigated in this study.

2. MODELS AND METER IMPLEMENTATION

In this section the GM02 is described first, followed by the CH12. Developments for the meter are subsequently outlined before performance profile reports are summarised. The structure of both loudness models within the context of the binaural meter is given in Figure 2. The time-domain input signal can be either from a microphone, head and torso simulator (HATS) or digital recording (mono or stereo). A calibration stage is required to anchor the signal to a known reference. In order to reduce both latency and computational load, a third order Butterworth high-pass filter is applied to approximate the low-frequency response of the middle ear and the remaining outer and middle-ear filtering is conducted in the frequency domain [32]. This is in contrast to the 4096 order FIR used in the original model [4].

Short-term multi-resolution power spectra are obtained via six parallel FFTs. The six segments are obtained using Hann windows, with each successive window half the length of the previous. The windows are aligned at their centres via zero-padding. The FFTs are updated every millisecond, but the window hop size (time-step) was exposed as a free parameter. A spectral compression stage was added to reduce the number of components in the composite spectrum. This achieves a compact power spectrum by summing components into composite bins, the width of which increases with frequency. More specifically, the algorithm tries to maintain a constant spacing between components on the Cam scale. This reduces the number of components significantly in order to simplify the computation of excitation patterns as suggested in [25] and is similar to the enhancement used in the GM02 implementation by [34].

The power spectrum is then weighted according to the presentation of the stimulus [22]: free-field, diffuse-field or middle ear only (HATS). An additional BeyerDynamic DT990 option is available for simulating headphone presentation. The transfer function

of the headphone capsules were measured on an artificial ear.

The power spectrum is transformed to an excitation pattern using a bank of rounded exponential (roex) filters equally spaced from 1.8 to 38.9 Cams in steps of 0.1, yielding a total of 372 filters [22]. The shape of the filters are level dependent and have to be re-evaluated on every frame. In our implementation, filter shapes are computed according to [15], but a lookup table is employed to save on computing exponentials. As with hop size and spectral compression criterion the filter spacing was also made variable.

Following [22], a separate stage is required to transform the excitation pattern to an SL pattern, which represents compressed intensity within the cochlea. Unlike the published version of the model which approximates overall binaural loudness by summation of loudness across the ears, the procedure of [28] was incorporated to improve estimates of loudness for stimuli presented binaurally by modelling inhibitory interactions between the ears. The binaural inhibition block produces an inhibited SL pattern for each ear.

The IL in each ear is given by summing the inhibited SL values. The overall IL is the sum of the ILs at each ear, which is then smoothed by two cascaded asymmetrical low-pass filters to give the short-term loudness (STL) and long-term loudness (LTL) respectively. The time constants are given in Table 1 in accordance with [4] and [5].

Table 1: *Time constants (seconds) used by the two models.*

Model	STL		LTL	
	τ_A	τ_R	τ_A	τ_R
GM02	0.022	0.050	0.1	2
CH12	0.016	0.032	0.1	2

The CH12 differs from the GM02 in the following ways:

1. It uses twice the frequency resolution (largest window size of 128 ms compared to 64 ms for the GM02).
2. The pre-cochlear filter is performed entirely by weighting the power spectrum.
3. The middle ear transfer function follows [23].
4. It uses a set of double roex filters [23] equally spaced from 1.5 to 40.2 in steps of 0.1 Cams (388 filters).
5. No SL transformation is required.

Chen and Hu decided to double the frequency resolution because the DFT specification reported by Glasberg and Moore resulted in a spectral bandwidth that exceeded the critical bandwidth at 1 kHz. Consequently, the predicted total loudness of pure tones was larger than expected. Unlike the GM02, the filter bank used by the CH12 incorporates a compressive nonlinearity and thus the area under the *excitation* pattern is proportional to loudness. The output of the filter bank was scaled by the constant of proportionality in order to arrive at the SL pattern, which can then be fed to the binaural inhibition stage. Although the binaural procedure has been integrated into the GM02 [20], our modification to the CH12 to account for binaural loudness requires validation.

2.1. Modifications Specific to the Meter

As stated above, the pre-cochlear filter of the original GM02 has been modified to simplify computation and a binaural inhibition procedure has been incorporated into both models. The loudness meter displays inhibited SL on a logarithmic scale as a function of frequency on the Cam scale. Both left and right patterns are shown to facilitate visual comparison. In addition, bargraphs have been added to show the STL (in sones) for each ear, which are useful for comparing and aligning loudness between the ears. Finally, the overall LTL and peak STL are displayed for purposes of measuring global binaural loudness. A third number box shows the average unweighted sound pressure level (SPL) which is useful for system calibration.

The user interface, shown in Figure 3, also features two additional screens which can be activated using the appropriate buttons. The first is a settings screen for configuring the models in accordance with the exposed parameters highlighted in bold in Figure 2. The second is a calibration window, allowing the user to enter SPL measurements and estimate calibration gains.

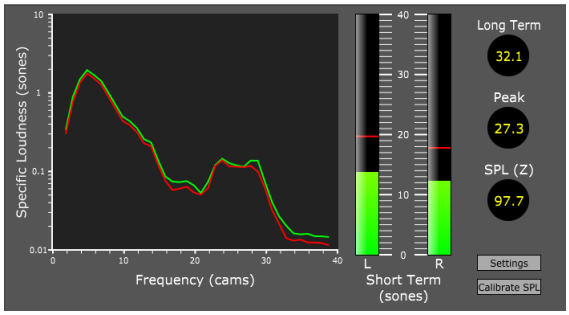

Figure 3: *Loudness meter user interface.*

2.2. Profiling the Models

The three parameters that govern the speed-accuracy trade-off are: hop size (R), spectral compression criterion (α) and filter spacing (β). A program to extract the loudness from 30 stereo sounds was developed for both models, each configured with their respective default parameter set (see Section 3). The sample-based profiler OProfile[2] was used to identify the most computationally intensive parts of each model. The programs were run on an idle Intel Core 2 Quad CPU Q8200 operating at 2.33 GHz with 4 GB RAM running Ubuntu 14.04 over 10 executions to increase the sample size. The profile reports are summarised in Table 2.

Table 2: *Program profile report - absolute sample count (and percentage of total samples collected).*

Process	GM02	CH12
Excitation transformation	2.7×10^8 (87.3)	1.3×10^8 (67.2)
FFTW [35]	1.4×10^7 (4.6)	3.4×10^7 (17.6)
Binaural Inhibition	8.1×10^6 (2.6)	8.8×10^6 (4.6)

For both models, the majority of CPU time was spent at the excitation transformation stage. Compared to the CH12, the GM02 collected a larger number of samples in this module, likely due to the added overhead of calculating variables pertaining to the roex filter shapes. Furthermore, the number of frequency points used by the CH12 is twice that of its predecessor and thus the time to compute the multi-resolution DFT is greater. The time required to compute the inhibited SL values was longer for the CH12, which can be attributed to the use of slightly more auditory filters.

[2] http://oprofile.sourceforge.net/

Importantly, the time complexity of the stages involved in excitation transformation of both models are linear functions of the number of frequency components (N) and auditory filters (M). For both models, the inner product of the auditory filters and power spectrum has the highest complexity of the excitation transformation module being $O(NM)$ (see [30]). Thus, although increasing the hop size reduces overall process time by computing fewer estimates, both spectral compression and filter spacing parameters target the bottleneck directly.

3. EVALUATION PROCEDURE

For each model, the parameter values were varied to reduce the computational workload and the errors between the reference and approximations were examined. The values used for each reference model and the approximations are given in Table 3 using Matlab notation ([start: delta: end]). The models were configured to process stereo recordings using the combined outer ear (free-field) [22] and middle ear [22, 23] transfer functions.

Table 3: *Reference and evaluation parameter values.*

Parameter	Reference	Approximation
R	1 ms	[2:2:16] ms
α	NA	[0.1:0.1:1] Cams
β	0.1 Cams	[0.25:0.25:4] Cams

The parameters were varied both independently and in combination. The effect of filter spacing was studied with and without interpolation applied to the excitation pattern. From experimenting with various input signals and interpolation schemes, cubic spline interpolation applied to the log excitation pattern was found to capture the smooth shape of the reference high-resolution patterns with good accuracy.

3.1. Stimuli

Each of the configurations were evaluated by extracting loudness features from 30 stereo tracks taken from the Sound Quality Assessment Material (SQAM) database [36] and assessing the error introduced. All tracks were sampled at 44.1 kHz. Each recording belonged to one of the following stimulus categories: single instruments, vocal, speech, solo, orchestra and pop music. Representative sound segments of the selected tracks were edited manually (average duration of 4.4 s (standard deviation (s.d.) 0.5 s)) and then peak calibrated to a random level between 84 and 94 dB SPL giving a spread of typical listening levels across the 30 sources (average RMS of 72.2 dB SPL (s.d. 4.4 dB) and peak of 89.1 dB SPL (s.d. 2.7 dB)).

3.2. Errors

The primary interest was to investigate the error between the reference and estimated short-term loudness time series of a given stimulus, denoted STL and $\hat{STL}$ respectively. The normalised root-mean-square error (nRMSE) between the reference and approximation is given by

$$nRMSE_{STL} = \frac{1}{STL_\mu}\sqrt{\frac{1}{F}\sum_{f=1}^{F}(STL_f - \hat{STL}_f)^2}, \quad (1)$$

where F is the number of non-zero frames and STL_μ is the average of the reference loudness. This normalisation was used to obtain a scale-free evaluation metric and corresponds to the coefficient of variation of the RMSE [37].

The following metric was used to measure the error introduced in the specific loudness patterns,

$$nRMSE_{SL} = \frac{1}{SL_\mu}\sqrt{\frac{1}{F\times M}\sum_{f=1}^{F}\sum_{m=1}^{M}(SL_{f,m} - \hat{SL}_{f,m})^2}, \quad (2)$$

where $SL_{f,m}$ is the total specific loudness in auditory filter m obtained by summing the corresponding inhibited specific loudness values in both ears. The RMS error is normalised by the average specific loudness value over time. In order to calculate the error at hop sizes greater than 1 ms a sample and hold procedure was applied to the approximated STL and SL time series.

It is also insightful to explore the perceptual implications of model performance for different parameter sets. For each stimulus, the level change (ΔL) required for equal overall loudness between the reference and approximated loudness predictions was estimated using the sone ratios obtained from three global loudness descriptors: average LTL (LTL_μ) [4], peak STL (STL_{pk}) [38] and the 95th percentile of the STL distribution (STL_{95}) [39]. Because of the nonlinear relationship between intensity level and loudness, an iterative procedure was employed to find the optimal gain required for equal loudness. More specifically, ΔL is obtained for each descriptor by minimising the sone ratio in decibels using a tolerance of 0.01 dB. For example, the level change required for equal loudness according to the peak STL is found by minimising the magnitude of

$$\epsilon = 10 \times \log_{10}\left(\frac{STL_{pk}}{\hat{STL}_{pk}}\right), \quad (3)$$

where STL_{pk} and $\hat{STL}_{pk}$ is the peak STL of the reference and approximation. For convenience, ΔL is referred to as the global loudness descriptor error and is expressed in decibels.

In addition to profiling the models, the computational savings are also analysed in terms of the reduction in the number of frequency components and auditory filters with respect to the reference quantities.

4. RESULTS

Figures 4, 5 and 6 show the nRMSE for parameters hop size (R), spectral compression criterion (α) and filter spacing (β). The base 10 logarithm of the stimulus errors (see ordinate) was taken in order to reduce positive skew in the error distributions (and thus biassing the arithmetic mean), as well as providing visual clarity throughout the plots. Data points are the arithmetic means of the log nRMSE across stimuli for either the STL or SL. The shaded areas surrounding the lines are the 95% confidence intervals of the means and were estimated using a percentile bootstrapping procedure based on 5000 samples [40]. These intervals have been corrected to eliminate between-stimuli variability which bias the sampling error in repeated-measures designs and are useful for observing patterns in the population means [41]. For example, the intervals shown in subplot (a) of Figure 6 were calculated according to a 2 (models) x 2 (interpolants) x 16 (filter spacings) within-stimulus design.

4.1. Hop Size

Increasing the hop size widens the interval between instantaneous loudness samples, giving rise to greater error in the time integrated

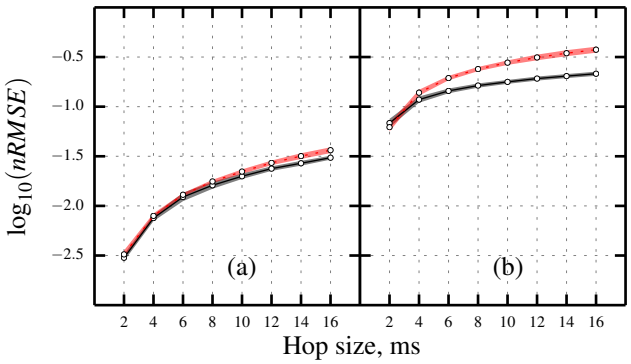

Figure 4: *Log normalised RMSE plotted as a function of hop size for (a) the STL and (b) SL. The data points are the arithmetic means of the log stimulus errors and shaded areas surrounding the lines represent the 95% confidence intervals of the estimates for the GM02 (black solid lines) and CH12 (red dotted lines).*

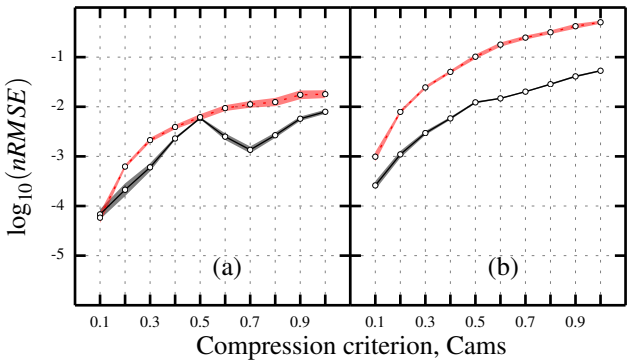

Figure 5: *Log normalised RMSE plotted as a function of compression criterion for (a) the STL and (b) SL. The data points are the arithmetic means of the log stimulus errors and shaded areas surrounding the lines represent the 95% confidence intervals of the estimates for the GM02 (black solid lines) and CH12 (red dotted lines).*

measurements. Although the STL error functions show a similar trajectory for both models, the errors appear to increase at a slightly faster rate for the CH12. A doubling of the hop size increases the geometric mean of the STL nRMSEs by an approximate factor of 2.1 for the GM02 and 2.2 for the CH12. Subplot (b) of Figure 4 shows the error introduced when SL patterns from previous processing frames are used as estimates for the current frame output by the reference model. The SL errors are higher than STL errors because they are instantaneous calculations whereas the STL is the combined result of integrating the SL patterns both in frequency and over time. For a hop size of 2 ms, the SL error is lower for the CH12, but rises above the GM02 at larger values, indicating that the CH12 sees greater variation in the auditory patterns over time.

4.2. Compression Criterion

For both models, spectral compression distorts the loudness patterns as shown in subplot (b) of Figure 5. This is to be expected, given that the input components are summed into sub-bands prior

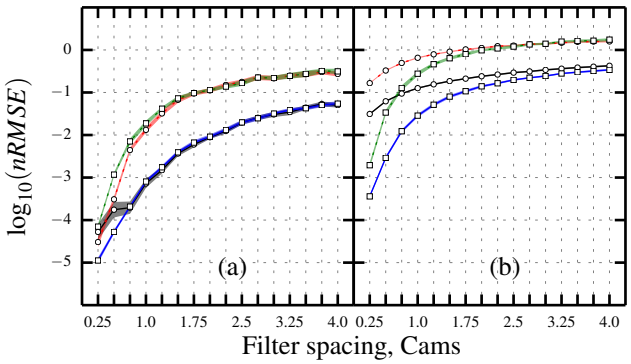

Figure 6: *Log normalised RMSE plotted as a function of filter spacing for (a) the STL and (b) SL. The data points are the arithmetic means of the log stimulus errors and shaded areas surrounding the lines represent the 95% confidence intervals of the estimates for the GM02 (black and blue solid lines) and CH12 (red and green dotted lines). The lines with square markers are the errors introduced when the approximated excitation patterns are interpolated at filter locations defined by the reference models.*

to auditory filtering. In general, the CH12 is more sensitive to spectral modifications and gives consistently larger errors in the SL domain where the average ratio of the nRMSE geometric means between the CH12 and GM02 was 9.1. The only criterion in which this model produced a lower average error (STL only) was 0.1 Cams although there is strong overlap between the two confidence intervals. For the GM02, summing the erroneous SL pattern leads to a non-monotonic STL error function; the error function shows a peak at 0.5 Cams and continues to increase above 0.7 Cams. This is an important consideration when optimising the GM02 based on spectral compression.

4.3. Filter Spacing

Figure 6 shows the STL and SL errors as a function of filter spacing, with interpolation type as parameter. With no interpolation, doubling the filter spacing increases the geometric mean of the STL nRMSEs by an approximate factor of 7.0 for the GM02 and 9.9 for the CH12, which is considerably higher than the relative increase in error when doubling the hop size. The STL error functions of the CH12 are especially steeper than those of the GM02 over 0.25-0.75 Cams, indicating that the former is more sensitive to modifications in this range. The only spacing in which the STL error of the CH12 was lower than the GM02 was 0.25 Cams.

The geometric mean of the STL nRMSEs, averaged across 0.25 and 0.5 Cams with cubic interpolation, was reduced by 74% for the GM02, though no such improvement was observed for the CH12. It is important to highlight that the nRMSE is a relative error metric; plotted data points below -3 translate to central RMS errors less than 0.1% of the mean loudness time series. The application of cubic interpolation to the SL patterns show an improvement in the approximations, though the benefit does not carry over to the STL measurements for the CH12.

4.4. The Bottleneck

The STL and SL errors have been presented for model parameters hop size (R), spectral compression criterion (α), and filter spac-

Table 4: *Number of components and filters (and percentage reduction) obtained with the optimal combination of compression criterion (α) and filter spacing (β) at $R = 1$ ms. The maximum global loudness descriptor error is also given.*

Model	α	$\hat{N}$	β	$\hat{M}$	$\max \lvert\Delta L\rvert$ (dB)
GM02	0.7	63 (95 %)	1.25	30 (92 %)	0.08
CH12	0.3	148 (95 %)	0.5	78 (80 %)	0.11

Table 5: *Program profile report after optimisations.*

Process	GM02	CH12
FFTW [35]	1.4×10^7 (73.7)	3.4×10^7 (78.9)
Excitation transformation	1.6×10^6 (8.1)	1.6×10^6 (3.8)
Binaural inhibition	1.1×10^5 (0.6)	4.7×10^5 (1.1)

ing (β). When configured with the reference parameter sets (see Table 3), the models did not execute in real-time. The computational speed can be increased approximately two-fold by simply doubling the hop size, however Table 2 shows that the bottleneck resides in the excitation transformation stage. The findings indicate that initial values of the hop size introduce significantly larger STL error when compared to a range of α and β values.

Can a speedup factor greater than two with a lower average STL error than that introduced by a hop size of 2 ms be achieved by targeting the bottleneck directly?

The number of frequency components (N) was 1393 and 2971 for the GM02 and CH12. The number of auditory filters (M) was 372 and 388. Let $\hat{N}$ and $\hat{M}$ denote the number of components and filters resulting after increasing parameters α and β beyond the reference values. For each model, the error surface generated by different combinations of α and β (with $R = 1$ ms) was searched for all geometric average STL errors less than the error introduced by $R = 2$ ms alone. The target error was 0.3% for both models. All combinations which lead to average errors exceeding this threshold were discarded. Of the remaining parameter values, the combination that maximised the complexity reduction at the excitation transformation stage ($1 - (\hat{N}\hat{M})/(NM)$) was selected.

Table 4 shows the performance of the models with the optimal parameter sets, operating with a hop size of 1 ms. The average STL nRMSEs were less than 0.25% and lead to a total complexity reduction of approximately 99% for both models. Indicative of worst-case performance, the final column shows the maximum absolute level change required for equal loudness between the reference and approximation predictions across all stimuli for all three features. The largest deviation from the global loudness predictions given by the reference models was 0.08 dB for the GM02 and 0.11 dB for the CH12.

The profile reports associated with the optimal parameter sets are shown in Table 5. It can be seen that the processing time consumed by the excitation transformation has been significantly reduced (by two orders of magnitude for the GM02) and consequently, the bottleneck has shifted to the computation of the multiresolution DFT. The mean stimulus speedup factor (CPU time of the reference divided by the CPU time of the approximation averaged across ten executions) was 16.09 (s.d. 0.01) for the GM02 and 4.55 (s.d. 0.02) for the CH12. This demonstrates that large speedups can be achieved before having to resort to incrementing the hop size independently. It should be noted that these parameters were selected based on a target STL nRMSE at $R = 2$ ms. Running the same optimisation procedure using the LTL yields slightly lower values for α and β because this feature is less affected by hop size and thus lowers the target error to 0.1% for both models.

Table 6: *Performance of the optimised models at four hop sizes (R) in terms of maximum global loudness descriptor error and minimum speedup factor across all stimuli. The first and second row of each cell correspond to the GM02 and CH12 respectively. The row in bold indicates the only parameter set in which real-time processing was not achieved.*

R	$\max \lvert\Delta L\rvert$ (dB)			Speedup
	LTL_μ	STL_{pk}	STL_{95}	
1	0.05	0.08	0.08	15.4
	0.06	**0.03**	**0.11**	**4.4**
2	0.05	0.07	0.10	30.5
	0.06	0.03	0.11	8.7
4	0.12	0.81	0.22	63.2
	0.08	0.21	0.20	17.7
8	0.29	2.57	0.24	124.8
	0.12	1.54	0.42	34.4

Finally, maximum level differences required for equal loudness readings across all sounds according to three global loudness descriptors are given in Table 6 for the two models configured with the parameter values listed in Table 4. The performance is evaluated at four hop sizes and worst-case speedup factors across all program executions and all stimuli are given. The row in bold indicates that the CH12 did not achieve real-time performance on the computer used in this study at the reference hop size.

5. DISCUSSION

The two dynamic loudness models presented in this study have been parameterised by hop size, spectral compression and filter spacing. The effect of hop size on STL error was comparable in both models and a hop size of 2 ms introduced a larger STL error than using the lower compression and filter values tested with a hop size of 1 ms. As shown in Table 6, large hop sizes can be detrimental when estimating the loudness of short duration sounds, which rely on accurate estimates of peak loudness [4]. Zwicker [38] highlighted the importance of maximum loudness for quantifying the perceived loudness of impulse and speech sounds. Indeed, of all the stimuli tested, the castanets recording (track 27 of the SQAM CD) occurred most frequently in the top 5% of the STL nRMSE distributions across all hop sizes tested for both models. Compared to the peak STL descriptor, the LTL and percentile measures stay within 0.5 dB of the reference values for hop sizes up to 8 ms.

A compressed spectrum can be incorporated into both models to reduce the number of components whilst attempting to maintain sufficient energy in the auditory filters to preserve the loudness density. The grouping of component intensities does introduce some error in the auditory patterns, the extent of which is dependent on the bandwidth used to average the spectrum. The GM02 requires a wider bandwidth to achieve a similar percentage reduction in the number of components as the CH12. This is because the GM02 uses half the frequency resolution as the CH12. When $\alpha = 0.1$ the compression criterion is satisfied above 264 Hz for the CH12 but not until 764 Hz for the GM02. In combination with the fact that most musical sounds are dominant in the low-mid frequency range, this explains why the average errors of the CH12 shown in Figure 5 are generally higher.

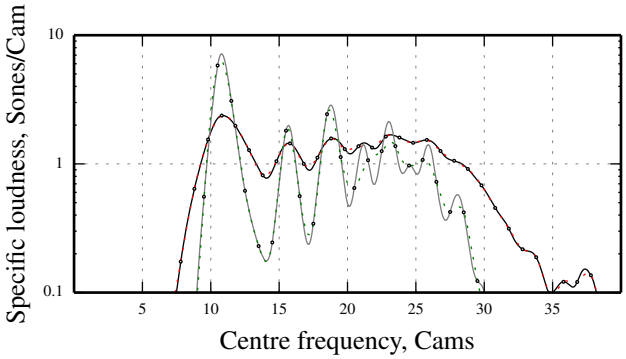

Figure 7: *Specific loudness pattern of a segment of track 42 (Accordion) from the SQAM CD using the GM02 (solid black line) and CH12 (solid grey line) with the reference parameter sets. Data points correspond to values obtained by sampling at increments of 1 Cam and dotted lines show the result of cubic interpolation.*

The choice of filter spacing determines the accuracy of the excitation and SL patterns. Too few filters per critical band can be severe for tonal signals where peak excitations that dominate perception are unlikely to be captured. Large intervals can also have a marked effect on broadband sounds in that the area under the SL pattern is estimated from an undersampled excitation pattern. The results showed that the CH12 is more sensitive to changes in filter spacing and introduced larger deviations from the target data compared to the GM02. This can be attributed to differences in DFT frequency resolution and filter bank architecture. For the CH12, the DFT mainlobe width is narrower and so the importance of filter spacing is greater. This issue was investigated by running the GM02 with twice the frequency resolution at multiple filter spacings (with the reference reconfigured) and indeed larger errors resulted. The SL patterns output by the two reference models were also compared across stimuli and it was observed that those of the CH12 had much larger peak-to-valley ratios compared to the GM02 (see Figure 7). The excitation pattern must be sampled more frequently to accurately capture peak excitations. Considering that the RMSE metric emphasises large deviations, it is of no surprise that the SL errors produced by the CH12 are notably high. As shown in Figure 7, large excitatory oscillations generated by the CH12 can be problematic when interpolating low-resolution patterns. For this particular analysis frame, the instantaneous loudness predicted by the reference GM02 was 30.86 sones and, with cubic interpolation applied to the excitation pattern sampled at 1 Cam intervals, 30.87. The CH12 predicted 25.41 sones and, for the same approximation, 24.68. Although the CH12 did not benefit from interpolation in terms of the integrated loudness, Figure 7 indicates that cubic interpolation should be preferred over linear interpolation when displaying SL versus frequency.

The parameter sets listed in Table 4 provide a good guideline for speeding up the models based on the initial bottleneck. Both models see a significant reduction in execution time by combining spectral compression with filter spacings above 0.1 Cams. Further speedup factors of ~ 2 can be obtained by doubling the hop size. The results in Table 6 indicate that for a hop size of 2 ms, combined with the parameters values given in Table 4, real-time performance can be achieved with estimates of global loudness within 0.15 dB of the reference predictions for a range of music and speech. Considering that discrimination thresholds for intensity can be as low as 0.2 dB for pure tones and 0.5 dB for most broadband noises [39], it is unlikely that there will be noticeable differences between identical stimuli aligned in loudness by the reference models and proposed approximations. More importantly, [42] found the average reproducibility of subjective relative loudness judgements involving *different* program material to be 1.24 dB. This suggests that greater error may be tolerable when the faster implementation is used for purposes of balancing the loudness of typical program material.

The authors in [33] showed that, based on STL errors of the GM02, a filter spacing of 1 Cam could be used to achieve real-time performance on their test machine. In the current study, a finer range of parameter values have been explored and for the GM02, greater computational savings and lower average STL nRMSE can be obtained using an interval of 0.75 Cams combined with a compression criterion of 0.2. Furthermore, the real-time proposal in [33] included a frequency domain weighting function to replace the FIR filter used by Glasberg and Moore. However, it has been shown [32] that spectral weighting is problematic at low-frequencies, mainly in terms of absolute threshold predictions, and hence why the implementation here incorporates a high-pass filter to improve the response. Our implementation can be applied to both laboratory and everyday sounds without having to switch filtering techniques. Finally, the authors in [31] employed the Hopping Goertzel DFT algorithm to optimise the parallel FFTs, which was then used by [32] to simplify the calculation of excitation patterns by means of nonuniform spectral sampling. Although faster performance can be obtained with that implementation at very low hop sizes, larger time-steps are unlikely to yield substantial computational gains due to the processing requirements of the Hopping Goertzel DFT. Preliminary experiments showed that for the GM02, nonuniform spectral sampling introduced larger average STL error compared to all compression criteria tested in this study.

6. CONCLUSIONS

Efficient implementations of the loudness models of Glasberg and Moore [4] and Chen and Hu [5] have been developed for the purpose of real-time binaural loudness metering. The meter exposes key parameters that govern model performance, enabling the user to control the speed-accuracy trade-off to meet the demands of a given application. The bottleneck of both models was identified as the transformation from power spectrum to excitation pattern. By incorporating a perceptually inspired method to obtain a compact spectrum and experimenting with different filter spacings, the total complexity of the excitation transformation stage was reduced by 99%, yielding significant speedup in execution time. The largest deviation was 0.11 dB when measured in terms of short-term and long-term loudness metrics commonly used to quantify overall loudness of time-varying sounds. The effect of hop size was also investigated both independently and in combination with the parameter sets that lead to high computational savings. In the latter case, a hop size of 2 ms was required to achieve real-time performance.

Future work is required to validate the predictions of the binaural models against empirical data, especially on sounds with complex spectro-temporal behaviour. Derivation of parameter sets that achieve maximum speedup whilst maintaining error limits to match required perceptual criteria would also be useful. Our meter does not currently simulate the effects of cross-talk between loudspeakers in multichannel setups typically used by sound engineers. This would require a more sophisticated filtering stage that accommodates appropriate head-related transfer functions.

7. REFERENCES

[1] N. Zacharov, "An Overview of Multichannel Level Alignment," in *Proceedings of the 15th AES International Conference on Audio, Acoustics & Small Spaces*, 1998.

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

[3] D. Ward, J. D. Reiss, and C. Athwal, "Multi-track mixing using a model of loudness and partial loudness," in *Proceedings of the 133rd Audio Engineering Society Convention*, 2012.

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

[5] Z. Chen and G. Hu, "A Revised Method of Calculating Auditory Excitation Patterns and Loudness for Time-varying Sounds," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2012.

[6] ITU, "ITU-R BS.1770-3: Algorithms to measure audio programme loudness and true-peak audio level," Tech. Rep., International Telecommunication Union, 2012.

[7] EBU, "EBU-R 128: Loudness normalisation and permitted maximum level of audio signals," Tech. Rep., European Broadcast Union, 2014.

[8] EBU, "EBU-Tech 3341: Loudness Metering: 'EBU Mode' metering to supplement loudness normalisation in accordance with EBU R 128," Tech. Rep., European Broadcast Union, 2011.

[9] ATSC, "ATSC Recommended Practice: Techniques for Establishing and Maintaining Audio Loudness for Digital Television," Tech. Rep., Advanced Television Systems Committee, 2013.

[10] E. Skovenborg and S. H. Nielsen, "Evaluation of Different Loudness Models with Music and Speech Material," in *Proceedings of the 117th Audio Engineering Society Convention*, 2004.

[11] P. D. Pestana, J. D. Reiss, and A. Barbosa, "Loudness Measurement of Multitrack Audio Content using Modifications of ITU-R BS.1770," in *Proceedings of the 134th Audio Engineering Society Convention*, 2013.

[12] E. Skovenborg and S. H. Nielsen, "Real-time Visualisations of Loudness Along Different Time Scales," in *Proceedings of the 10th International Conference on Digital Audio Effects*, 2007.

[13] E. Zwicker, G. Flottorp, and S. S. Stevens, "Critical Band Width in Loudness Summation," *Journal of the Acoustical Society of America*, vol. 29, no. 5, pp. 548–557, 1957.

[14] R. D. Patterson, "Auditory Filter Shapes Derived with Noise Stimuli," *Journal of the Acoustical Society of America*, vol. 59, no. 3, pp. 640–654, 1976.

[15] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research*, vol. 47, pp. 103–138, 1990.

[16] W. Hartmann, *Signals, Sound and Sensation*, Springer, 1997.

[17] S. S. Stevens, "The Measurement of Loudness," *The Journal of the Acoustical Society of America*, vol. 27, no. 5, pp. 815–829, 1955.

[18] E. Zwicker and B. Scharf, "A Model of Loudness Summation," *Psychological review*, vol. 72, no. 1, pp. 3–26, 1965.

[19] B. C. J. Moore and B. R. Glasberg, "A Revision of Zwicker's Loudness Model," *Acta Acustica united with Acustica*, vol. 82, no. 2, pp. 335–345, 1996.

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

[21] B. C. J. Moore and B. R. Glasberg, "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," *Journal of the Acoustical Society of America*, vol. 74, no. 3, pp. 750–753, 1983.

[22] ANSI, "ANSI S3.4-2007 Procedure for the Computation of Loudness of Steady Sounds," Tech. Rep., American National Standards Institute, 2007.

[23] Z. Chen, G. Hu, B. R. Glasberg, and B. C. J. Moore, "A new method of calculating auditory excitation patterns and loudness for steady sounds," *Hearing research*, vol. 282, no. 1-2, pp. 204–15, 2011.

[24] L. Robles and M.A. Ruggero, "Mechanics of the mammalian cochlea," *Psychological review*, vol. 81, pp. 1305–1352, 2001.

[25] M. A. Stone, B. C. J. Moore, and B. R. Glasberg, "A Real-Time DSP-Based Loudness Meter," in *Contributions to Psychological Acoustics*, A. Schick and M. Klatte, Eds., pp. 587–601. Bibliotheks- und Informationssystem der Universität Oldenburg, 1997.

[26] B. C. J. Moore, B. R. Glasberg, and M. A. Stone, "Why Are Commercials so Loud ? - Perception and Modeling of the Loudness of Amplitude-Compressed Speech," *Journal of the Acoustical Society of America*, vol. 51, no. 12, pp. 1123–1132, 2003.

[27] O. Tuomi and N. Zacharov, "A Real-Time Binaural Loudness Meter," in *Proceedings of the 139th meeting of the Acoustical Society of America*, 2000.

[28] B. C. J. Moore and B. R. Glasberg, "Modeling Binaural Loudness," *The Journal of the Acoustical Society of America*, vol. 121, no. 3, pp. 1604–1612, 2007.

[29] H. Krishnamoorthi, V. Berisha, and A. Spanias, "A Low-Complexity Loudness Estimation Algorithm," in *Proceedings of the 2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008.

[30] H. Krishnamoorthi, A. Spanias, and V. Berisha, "A Frequency / Detector Pruning Approach for Loudness Estimation," *IEEE Signal Processing Letters*, vol. 16, no. 11, pp. 997–1000, 2009.

[31] R. J. Cassidy and J. O. Smith III, "Efficient Time-Varying Loudness Estimation via the Hopping Goertzel DFT," in *Proceedings of the 50th Midwest Symposium on Circuits and Systems*, 2007, pp. 421–422.

[32] D. Ward, C. Athwal, and M. Kokuer, "An Efficient Time-Varying Loudness Model," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2013.

[33] E. Burdiel, L. Vetter, A. J. R. Simpson, M. J. Terrell, A. McPherson, and M. Sandler, "Real-time Implementation of Glasberg and Moore's Loudness Model for Time-Varying Sounds," in *Proceedings of the 133rd Audio Engineering Society Convention*, 2012.

[34] D. Cabrera, S. Ferguson, and E. Schubert, "PsySound3: An integrated environment for the analysis of sound recordings," in *Proceedings of ACOUSTICS*, 2008.

[35] M. Frigo and S. G. Johnson, "The Design and Implementation of FFTW3," *Proceedings of the IEEE*, vol. 93, no. 2, pp. 216–231, 2005.

[36] EBU, "EBU-Tech 3253 Sound Quality Assessment Material recordings for subjective tests," Tech. Rep., European Broadcast Union, 2008.

[37] A. T. Reddy, *Applied Data Analysis and Modeling for Energy Engineering and Scientists*, Springer, 2011.

[38] E. Zwicker, "Procedure for calculataing loudness of temporally variable sounds," *Journal of the Acoustical Society of America*, vol. 62, no. 3, pp. 675–682, 1977.

[39] H. Fastl and E. Zwicker, *Psychoacoustics: Facts and Models*, Springer, third edition, 2007.

[40] K. Singh and M. Xie, "Bootstrap : A Statistical Method," Tech. Rep., Rutgers University, USA, 2010.

[41] F. O'Brien and D. Cousineau, "Representing Error bars in within-subject designs in typical software packages," *The Quantitative Methods for Psychology*, vol. 10, no. 1, pp. 56–67, 2014.

[42] G. A. Soulodre, M. C. Lavoie, and S. G. Norcross, "The Subjective Loudness of Typical Program Material," in *Proceedings of the 115th Audio Engineering Society Convention*, 2003.