PARTIAL LOUDNESS IN MULTITRACK MIXING

ZHENG MA[1], JOSH REISS[1], AND DAWN BLACK[1]

[1] *Center For Digital Music (C4DM), Queen Mary, University of London, London, UK*
zheng.ma@eecs.qmul.ac.uk
josh.reiss@eecs.qmul.ac.uk
dawn.black@eecs.qmul.ac.uk

ABSTRACT:

Partial loudness can be used as high-level, perceptually relevant metadata in the context of semantic audio, especially in multitrack mixtures or wherever masking scenarios are desired. Subjective evaluation of the partial loudness model of Glasberg and Moore on multitrack signals in the form of equal loudness matching experiment is presented. The observed results imply that the current model underrates the partial loudness perceived by the subjects. We analyze the underlying features and propose a parameter modification in the implementation of the partial loudness model that yields better compliance for musical signals.

INTRODUCTION

Auditory masking occurs when the perception of one sound is affected by the presence of another sound [11]. The term 'Complete masking' is used when the presence of a sound can make another sound inaudible. Partial masking is a situation where the accompanying sound influences the perception of a given sound even though it is still audible. Partial loudness thus refers to the actual perceived loudness of a sound against a background of other sounds.

Loudness is a hot research topic in both academia and industry. Different loudness models have been proposed in the past few decades (see [1] where several commonly used loudness models were briefly explained and evaluated). However, models to predict partial loudness are relatively unexplored. Moore, Glasberg and Baer were the pioneers to propose a partial loudness model for steady sound in [2], which later extended to [3] for predicting the audibility of time-varying sounds in the presence of background sounds. In [3], a series of experiments were conducted to evaluate the model by measuring the detection thresholds for different signal and background combinations. However, most tested audio samples were laboratory stimulus such as tones and noises of duration less than 1s. No musical signal were ever used, which can be highly time varying and contains complex spectral patterns.

Partial loudness can be used as high-level, perceptually relevant metadata in the content of semantic audio, especially in the multitrack mixture or wherever masking scenarios are desired. In multitrack mixing, as long as audio signals are mixed together they inevitably mask on another. Aforementioned partial loudness model has been explored in the research of new mixing interface design [5] and the intelligent mixing system [6-9]. [7] introduced a method to quantify the masking using a signal-to-masker ratio calculated from excitation patterns. Similarly, [8] proposed a partial loudness based masked-to-unmasked ratio to describe the transparency of mixdowns. Later [9] proposed an automatic multitrack mixing algorithm to achieve an equal loudness of all instruments. Unfortunately, none of the previous works provided any formal evaluation of the partial loudness model on musical signals in the content of multitrack mixing against human perception.

In this paper, we first describe a series of loudness matching listening experiments at the point of equal loudness on instrumental stems[1] for the assessment of using partial loudness model on multitrack mixtures. Later, novel modification to the partial loudness model is suggested to achieve a better compliance with the observed human perception. The results of the research can be used in the development of intelligent mixing systems (such as the concept of semantic approach to autonomous mixing proposed in [13]) or any application where the masking scenarios are needed.

1 LOUDNESS MODELS

[1] Stem: a sub-mix of the tracks that represent the same instrument in the process of audio mixing.

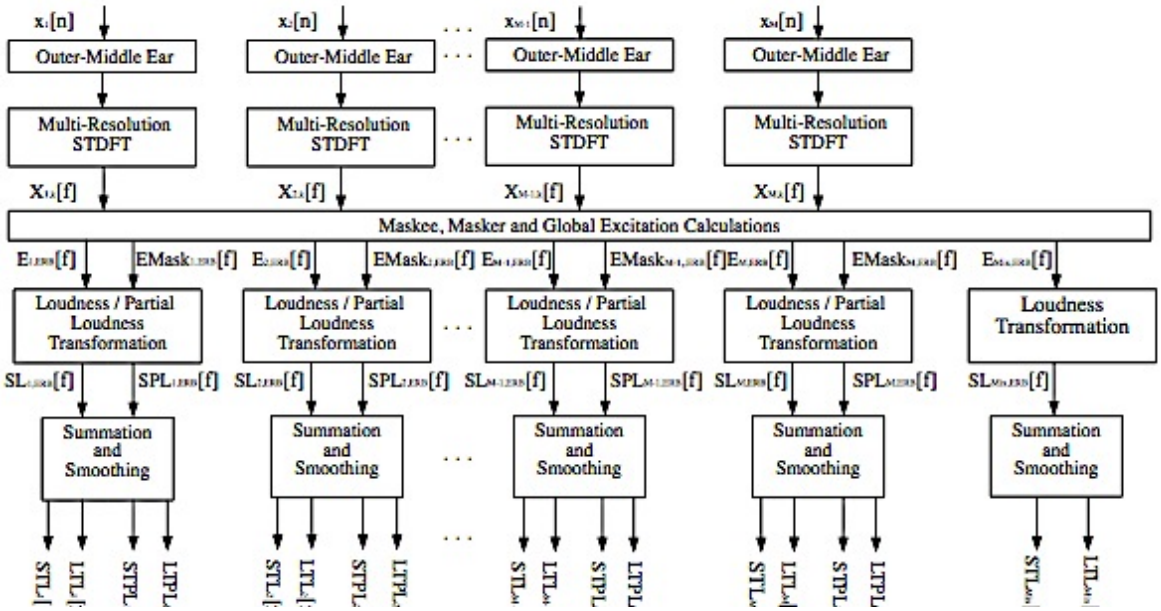

Figure 1 Block diagram of multi-channel loudness model for M input signals [9].

The multitrack loudness model that is evaluated in this paper adapts the loudness models of Glasberg and Moore [3] into a cross-adaptive architecture [10] to estimate the loudness and partial loudness of musical instruments where each input may be masked by the combination of every other input. The structural overview of the model is depicted in Figure 1. System calibration is crucial and performed by measuring the sound pressure level of a 1kHz full-scale tone at eardrum. The same headphone was used during all experiments.

The procedure to obtain the loudness and partial loudness using proposed model is summarised as follows:

1. All inputs, $x_m[n]$ to the model are first passed through a 4097 coefficient FIR filter simulating the combined outer-middle ear magnitude response [4].
2. A multi- resolution Short Time Discrete Fourier Transform (STDFT), comprising 6 parallel FFT's performs the spectral analysis.
3. Each spectral frame $X_{m,k}[f]$ is filtered by a bank of level-dependent roex filters whose centre frequencies range from 50Hz to 15kHz, the output of which yields the excitation pattern $E_{m,ERB}[f]$, where the frame number f is updated every millisecond. Such spectral filtering represents the displacement distribution and tuning characteristics across the human basilar membrane.
4. The excitation pattern is then transformed to a specific loudness pattern $SL_{m,ERB}[f]$ that represents the loudness at the output of each auditory filter. The summation of $SL_{m,ERB}[f]$ across the perceptual scale produces the total unmasked instantaneous loudness $IL_{m,ERB}[f]$.
5. To account for masking, each excitation pattern is recalculated as described in [2] along with an additional M excitation patterns required to formulate the background maskers for every channel, $EMask_{m,ERB}[f]$. The current implementation allows any combination of inputs to be used when generating the maskers. All excitation patterns are then transformed to a specific partial loudness pattern $SP_{m,ERB}[f]$ that describes loudness under inhibition [2]. This is summed to produce the total partial loudness $IPL_{m,ERB}[f]$.
6. All of the above instantaneous loudness frames are smoothed by two separate temporal integration stages resulting in two perceptual measures; the short-term loudness $STL_m[f]$, describing the loudness perceived at any moment, and the long-term loudness $LTL_m[t]$ reflecting overall loudness judgment and memory effects. Both the short-term partial loudness $STPL_m[f]$ and long-term partial loudness $LTPL_m[f]$ represent the same respective features, but under masked conditions.
7. Finally, two single values STL_m and $LTPL_m$ are computed by averaging $STPL_m[f]$, $LTPL_m[f]$ over whole period to present the averaged

perceptual unmasked and masked loudness of each stem input.

Detailed mathematical description of the loudness model is given in the original papers [2-4].

2 EVALUATION: LOUDNESS MATCHING EXPERIMENT BETWEEN SOLO STEM [2] AND MIXED STEM[3]

2.1 Procedure

A preliminary listening test was performed before the actual loudness matching experiment. Subjects were required to listen to all the mixes and to identify every instrument contained in each mix. Subjects need to pass this preliminary test in order to continue to the next formal experiment.

All tests were performed in a quiet listening room, where the environmental noise is minimized. For each loudness matching trial, both solo stem and mixed stem were presented in a regular alternation with two seconds silent intervals between successive sounds played through the same calibrated headphone. The order of the trials was randomized for each subject to minimize the bias that subjects become familiar with the same song and judge the loudness based on memory. Within a given trial, either the solo stem or the mixed stem level was fixed and the level of the other was varied to determine the level corresponding to equal loudness in perception. By varying the level of the mixed stem, it means subjects were only allowed to adjust the same instrumental stem in the mix as the solo stem while the level of other stems in the mix were kept unchanged. The starting level of the variable sound was chosen randomly from within a certain range. The starting level was chosen randomly from within a range of ±10 dB around the level of the fixed sound.

The loudness matching experiments were designed using the method of adjustment methodology similar to the method in [12]. The levels of the stems were adjusted using the built-in fader (in dB scale) tool in Apple's Logic Software. The reference stem assigned for each run (the stem that is fixed in level) could be either the solo stem or the same stem in the mix. A lock shape sign was attached to target stem (the stem that can be adjusted in level) for indication. Subjects were told to adjust the fader level of target stem until it perceived as equally loud as the reference stem. The difference between the target stem and the reference stem was recorded after each trial, which was expressed as the Root-Mean-Square level (RMS). The average differences for each stem across subjects were then calculated as a measure of partial masking/partial loudness. Model predictions were then computed in both conditions in a similar way.

[2] Solo Stem: Stem that is played separately.

[3] Mixed Stem: Stem that is played in a mixture together with other stems.

2.2 Stimuli

Four multitrack songs of different genres were selected and 10s segments of each song were extracted from the whole un-processed waveform signals. Each consisted 4 or 5 different instrument stems, all in mono and running at a typical sampling rate of 44.1 kHz. The specifications of the testing samples are presented as follows:

Table 1 The specifications of the testing samples.

	Genre	Instrumentation	RMS level (dB)
Song 1	Classical	Bassoon	64
		Clarinet	64
		Saxophone	67
		Violin	68
Song 2	Metal	Bass	67
		Electric Guitar	70
		Drum set	65
		Vocal	70
Song 3	Punk	Bass	60
		Electric Guitar	73
		Drum set	54
		Vocal	67
Song 4	Alternative rock /Electronic	Bass	52
		Drum set	65
		Acoustic Guitar	64
		Vocal	71
		Piano	62

2.3 Subjects

In total 12 participants whose age ranged from 21 to 32 had taken part in the experiments. Before commencing, subjects were asked to complete a questionnaire. The summary is displayed in Table 2. The results show that the majority of subjects had at least some experience in critically audio analysing, and no one has hearing impairment.

Table 2 Results of the informational questionnaire.

Gender	Male	9
	Female	3
Critical listening skill? / Listening tests experience?	No	2
	Some	2
	Yes	8
Hearing impairment?	No	12
	Yes	0

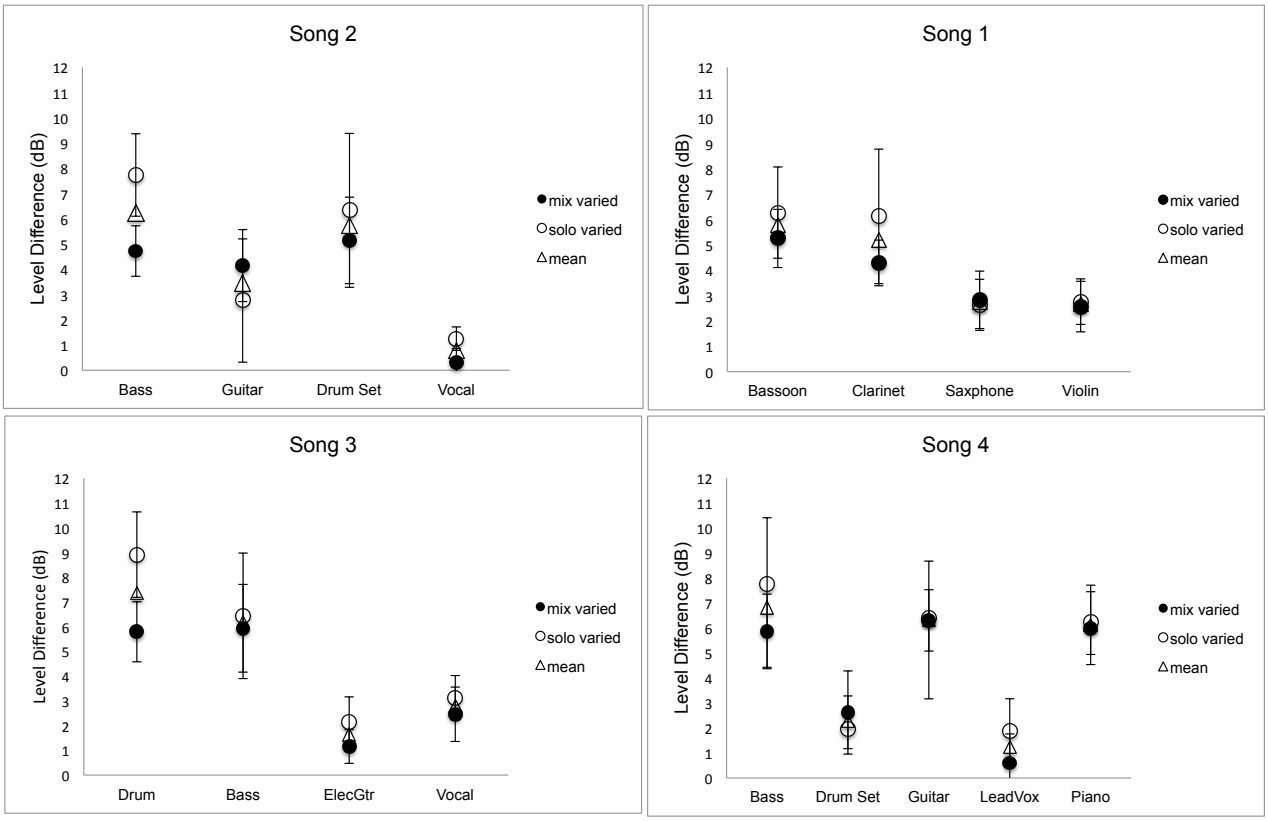

Figure 2 The measured results of both cased at the point of equal loudness across all the subjects.

All 12 subjects successfully passed the preliminary tests suggesting that subjects were able to identify and judge the partial loudness of an instrument stem when mixed with other musical sources. To present the results, the level difference between the solo stem and the mixed stem at the point of equal loudness are calculated as follows:

$$L_\Delta = L_m - L_s \quad (1)$$

Where L_m, L_s are the levels of the mixed stem and the solo stem respectively. Positive level difference L_Δ indicates the mixed stems require a larger RMS level to reach the point of equal loudness with the solo stems. This agrees with the concept of partial masking, i.e., the loudness of an audio signal is generally reduced in the presence of a background of other sounds. However, unusual negative L_Δ is less common and considered an error due to subjects' mistakes in the experiment or the sensitivity limit of human ears, which is generally within ± *2 dB*.

The mean subjective results of the loudness matching experiments across all the subjects are shown in Figure 2. Results are plotted separately for the case where the mixed stem is varied (filled circle) and the case where the solo track is varied (open circle). The triangles represent the mean value of both cases.

As Figure 2 shows, the evaluation results for both conditions (open circles and filled circles) shared a good degree of consistency. There is a very small bias related to whether the mixed stem or the solo stem was varied. The open circles lie above the filled circles at most instrument stems indicating that subjects tend to assign a lower level to the solo stem when matching loudness against the mixed stem. The mean of the consistent bias across all conditions and subjects is about *+1.2 dB*. We believe the bias results from the difficulty of judging the loudness of the mixed stem as a reference out of the mixture.

Discounting the bias by looking at the mean for both conditions (triangles), all values are positive, above the 0 dB line, which means that at the point of equal loudness the RMS level of mixed stems are higher than the solo stems. This implies that partial masking occurs. The level difference L_Δ at the point of equal loudness could be seen as a measurement of partial loudness.

We can also observe some variations across different instrument stems for every song. The drum set stem in song 3 scored the highest level-difference of 7.4 dB

while the vocal tracks in song 2 and song 4 have the lowest average of variation of 0.8 dB and 1.2 dB respectively. It means some instruments suffer less partial masking while other instruments suffer significant partial masking resulting larger loudness reduction. It also confirmed that masking is source dependent. The level and frequency interactions between the masker and masked sounds decide the degree of simultaneous masking.

2.4.1 Model Prediction

Next, we employed the adapted loudness model described in the previous section (see section 1) to predict the same level difference at the point of equal loudness as in the listening experiment. Theoretically, the point of equal loudness for model prediction should be:

$$LTL_m = LTPL_m \quad (2)$$

The separate loudness of the *m-th* stem, LTL_m equals to the partial loudness of the same stem when presented in the mix, $LTPL_m$. Model predictions were obtained separately for both cases corresponding to the loudness matching experiments. For instance, in the case of varying the level of the solo stem: an average long-term partial loudness of the mixed stem, $LTPL_m$, regarding the sum of the other stems as a masker is calculated. This $LTPL_m$ served as a loudness reference for equal loudness matching. The average long-term loudness of the solo stem, LTL_m is calculated and compared against $LTPL_m$. Iterations of applying boost or attenuation (in *dB* scale) are performed to the solo stem. New LTL_m is then re-calculated and compared to the reference again. The iteration continues until the condition $|LTPL_m - LTL_m| \leq T$ is fulfilled, where T =1.5 phons is the tolerance of error. When reaching the point of equal loudness, the value of attenuation offset (in *dB*) is then

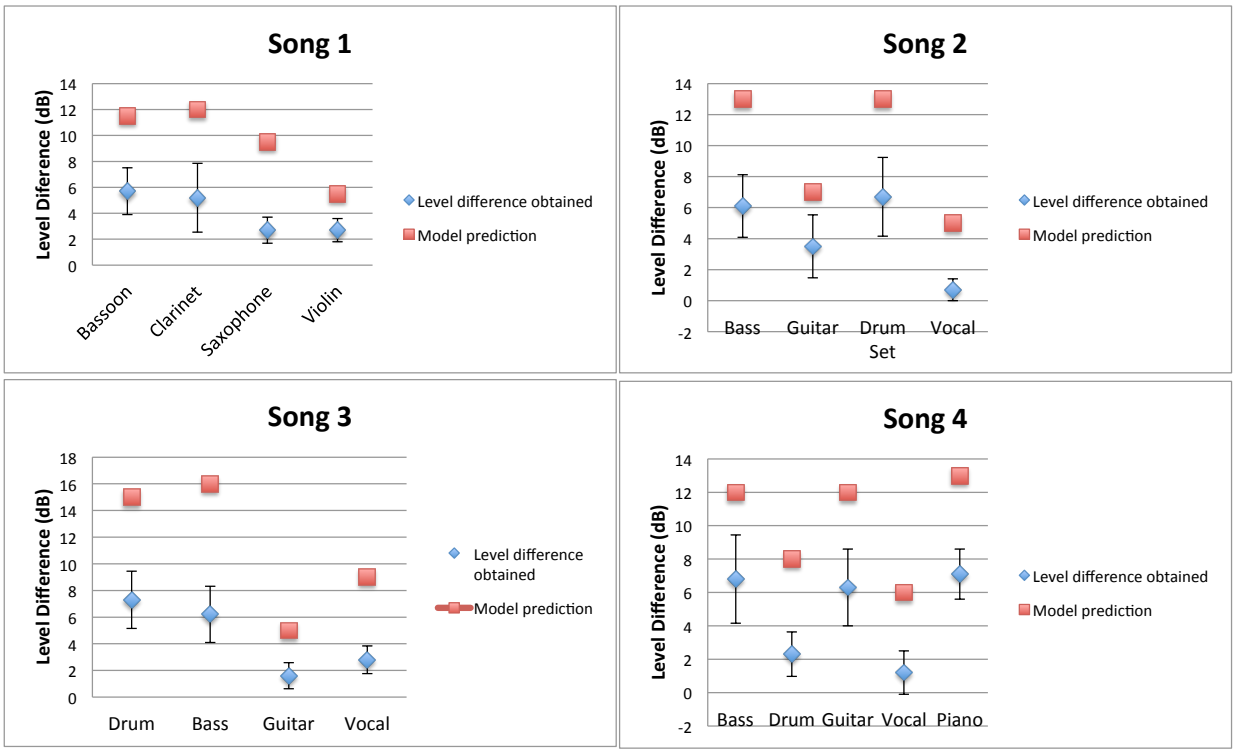

Figure 3 Visualized data presentation of model prediction against measured values

recorded as the model prediction of the difference in level. A similar scenario is performed for the case of varying the mixed stem, where LTL_m of the solo stem acts as the loudness reference and $LTPL_m$is continually calculated until it satisfies the equal loudness condition within the tolerance of error.

Table 3 Results of model prediction for level difference.

	Instrument	Measured Level Difference (dB)	Model Prediction (dB)	Prediction Error (dB)
Song 1	Bassoon	5.7	11.5	5.8
	Clarinet	5.2	12	6.8
	Saxophone	2.7	9.5	6.8
	Violin	2.7	5.5	2.8
Song 2	Bass	6.1	13	6.9
	Guitar	3.5	7	3.5
	Drum Set	6.7	13	6.3
	Vocal	0.7	5	4.3
Song 3	Drum	7.3	15	7.7
	Bass	6.2	16	9.8
	Guitar	1.6	5	3.4
	Vocal	2.8	9	6.2
Song 4	Bass	6.8	12	5.2
	Drum	2.3	8	5.7
	Guitar	6.3	12	5.7
	Vocal	1.2	6	4.8
	Piano	7.1	13	5.9

Table 3 presents the level difference predicted by the proposed loudness model and comparison with the measured mean results from loudness matching experiments. The final column shows prediction error.

As Figure 3 and Table 3 show, the model prediction values correlate well with the overall trend of the level difference obtained from loudness matching experiments. However, the model predictions are much higher than the empirical results. The biggest prediction error of 9.8 dB is found at the bass stem in song 3. Even the lowest difference at the violin stem in Song 1 still shoots up to *2.8 dB*. These errors values (see last column in table 3) are significantly larger than the minimum perception sensitivity of human hearing system of loudness variations.

Overall, results suggest the proposed loudness model overrated the loudness reduction caused by partial masking. The cause of the problem could be the nature of music signals. Unlike laboratory stimuli such as tones and noises, music signals could contain distinct spectral components and rhythm and melody structures, which could make it easier to distinguish. As a result, it reduces the effect of partial masking in the mix. However, it's more likely that the model prediction errors arise from the partial loudness model. Looking into the process of obtaining the model prediction at the point of equal loudness: LTL_m=$LTPL_m$. As previous research [3] shows that loudness performs well in predicting the loudness of the sounds without the presence of other sounds, which suggests LTL_m values corresponds well to perception. Then all errors are positive indicating that the partial loudness predicted by the model, $LTPL_m$ is lower than the loudness that subjects perceived. That is, the partial loudness model underrates the loudness of musical signal in the presence of other sounds. In addition, the model does not take into account the fact that the audibility of a signal may be improved when the masker contains amplitude fluctuations that are correlated in different frequency regions. Therefore authors believe that some minor modification in the partial loudness implementation could be made to better describe the masking scenario in musical signals.

3 MODIFICATION

Following the previous results and discussion, we look into the implementation of the partial loudness model and adjust the model to produce more accurate partial loudness prediction for music signals.

3.1 Parameter K in Partial Loudness Model

A parameter K was introduced in the process of transformation of the excitation pattern to a specific partial loudness pattern [2]. The parameter K has a crucial influence on the calculation on partial loudness. It is used to obtain the signal's excitation at its masked threshold. The lower the values of K, the higher the predicted partial loudness value.

According to (Moore 1997), specific partial loudness is assigned at different signal levels in four situations namely:

1) $E_{SIG} \geq E_{THRN}$ *and* $E_{SIG}+E_M \leq 10^{10}$
2) $E_{SIG} \geq E_{THRN}$ *and* $E_{SIG}+E_M < 10^{10}$

3) $E_{SIG}<E_{THRN}$ *and* $E_{SIG}+E_M \leq 10^{10}$
4) $E_{SIG}<E_{THRN}$ *and* $E_{SIG}+E_M>10^{10}$

Where E_{SIG}, E_M denote the signal and masker excitation in quiet, E_{THRN} is the peak excitation of signal at its masked threshold in the presence of background sounds. It is calculated from the equation: $E_{THRN}=K \cdot E_M+E_{THRQ}$, *where* E_{THRQ} is the signal's excitation at absolute hearing threshold. K is then defined as the signal-to-noise ratio at the output of the auditory filter required for threshold at high masker levels. The values of K as a function of frequency are estimated by pooling data from relatively old research work [2]. Nevertheless, there are no estimates of K for centre frequencies below 100Hz, so values from 50 to 100 Hz are based on extrapolation.

3.2 Adjustment of Parameter *K* and Evaluation

In [8], threshold detection experiments using an adaptive two-alternative forced-choice (2AFC) task to adjust the partial loudness model were performed. The results showed that if K was reduced by 5 dB the compliance of the prediction and the measurement is improved. However, once again, the stimuli used in the experiment were laboratory tones and noise rather than musical signal. Thus model adjustment based on K is further explored. We perform the same model prediction process as in Section 3 using different partial loudness models with different K values. The values of K are chosen to be the original K series, with 5 dB attenuation, with 10 dB attenuation and 15 dB attenuation. The results of the different model predictions are compared to the evaluation results obtained from the loudness matching experiments. See Figure 4 below.

The blue diamond indicates the mean result obtained from the loudness matching experiment with error bars corresponding to the standard deviation across all subjects. Blue circle, red square, green triangle, purple cross indicate the model predictions with 15 dB reduction in K, 10 dB reduction in K, 5 dB reduction in K and its original, suggested values respectively.

The model predictions by the original K values, -5dB K values are all above the upper standard deviation of the obtained subject's data, which mean that these two models overestimate the effect of partial masking. The

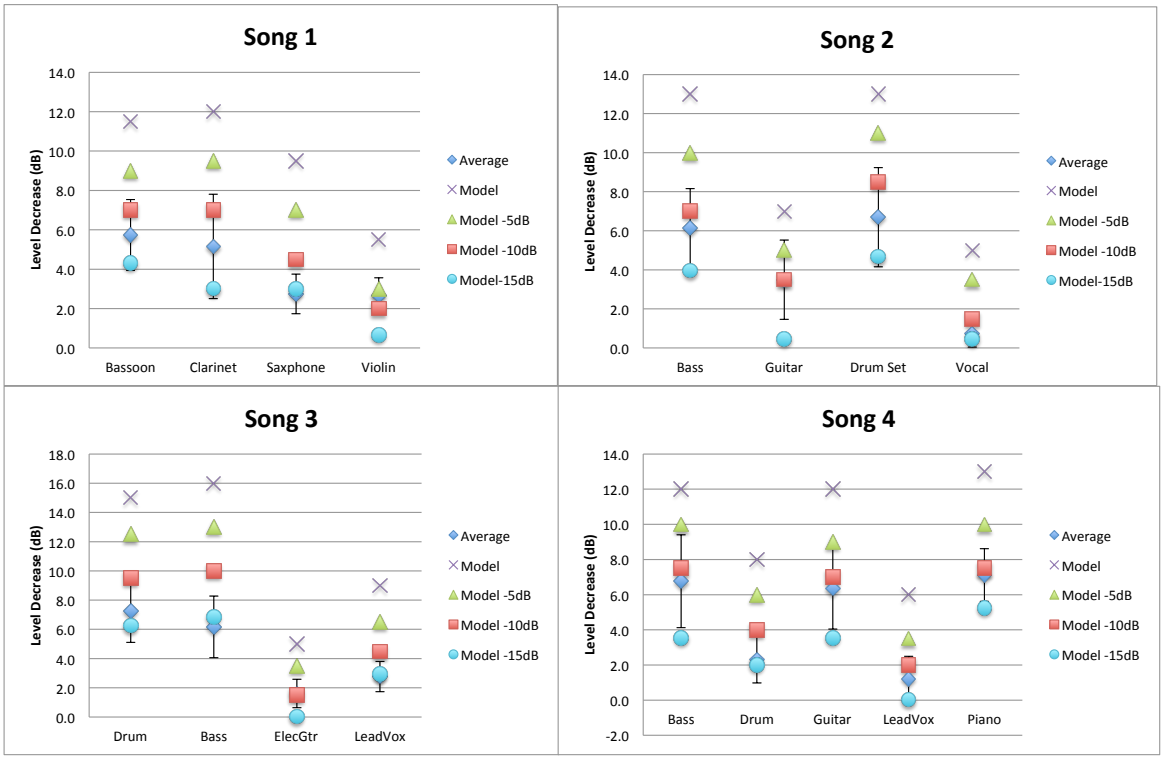

Figure 4 Comparison of different model predictions with obtained results

Model with -15 dB *K* values on the other hand underrates the effect of partial masking. Overall, model adjustment with 10 dB attenuation applied to *K* produces the best compliance with the empirical data as most model predictions values (19 out of 21) are within the standard deviation area of the empirical data. Detailed results of the model with adjustment of -10dB in *K* compared with the empirical data are shown in Table 4. As the last column, prediction error shows in Table 4, most errors are within 0 - 1.5 dB variation which are barely perceivable through the human hearing system.

Table 4 Comparison between -10dB model predictions with the obtained data

	Instrument	Measured Level Difference (dB)	-10dB Model Prediction (dB)	Prediction Error (dB)
Song 1	Bassoon	5.7	7	1.3
	Clarinet	5.2	7	1.8
	Saxophone	2.7	4.5	1.8
	Violin	2.7	2	-0.7
Song 2	Bass	6.1	7	0.9
	Guitar	3.5	3.5	0.0
	Drum Set	6.7	8.5	1.8
	Vocal	0.7	1.5	0.8
Song 3	Drum	7.3	9.5	2.2
	Bass	6.2	10	3.8
	Guitar	1.6	1.5	-0.1
	Vocal	2.8	4.5	1.7
Song 4	Bass	6.8	7.5	0.6
	Drum	2.3	4	1.7
	Guitar	6.3	7	0.7
	Vocal	1.2	2	0.8
	Piano	7.1	7.5	0.4

4 CONCLUSIONS AND FUTURE WORK

A loudness matching experiment on real musical signals using the method of adjustment was conducted to evaluate the performance of proposed partial loudness model. Empirical results suggest an adjustment of the parameter K in the partial loudness implementation can be made to obtain a better compliance between model predictions and subjective evaluation of human hearing. The results are summarized as follow:

1. We have proved that when mixing instrument stems together, the perceptual loudness of individual tracks is reduced due to an effect called partial masking

2. The results show that the effect of partial masking on the perception of the overall loudness is significant. The loudness reduction varies across different instruments, which indicates that the partial loudness of the musical signal depends on the sonic interaction between the stems being mixed together. The results shared a trend across subjects.

3. There was a small consistent bias effect related to whether the track in the mix or the solo track was varied. The differences at the point of equal loudness obtained in the case of varying the solo track were slightly higher.

4. The model prediction produced by the partial loudness model of [2-4] with an adjustment of reducing the *K* parameter by 10 dB yields a better compliance with the measured loudness reduction.

A larger scale listening test using more subjects and more diverse music signals as future work will improve the performance of employing the partial loudness model on musical signals. The improved partial loudness model can be used in any situation where masking scenarios between complex signals is desired such as intelligent mixing production, audio quality evaluation and audio broadcasting.

5 REFERENCES

[1] Nielsen, Soren H. and Skovenborg, Esben. "Evaluation of Different Loudness Models with Music and Speech Material." *117th Audio Engineering Society Convention* . San Francisco, 2004.

[2] Moore, B. C. J., Glasberg, B. R. and Baer T. "A model for the prediction of thresholds, loudness, and partial loudness." *J. Audio Eng. Soc.* 45 (1997): 224-240.

[3] Glasberg, Brian R. and Moore, Brian C. J. "Development and Evaluation of a Model for Predicting the Audibility of Time-Varying Sounds in the Presence of Background Sounds." *J. Audio Eng. Soc* 53 (2005): 906-918.

[4] Glasberg, Brian R. and Moore, Brian C. J. "A Model of Loudness Applicable to Time-Varying Sounds." *J. Audio Eng. Soc* 50 (2002): 331-342.

[5] Terrell, Michael J. and Simpson, Andrew J. R. and Sandler, Mark B. "A Perceptual Audio Mixing Device." *134th Audio Engineering Society Convention.* Rome, 2013.
[6] Reiss, J. D. "Intelligent Systems for Mixing Multichannel Audio." *17th International Conference on Digital Signal Processing (DSP2011).* Corfu, Greece, 2011. 1-6.
[7] Vega, S., and Janer J. "Quantifying Masking in Multi-track Recordings." *Proceedings of SMC Conference.* Barcelona, 2010.
[8] Aichinger, P., A. Sontacchi, and B. Schneider-Stickler. "Describing the Transparency of Mixdowns: The Masked-to-Unmasked-Ratio." *130th Audio Engineering Society Convention.* 2011.
[9] Ward, Dominic and Reiss, Joshua D. and Athwal, Cham. "Multitrack Mixing Using a Model of Loudness and Partial Loudness." *133th Audio Engineering Society Convention.* New York, 2012.
[10] U. Zolzer, V. Verfaille, D. Arfib. "Adaptive Digital Audio Effects (A-DAFx): A New Class of Sound Transformations." *IEEE Transactions on Audio, Speech and Lan- guage Processing* 14 (2006): 1817-31.
[11] Gelfand, S.A. *Hearing- An Introduction to Psychological and Physiological Acoustics.* 4th. New York: Marcel Dekker, 2004.
[12] Brian C. J. Moore, Deborah A. Vickers, Thomas Baer, and Stefan Launer. "Factors affecting the loudness of modulated sounds." *J. Acoust. Soc.* 105, no. 5 (1999): 2757-2772.
[13] De Man, B, Reiss, Joshua D. “A semantic approach to autonomous mixing” APR13, 2013.