Article

Musical Mix Clarity Prediction Using Decomposition and Perceptual Masking Thresholds

Andrew Parker * and Steven Fenton

Centre for Audio and Psychoacoustic Engineering (CAPE), University of Huddersfield,
Huddersfield HD1 3DH, UK; S.M.Fenton@hud.ac.uk

* Correspondence: Andrew.Parker@hud.ac.uk

Featured Application: The model proposed in this work is expected to form the basis of a perceptually motivated objective mix clarity model to be used in music production and music information retrieval applications.

Abstract: Objective measurement of perceptually motivated music attributes has application in both target-driven mixing and mastering methodologies and music information retrieval. This work proposes a perceptual model of mix clarity which decomposes a mixed input signal into transient, steady-state, and residual components. Masking thresholds are calculated for each component and their relative relationship is used to determine an overall masking score as the model's output. Three variants of the model were tested against subjective mix clarity scores gathered from a controlled listening test. The best performing variant achieved a Spearman's rank correlation of $\rho = 0.8382$ ($p < 0.01$). Furthermore, the model output was analysed using an independent dataset generated by progressively applying degradation effects to the test stimuli. Analysis of the model suggested a close relationship between the proposed model and the subjective mix clarity scores particularly when masking was measured using linearly spaced analysis bands. Moreover, the presence of noise-like residual signals was shown to have a negative effect on the perceived mix clarity.

Keywords: mix clarity; clarity; auditory masking; perception; psychoacoustic model; MPEG; music information retrieval

Citation: Parker, A.; Fenton, S. Musical Mix Clarity Prediction Using Decomposition and Perceptual Masking Thresholds. *Appl. Sci.* **2021**, *11*, 9578. <https://doi.org/10.3390/app11209578>

Academic Editor: Sławomir
K. Zieliński

Received: 8 September 2021

Accepted: 12 October 2021

Published: 14 October 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Terms such as 'clarity', 'punch', 'warmth' and 'brightness' are semantics often used to describe perceptual features found in musical mixes. These features are often subconsciously combined by a listener when assessing the overall quality of a musical mix. Whilst some of the perceptual features represented by the semantics outlined have an objective counterpart [1–5], clarity does not.

Clarity in the context of this work is related to Pedersen and Zacharov's sound wheel term 'clean' [6], which is defined as: "It is easy to listen into the music, which is clear and distinct. Instruments and vocals are reproduced accurately and distinctly. The opposite of clean: dull, muddy." Other similar definitions can be found in the literature, for example [7–11]. Although none of these definitions are uniform in wording, they have a similar focus on the separability of the component parts of the mix, such that each part is distinctly audible. Potential links between the perception of single instrument clarity and brightness, measured using centroid and harmonic centroid, have been suggested [7,12]. This work evaluates a model-based approach to objective mix clarity prediction. Perceptually motivated metrics have their use in metering and control applications, as well as automatic mixing [10,13–15]. Additionally, understanding the underlying related signal characteristics of these perceptual features facilitates the proposal of formalised definitions.

It is noted that the idea of 'clarity' and 'quality' should not be conflated here. The perceived level of clarity for a given multitrack mix may not correlate directly to the

perception of the overall quality, as this will depend on the content and intention of the multitrack mix itself. Furthermore, overall quality judgement is likely to be made based upon multiple perceptual features such as loudness [5] and punch [1] alongside that of clarity.

Considering the acoustic characteristics of a space, it is possible to objectively determine the clarity and intelligibility that can be achieved. Measures such as C50 (ratio between early and late arriving reflections) and early decay time (EDT) can be combined for this purpose. The direct to reverberant ratio (D/R) has also been linked to acoustic clarity [16]. Furthermore, in a proposal for objective measurement for loudspeaker quality [17], ‘clearness’ is considered to be a perceptual dimension. This is calculated based on perceived degradation imparted on a signal by the loudspeaker under test in reference to an ‘ideal’ loudspeaker signal. This approach is similar to established standards to objectively measure music and speech quality, PESQ [18] and PEAQ [19]. These compare encoded signals with their unencoded counterparts to estimate a perceived error signal caused by encoding, which can then be measured using a number of features to determine how disturbing it is. What is common between these measures is that they all compare a signal with an affected version of itself, for example altered by a room or a loudspeaker’s response. In the proposed model, there is not a processed version of the signal under test to compare to. Instead, the signal is decomposed and a comparison made between its constituent components to determine if they are suitably balanced.

In previous research [20,21], it has been suggested that masking between signals constituting a multitrack mix could be related to the perception of clarity of the overall mix. Auditory masking is a phenomenon of hearing in which energy present at the ear is not perceived, due to stronger neighbouring energy in frequency or time, known as frequency masking and temporal masking, respectively [22]. The model proposed for mix clarity prediction is based on an analysis of the masking relationship between transient, steady-state, and residual components of the signal, utilising the MPEG Psychoacoustic Model II [23,24]. The model’s performance is assessed using a correlation test against subjective mix clarity scores elicited from a controlled listening test. In addition, further analysis of the model’s response to an independent dataset consisting of musical excerpts with varying degrees of controlled degradation is undertaken. Finally, the performance is discussed, including comparison to the model defined in previous research [20,21].

2. Proposed Model

The proposed model has some similarities to a cross-adaptive approach employed to minimise masking between composite signals of a multitrack mix in automatic mixing systems [10,14]. The automatic mixing system employed in [10] was shown to increase the perceived clarity of the mixes where the amount of masking had been reduced. To achieve this, a cross-adaptive masking metric was used to calculate the level by which a given signal in the multitrack was masked by the sum of all other signals of the multitrack mix. The signals contributing to the given multitrack mix are then processed in such a way to minimise the amount they are masked, thus lowering the amount of masking occurring in the multitrack mix overall. Their work was congruent with [25] which proposes a cross-analysis based model. This compares the partial loudness (frequency dependent loudness) of a signal in isolation to the partial loudness of the signal when masked by the other signals present in the multitrack mix. Subjective testing showed the greater the disparity in partial loudness between the masked and unmasked forms of a signal, the lower the probability of it being successfully identified in the multitrack mix. A similar approach was also proposed as a Hierarchical Perceptual Mixing (HPM) system [26], which first determines the most important signals present in the mix as a function of time based on a user parameter, then calculates the perceptual masking threshold of these dominant signals and removes the masked energy of the non-dominant signals present in the mix. It is suggested this approach may improve the clarity of the resulting mix. However, this was not related to any subjective testing.

In typical MIR applications, a multitrack based approach cannot be used directly as it assumes access to all of the signals present which make up the multitrack mix. Additionally, it is unable to calculate masking occurring contained within a single signal, for example, if multiple instruments were captured by a single microphone. Whilst there is an emerging field of source separation methods utilising deep learning techniques to ‘unmix’ mixed signals, such as Spleeter [27], these systems are unsuited in this case as they are only able to classify a limited number of instruments. This results in the cross-masking being poorly represented in cases where the given mix is split into only 1 or 2 instrument categories.

The proposed model assumes no knowledge of the individual signals making up the multitrack mix. It utilises a novel approach which calculates the level of perceived masking between the separated transient, steady-state, and residual (TSR) components of the multitrack mix. The model outline is shown in Figure 1 and consists of the TSR separation stage, a component energy and masking calculation stage (utilising the MPEG Psychoacoustic model II), and finally a statistical output calculation stage which generates the representation of the masking occurring between the separated components.

Figure 1. Block diagram of the proposed transient, steady-state and residual based masking metric.

2.1. Transient, Steady-State, and Residual Separation Stage

When considering a spectrogram representation, the TSR components are characterised as follows:

- Transient components: Spectrally broadband and temporally transient bursts of energy which form strong vertical beams on the magnitude spectrogram [28], and unpredictable changes in phase between consecutive frames [29].
- Steady-state components: Slowly evolving harmonic partials with predictably evolving phase between consecutive frames [29], forming strong horizontal beams on the magnitude spectrogram [28].
- Residual components: Cannot be classified as transient or steady-state. These are noise-like components of the signal, described as the ‘texture’ of the sound [30], which are generally broadband and stochastic in terms of magnitude and phase.

The TSR model gives a more general description of signal components that do not require specific instrument classification. Following the suggested negative impact of noise-like signals on mix clarity [20,21], greater masking of the residual (R) component would indicate a less audible residual component and therefore a potentially higher perceived mix clarity. Additionally, and perhaps more importantly, less masking of the transient and steady-state (TSS) components would indicate greater audibility of the percussive (rhythmic) and harmonic (pitched) parts of the signal. Transient onsets have also been linked to instrument identification [31], suggesting potentially greater mix clarity perception where TSS components are less masked. Considering the HPM system [26], masked residual

energy bears some resemblance to the idea of masked non-dominant signals, whose presence may be unnecessary or even detrimental to the perceived clarity of the mixed signal. Moreover, the importance of TSS and R components can be thought of as a simple hierarchy, where the presence of the TSS components take priority over R components. In this regard, the residual energy may not necessarily be detrimental to the perception of clarity where it is below a level which would mask the TSS energy. On the contrary, as the residual information contains cues which dictate the texture and timbre of instruments, its presence may be beneficial to a mix where it is properly balanced.

The TSR separation stage adopts a median filter based approach [30], chosen for its efficiency in the perceptual measure of punch [1,32] and its relative simplicity. For this, the magnitude spectrogram ($S(k, t)$) of the input signal is taken using a short time Fourier transform (STFT), where k indicates bin index, and t indicates frame index. In this case, the STFT was performed on a 2048 sample window, multiplied with a Hanning window function, and a 50% overlap between frames. The magnitude spectrogram is then filtered with a sliding 17th order median filter. Filtering is performed vertically across the bins, yielding the transient emphasised spectrogram (S_T), and horizontally across the frames yielding the steady-state emphasised spectrogram (S_{SS}). Binary masks (BM) were then defined for the transient and steady-state components as:

$$\begin{aligned} BM_T(k, t) &= \left(\frac{S_T(k, t)}{S_{SS}(k, t) + \epsilon} \right) > \beta_T \\ BM_{SS}(k, t) &= \left(\frac{S_{SS}(k, t)}{S_T(k, t) + \epsilon} \right) > \beta_{SS} \end{aligned} \quad (1)$$

where ϵ is a small offset to avoid division by 0, β_T is the transient threshold set at 1.75, and β_{SS} is the steady-state threshold set at 1. These parameter values were chosen based on previous works, namely the perceptual punch meter [1,32].

The residual component mask is determined to be any bins which do not appear in the transient or steady-state masks, as:

$$BM_R(k, t) = 1 - (BM_T(k, t) \mid BM_{SS}(k, t)) \quad (2)$$

Using the masks, the spectrograms of the transient (T), steady-state (S) and residual (R) components can then be extracted:

$$\begin{aligned} T(k, t) &= S(k, t) * BM_T(k, t) \\ S(k, t) &= S(k, t) * BM_{SS}(k, t) \\ R(k, t) &= S(k, t) * BM_R(k, t) \end{aligned} \quad (3)$$

The separated component signals are then resynthesised by inverse STFT and overlap add procedure, before being combined by addition without the residual to form the TSS signal. The TSS and R signals are then passed to MPEG Psychoacoustic Model II stage.

It is worth noting that ideally the separation stage incorporated in the model would leave no residual energy in the TSS component and vice versa. In practice, however, with a purely white noise input, the output of energy to both TSS and R components is approximately even. This is not an issue in the present application as the white noise example simply represents the upper-bound of overall masking scores.

2.2. Component Energy and Masking Calculation Stage

The MPEG Psychoacoustic Model's output, the signal-to-mask ratio (SMR), is employed in the cross-adaptive automatic mixing system [10]. It is also utilised in the proposed model where it is used to represent masking occurring across sub-bands of the frequency spectrum, called 'scale-factor bands' (sb) [23]. SMR indicates the ratio of the energy (E) of

the input signal to a masking threshold (MT), which is calculated as a function of frequency and time [23,24]. Thus, the SMR is defined as:

$$SMR(sb) = \frac{E(sb)}{MT(sb)} \quad (4)$$

where MT is calculated by grouping spectral bins into threshold calculation partitions, which represent approximately a third of a critical bandwidth. Energy in these partitions is spread and then weighted based on a tonality measure. The tonality measure is a sliding scale indicating how noise-like or tone-like the input is. More noise-like partitions result in an increased masking threshold as they are more effective maskers. This weighted masking threshold is then compared with the threshold in quiet and the largest of the two is taken as the final masking threshold in each partition. By calculating the masking threshold of an external signal, the SMR reflects the level at which the external signal will be perceived to mask the input signal.

Three implementations of the MPEG Psychoacoustic Model were explored in the objective testing of the clarity model, these were the layer 2 (L2PM), layer 3 (L3PM) and modified L3PM. This allowed for the performance of the implementations to be compared, with a view to establishing which would perform optimally in the proposed mix clarity model. There are three key differences between SMR calculation in the L2PM and L3PM. Firstly, the L3PM employs two window lengths, a long window, and a short window. This improves the temporal resolution of the tonality measure at high frequencies using the short windows, while retaining spectral resolution at low frequencies using the long windows. The L3PM can also employ a window switching technique, such that SMR calculation for particularly transient information is performed solely using the short window length, to reduce pre-ringing. When using only short windows the number of scale-factor bands output is reduced as the masking calculation has lower frequency resolution. Secondly, the L2PM masking thresholds are calculated in approximately one third of critical bands which are then spread back over the spectral bins of the Fourier transform. The spread masking threshold is then grouped into linearly spaced scale-factor bands, from which the SMR calculation is performed. In the L3PM, the masking thresholds are also calculated in approximately one third of critical bands. These are converted directly into scale-factor bands spaced to approximate the critical bands of human hearing. As such, while masking threshold calculations in both are perceptually derived, the L3PM scale-factor band spacing is more closely related to human perception than the L2PM scale-factor band spacing. Finally, due to the more complex nature of the L3PM, it requires more computational overhead than the L2PM.

The modified L3PM is a combined approach devised for the present work. It performs the masking threshold calculation from the L3PM, which is then spread back over the spectral bins of the Fourier transform and used to calculate the SMR in linearly grouped scale-factor bands employed in the L2PM. This allowed for assessment of the impact of the scale-factor band spacing on the model performance presented in Section 4.

In the case of the cross-adaptive masking metric [10], the energy and threshold calculation for the scale-factor bands is kept the same, though they define a masker-to-signal ratio (MSR) given in decibels as:

$$MSR(sb) = 10 \log_{10} \left(\frac{MT'(sb)}{E(sb)} \right) \quad (5)$$

where MT' is the threshold calculated for the sum of accompanying signals, and E is the energy of the signal of interest. They assume masking occurs in any band where $E < MT'$.

An overall masking metric is defined as the sum of scale-factor bands where $E < MT'$, which is then scaled by a predefined $Tmax$ value of 20 dB, giving the final masking metric as:

$$Masking\ Metric = \sum_{\substack{sb = 1 \\ sb \neq E < MT'}}^N \frac{MSR(sb)}{Tmax} \quad (6)$$

where N is the number of scale-factor bands.

The model proposed in this paper first calculates the MSRs for the TSS energy (E_{TSS}) compared to the masking threshold calculated from the R component (MT_R), and the R energy (E_R) compared to the masking threshold calculated from the TSS component (MT_{TSS}), following Equation (5):

$$\begin{aligned} MSR_{TSS}(sb) &= 10 \log_{10} \left(\frac{MT_R(sb)}{E_{TSS}(sb)} \right) \\ MSR_R(sb) &= 10 \log_{10} \left(\frac{MT_{TSS}(sb)}{E_R(sb)} \right) \end{aligned} \quad (7)$$

Then, the overall amount of the TSS component is masked by the R component (AM_{TSS}) and vice versa (AM_R) is calculated as the sum of MSRs, following Equation (6):

$$\begin{aligned} AM_{TSS} &= \sum_{\substack{sb = 1 \\ sb \neq E_{TSS} < MT_R}}^N \frac{MSR_{TSS}(sb)}{Tmax} \\ AM_R &= \sum_{\substack{sb = 1 \\ sb \neq E_R < MT_{TSS}}}^N \frac{MSR_R(sb)}{Tmax} \end{aligned} \quad (8)$$

Since the number of scale-factor bands output by the L3PM and modified L3PM can vary, depending on whether a long or short window is used, the overall masking scores for each component (AM_{TSS} and AM_R) are normalised by the number of scale-factor bands output:

$$\begin{aligned} AM_{TSS} &= \frac{AM_{TSS}}{N} \\ AM_R &= \frac{AM_R}{N} \end{aligned} \quad (9)$$

2.3. Statistical Output Calculation Stage

A statistical representation of each metric is taken to represent an overall score for the given analysis window, in this case 10 s, as this was the length of the stimuli included in the listening test. These are as follows: the 5th percentile of AM_{TSS} giving the points where the TSS component is minimally masked, and the 95th percentile of AM_R giving the points where the R component is maximally masked. The overall masking metric is defined as the ratio between the statistical representations of masking expressed in decibels:

$$Overall\ Masking\ Score = 10 \log_{10} \left(\frac{5thPercentile(AM_{TSS})}{95thPercentile(AM_R)} \right) \quad (10)$$

As such, where AM_R is low and AM_{TSS} is high a large overall score is given and in the opposing case a low overall score is given, thus the score is negatively correlated with mix clarity perception.

3. Test Stimuli

The model outputs were evaluated against subjective scores collected in a controlled listening test. The stimuli used in testing were widely stylistically varied in order to determine features of clarity that applied regardless of instrumentation or arrangement, such that any resulting metrics may be generally applicable across a wide range of music. In addition, an independent dataset was synthesised to investigate the effect of specific forms of signal degradation on the model's output.

3.1. Subjective Score Elicitation Test

A controlled listening test was conducted to elicit perceived mix clarity scores across a selection of stylistically different musical stimuli as presented in previous work [20]. In total, 18 listeners took part in the listening test, of which 11 were undergraduate students enrolled in a music technology course, 4 were postgraduate researchers, and 3 were Doctors in the field of Psychoacoustics.

The test stimuli were taken from the Free Music Archive (FMA) 'FMA small' dataset [33], which consists of 8000 30-s-long musical excerpts from 8 different parent genres. The dataset was processed and a random selection method was used to select 16 10-s-long 44.1 kHz 16 bit wave file stimuli which were monophonic and loudness normalised (−23LU) [5]. This process allowed the selection of 16 stimuli that were widely stylistically varied without the need for any personal selection. Details of the selected stimuli are given in Table 1.

Table 1. Table of selected listening test stimuli.

FMA ID	Parent Genre	Sub-Genre
144179 134826	Hip-Hop	Alternative Hip-Hop Hip-Hop
021895 040903	Rock	Indie-Rock Rock
067357 081895	Experimental	Audio College Field Recording
053727 062596	International	Polka Afrobeat
067357 121592	Pop	Experimental Pop Pop
094414 114556	Electronic	Ambient Electronic Electronic
122356 146639	Folk	Singer-Songwriter Psych-Folk
126410 137167	Instrumental	Soundtrack New-Age

3.2. Procedure

Listeners were asked to rate the stimuli for perceived mix clarity. This was a standalone and absolute judgement of mix clarity without reference or comparison to any other music, similar to how mix clarity would be judged were the listener to audition a piece of music on the radio or a music playlist.

The custom test interface used was designed in MAX [34] by modifying an existing HULTIGEN [35] interface. Each stimulus was presented individually along with a slider on which the listener was asked to rate the mix clarity they perceived between 0 and 100 in steps of 1. A score of 0 indicated the stimulus was perceived to be unclear (no clarity), and a score of 100 indicated the stimulus was perceived to be clear (highest clarity). The following labels were given according to a standard 5-point category rating scale [36]:

‘Excellent’, ‘Good’, ‘Fair’, ‘Poor’, and ‘Bad’, in descending order. In total, 6 repeats were performed where the order of the stimuli was randomised each time.

3.3. Subjective Listening Test Results

The listeners’ results were screened based on the consistency of their repeat scores, as a lack of consistency between repeats showed the listener may not have had a consistent perception of mix clarity or may not have understood the task at hand. Repeat consistency was measured using the interclass correlation coefficient (ICC) between a given listener’s repeats. Estimated ICCs and their 95% confidence intervals were calculated using SPSS [37]. The ICC was a mean-rating, absolute agreement, 2-way mixed effects model as recommended in [38]. Ratings from listeners whose repeat scores that had an ICC of below 0.75 were excluded from the final analysis, as it is suggested ratings with an ICC equal to or greater than 0.75 have a ‘good’ reliability [38]. In this case, the lower bound of the ICC estimates’ 95% confidence intervals were taken as the value which needed to exceed the 0.75 threshold, as this ensured with 95% confidence the listeners had ‘good’ reliability. After post-screening, only ratings from the 15 most consistent listeners remained.

The mean of each listener repeat rating was taken as the listener’s subjective clarity score for a given stimulus. The median of all listeners’ mean ratings for each stimulus was then taken as its median clarity score (MCS). Figure 2 shows these median subjective scores per stimuli, also indicating interquartile ranges, and outliers.

Figure 2. Box plot of median subjective clarity scores collected in the subjective listening test.

The assumption of homogeneity of variances for ANOVA was not met, indicated by both Levene’s and Barlett’s tests showing $p < 0.05$ for all stimuli. As such, a multiple pairwise Wilcoxon test was performed to identify significantly different MCSs. Holm correction was applied to account for multiple comparisons. The results affirm what is indicated by interquartile ranges shown in Figure 2, in that, stimuli whose interquartile ranges do not overlap were also shown as being significantly different ($p < 0.05$) in the pairwise Wilcoxon analysis.

3.4. Independent Dataset

An independent dataset was synthesised to investigate the effect of specific signal degradation and how it would impact the predicted clarity score. For this, the stimuli used in the subjective listening test (Section 3.1) were processed in 3 different ways, with 7 levels of severity, creating 3 test datasets. The processing was intended to gradually decrease the level of perceptual clarity at each level of severity by gradually introducing masking from noise-like signals common in music production. By forming the dataset from the stimuli which had been rated subjectively by listeners, the effect of the degradation on the proposed model could be evaluated relative to the MCSs the stimuli received.

The 3 sets included:

- Set 1—Addition of broadband pink noise in 6dB steps (−36dBFS, −30dBFS, −24dBFS, −18dBFS, −12dBFS, −6dBFS, 0dBFS). This degradation simulates a gradually raising noise floor or an extremely ‘busy’ mix where lots of conflicting signals become noise-like.
- Set 2—Addition of reverberation in 6dB steps of wet mix level (−36dBFS, −30dBFS, −24dBFS, −18dBFS, −12dBFS, −6dBFS, 0dBFS). In this case, the reverberation was meant to simulate a large hall, and thus had a diffuse tail with a decay time of 3 s, with decay times more pronounced for low frequency energy rather than high frequency energy. Artificial reverberation is commonly added in music production to enhance the sense of space and listener envelopment. However, it also disperses diffuse energy over time adding to the signal’s noise floor and can smear transient energy over time making them less clearly defined.
- Set 3—Clipping applied at a ceiling calculated as a percentile range of amplitude values about the 50th percentile (90%, 80%, 70%, 60%, 50%, 40%, 30%), such that clipping applied is uniform regardless of signal amplitude. Clipping is often used as a creative effect, though also occurs in signals recorded without an appropriate level of headroom. This can cause a reduction in dynamic range, decreasing the energy difference between the signal peaks and the noise floor. In addition, clipping can produce additional harmonic content making the signal more spectrally dense, thus increasing the potential for masking to occur. However, increasing spectral density and brightness of transient and steady-state components may be beneficial to the perception of clarity in some cases, whereby the masking potential of the transients and steady-state components in the signal are increased greater than the masking potential of the residual component.

Each of the test signals generated were loudness normalised to −23LU after the signal degradation had been applied.

4. Model Evaluation

To evaluate the model’s performance, the outputs of the clarity model variants were correlated against the MCSs obtained from the controlled listening test detailed in Section 3.3. Both Pearson’s correlation coefficient r and Spearman’s rank correlation coefficient ρ were calculated, as a lack of a bivariate normal distribution or linear relationship between the two variables can cause the Pearson’s coefficient to provide an inaccurate measure of association. Pearson’s coefficient is still provided, as a non-linear relationship is suggested in cases where ρ is greater than r , and a linear relationship where r is greater than ρ . The Spearman’s rank coefficient is non-parametric, so is more robust, and is the focus of the following analysis. Both correlation coefficients for each model variation are given in Table 2 for the sake of comparison.

Table 2. Pearson and Spearman rank correlations of tested metrics against MCSs.

Measure	<i>r</i>	<i>rho</i>
L3PM Clarity Model	−0.6682	−0.6882
Modified L3PM Clarity model	−0.7868	−0.8088
L2PM Clarity Model	−0.7884	−0.8382

All coefficients shown are significant where $p < 0.01$.

When correlating the ranked L2PM clarity model masking scores against the MCSs, a strong negative Spearman rank correlation of $\rho = -0.8382$, $p < 0.01$ was achieved. This correlation is shown in Figure 3, where the box plots indicate the distribution about the MCSs obtained, ordered by rank determined by the overall masking score calculated by the proposed model.

Figure 3. Spearman rank correlation between overall masking score ranks calculated using the L2PM clarity model median clarity scores.

This was the highest Spearman rank correlation achieved by any implementation of the model, with most deviations from the line of best fit still within the interquartile range of the MCSs. The least well ranked stimulus, '126410', was also poorly predicted by a different mix clarity model suggested in previous work [20]. This implementation also achieved $r = -0.7884$, $p < 0.01$, which suggests a strong and slightly non-linear parametric relationship between the model output and the subjective scores.

The model was also implemented using the L3PM, forming the L3PM clarity model. This achieved a significant, but weaker correlation of $\rho = -0.6882$, $p < 0.01$, shown in Figure 4. Again, a lesser Pearson correlation ($r = -0.6682$, $p < 0.01$) was seen, indicating a somewhat non-linear relationship.

Figure 4. Spearman rank correlation between overall masking score ranks calculated using the L3PM clarity model and median clarity scores.

This correlation shows some heteroscedasticity, where the stimuli given higher masking scores were predicted less accurately than those given low masking scores, though many of the interquartile ranges cross the line of best fit. The weaker correlation of this model compared to the L2PM clarity model was somewhat unexpected, as the L3PM is more efficient in coding applications, and the SMR values are calculated in scale-factor bands which approximate critical bands. These are more reflective of perception than the L2PM scale-factor bands which are linearly spaced. The better performance of the L2PM clarity model suggests a greater importance of masking occurring between higher frequency energy, due to the model calculating the overall masking of a given frame in each component as the sum of masking within the scale-factor bands (Equation (8)). Other work has suggested an importance of high frequency energy in clarity perception for some single instrument sounds [7,12]. The L3PM also employs window switching which the L2PM implementation does not [24]; whereby, shorter analysis windows and fewer scale-factor bands are used in determining the energy and masking thresholds of highly transient frames. For the stimuli used in the present testing, the perceptual entropy threshold was not crossed by any frames of any of the stimuli. Therefore, short analysis frames were not used in any of the masking calculations and thus could not be responsible for the difference in performance between the L2PM and L3PM clarity model variants.

To confirm that the difference in scale-factor bands was largely responsible for the weaker correlation of the L3PM clarity model, a modified version of the L3PM was devised. This calculated the energy and masking thresholds following the L3PM [23,24], however, rather than calculating the SMR directly in scale-factor bands which approximate critical bandwidth, the masking thresholds were spread back across the spectral bins of the Fourier transform and the SMR was calculated for the linearly distributed bands specified in the

L2PM [23]. This saw an increase in performance to a level similar to, but lower than the L2PM clarity model, achieving $\rho = -0.8088$, $p < 0.01$ and is shown in Figure 5.

Figure 5. Spearman rank correlation between overall masking score ranks calculated using the modified L3PM clarity model and median clarity scores.

The modified L3PM clarity model improves the position of the L3PM clarity model's most severe outlier, '94414', though a number of outliers remain. The remaining differences between the L2PM and modified L3PM clarity models were due to the difference in how the models calculate the masking threshold. Whilst this implementation of the clarity model had a slightly weaker Spearman and Pearson correlation ($r = -0.7868$, $p < 0.01$) than the L2PM clarity model, the Spearman and Pearson correlation coefficient values are similar, suggesting a more linear relationship between the modified L3PM clarity model and the subjective scores.

Further to testing the clarity model variations' correlations to the subjective data, they were also evaluated using the independent dataset (Section 3.4). Figures 6–8 show the scores calculated for audio examples at the various levels of degradation for each of the three degradation methods. Box plots show median and interquartile range as well as outliers at each degradation level.

Figure 6 shows clarity scores calculated by the L2PM clarity model for the independent dataset. The addition of pink noise sees the addition of broadband residual energy resulting in the masking of transient and steady-state components of the stimuli. As such, the masking scores progressively rise as more noise is added, until they reach a ceiling level where the TSS component is maximally masked. This ceiling causes a convergence of the clarity scores for the tested stimuli, with the tracks scoring a lower MCS in the subjective test tending to see less change when degraded by pink noise. This suggests that how noise-like the stimuli were may have had an influence on the MCSs they received from listeners.

Figure 6. L2PM clarity model overall masking scores (Equation (10)) of the independent dataset at the seven levels of degradation for each method (detailed in Section 3.4), including the reference level of the examples without degradation.

The addition of reverberation had a similar though less extreme effect as adding pink noise. Unlike pink noise, the reverberated signal is still derived from and therefore correlated to the original signal. If viewed using a spectrogram, reverberation is somewhat akin to blurring an image where the energy is smeared over time, and to a lesser extent frequency. This smeared energy is largely characterised by the separation system as residual energy, increasing the level to which the TSS component is masked by the R component, and decreasing the level to which the R component is masked by the TSS component. The addition of reverberation tended to increase the masking score of the stimuli. However, a greater effect is seen for stimuli which received a higher MCS similarly to the addition of pink noise. Moreover, masking scores for stimuli which did not contain strong transient onsets from things such as drum hits, such as ‘137167’, ‘15541’, and ‘94414’ also showed less difference than those which did. This suggests the masking effect of reverberation is more severe for transient sounds, which is in line with the greater effect reverberation has on the temporal axis of the spectrogram.

Clipping appears to cause two different responses in the model output whereby some stimuli masking scores increase with degradation level and others decrease. The effect of this is almost symmetrical, keeping the median masking score relatively constant throughout the degradation levels whilst the interquartile range increases. Clipping had the effect of reducing AM_R (Equation (8)), as the dynamic range was decreased and the difference between the R and TSS components was reduced. However, in cases where there was very little residual energy, the AM_{TSS} (Equation (8)) was also reduced. Clipping can increase the harmonic density of transients which emphasises their spectrally broadband characteristic and thus increases the level of transient energy separated by the separation system. In

cases where the present signal is largely steady-state energy, such as a sustained Rhodes piano chord, clipping can emphasise the temporally constant and more slowly evolving nature of steady-state energy and increase the level of steady-state energy separated by the separation system. Moreover, additional harmonics generated from clipping such a signal are spectrally narrowband and temporally constant, and as such are considered to be steady-state by the separation system, further contributing to the level of steady-state energy extracted. Therefore, stimuli that were more noise-like at the reference level, receiving high masking scores, received even higher masking scores as more degradation was applied. Conversely, stimuli that had a low level of residual energy at the reference level received even lower masking scores at higher degradation levels, as the reduction of the R component masking was lesser than the reduction of TSS component masking. Whilst the effects of clipping may potentially increase the perception of clarity, given the more complex effect compared to the addition of pink noise or reverberation, the clipping applied at the highest levels of degradation is extreme. The resulting signals are very noise-like, and as such were expected receive high masking scores, which they did not in the case of stimuli that scored middle or low masking scores at the reference level.

Figure 7. L3PM clarity model overall masking scores (Equation (10)) of the independent dataset at the seven levels of degradation for each method (detailed in Section 3.4), including the reference level of the examples without degradation.

Figure 8. Modified L3PM clarity model overall masking scores (Equation (10)) of the independent dataset at the seven levels of degradation for each method (detailed in Section 3.4), including the reference level of the examples without degradation.

The L3PM clarity model responded similarly to the L2PM clarity model for all degradation types. Figure 7 indicates that there was an increase of median masking score and convergence with increasing levels of degradation in the case of additional pink noise and reverberation. Moreover, there were somewhat diverging masking scores with a relatively constant median in the case of increasing levels of clipping. However, this response was muted in comparison to the response seen from the L2PM clarity model. Application of the signal degradation had a greater effect on the separated higher frequency energy than on the separated low frequency energy; this was as a result of the STFT based separation system employed having frequency resolution which increases with bin index. When measured using linearly spaced scale-factor bands of the L2PM, the high frequency bins represent a larger proportion of the scale-factor bands than when measured with more perceptually aligned scale-factor bands of the L3PM. Thus, less change in masking score is seen when degradation is applied in the case of the L3PM clarity model's perceptually spaced bands, as fewer of these scale-factor bands correspond to the high frequency bins which have the greatest difference in separated energy between degradation levels. While the L3PM scale-factor bands are more aligned with human perception, in this case, a negative effect on correlation to the subjective scores was seen (Figures 3 and 4).

Figure 8 shows the clarity model which employed the modified L3PM, using linearly grouped scale-factor bands like those used in the L2PM clarity model. The modified L3PM clarity model responded to the different types of degradation similarly to the L2PM clarity model (Figure 6). The similarity of these results, and difference to the L3PM clarity model results (Figure 7), suggests that the differences between the models' responses could have largely been caused by the different scale-factor bands employed. The remaining

differences then are due to the difference in masking threshold calculation between the L2PM and L3PM clarity models discussed previously, and detailed in [23].

5. Discussion

All tested correlations showed significant ($p < 0.01$) results, with a worst-case correlation of $r = -0.6682$, $\rho = -0.6882$ to the subjective scores. This suggests a link between the underlying concept of transient, steady-state and residual masking and mix clarity perception. However, as the tested subjective dataset contained only 16 stimuli, the model may be overfit, and a larger test containing more participants and stimuli should be performed to validate these results.

Previous work proposed a model of Inter-Band Relationship Analysis (IBR), which measures the correlation between dynamic ranges in three bands of the frequency spectrum over time [20,21]. The best performing L2PM model outperformed the IBR model scoring $\rho = -0.8382$ ($p < 0.01$) versus $\rho = 0.7882$ ($p < 0.01$). As the IBR model simply utilises the dynamic range across three bands, it only has an incidental relationship to human perception. Furthermore, the IBR model is prone to being ‘fooled’ by some signals. For example, it would rate sustained pink noise and sustained full-bandwidth harmonic signals equally. The present model supersedes IBR by negating these shortcomings, and by relating more directly to human perception of auditory masking. The model also incorporates a greater awareness of the makeup of the signal being measured due to the TSR component separation employed.

The L2PM clarity model showed the strongest ρ correlation, though was non-linear, with a Spearman’s rank correlation test showing a stronger relationship than the Pearson correlation. This implementation of the model is also the least computationally expensive, as the L2PM is less complex than the L3PM. During independent testing, the model responded as expected to the addition of pink noise and reverberation, with the masking score increasing (clarity decreasing) as the degradation levels increased. The model’s response to clipping was less uniform than that of the additional pink noise or reverberation, showing a more complex effect on the relationship of the R and TSS components of the signal. While the response is understandable in terms of how the model operates, it is not expected to be congruent with perception. Low perceived clarity scores, corresponding to high masking scores, would be uniformly expected for all stimuli at the highest levels of clipping degradation, showing a potential shortcoming of the model. However, given the strong correlation to the subjective scores, this shortcoming did not seem to have a greatly detrimental effect for the tested stimuli in this case. Though, improving the response to this kind of degradation would improve the robustness of the model in extreme cases.

While the L3PM has a more complex calculation of masking threshold and provides more efficient encoding in the context of MPEG compression compared to the L2PM, the L3PM clarity model unexpectedly had the weakest correlation to the subjective scores. It is suggested this was largely due to L3PM’s perceptually aligned scale-factor bands not providing emphasis on masking occurring in high frequency bands as in the L2PM and modified L3PM variations of the clarity model. This could indicate a greater importance of masking occurring between high frequency energy to the perception of clarity. In terms of independent testing, the L3PM clarity model had a similar but less extreme response to all three degradation types than the L2PM clarity model. The L3PM is capable of calculating masking thresholds using shorter windows when transients occur, providing a higher time resolution and reducing pre-ringing artefacts in MPEG coding. Similarly, this window switching could be used in the clarity model to provide greater temporal resolution for masking occurring relating to transient passages in the signal under test. In the present testing only long windows were used, as none of the stimuli under test had transient content capable of triggering the short frame calculation. If a more appropriate onset detection method was used, the application of shorter windows may improve the clarity model’s performance.

A modified version of the L3PM was devised to employ linearly grouped scale-factor bands like the L2PM. The performance of the modified L3PM clarity model response was similar to the L2PM clarity model in both correlation with the subjective scores and to the independent dataset, affirming the difference in scale-factor bands was largely responsible for the difference in performance between the L2PM and L3PM clarity models. While the ρ and r coefficients showed a slightly weaker correlation than that of the L2PM clarity model, the coefficients were more similar in value, which indicates a closer linear relationship to the subjective data. Additionally, being based on the L3PM clarity model, this model may also benefit from improving the onset detection system used for window switching.

6. Conclusions

A new perceptually motivated model for the prediction of mix clarity was proposed, based on the masking relationship between residual, transient and steady-state components of a musical signal. The model consists of a median filter-based separation system, which feeds the MPEG Psychoacoustic Model II used to calculate signal-to-mask ratios of the component parts which were then compared. Both layer 2 and layer 3 implementations of the MPEG Psychoacoustic Model II were tested along with a modified version of the layer 3 implementation, forming L2PM, L3PM, and modified L3PM variants, respectively.

Each variation was evaluated through both Pearson and Spearman's rank correlation to subjective scores gathered in a controlled listening test. Their response to an independent dataset of stimuli degraded through the addition of pink noise, reverberation, and clipping was also evaluated. The L2PM clarity model showed the strongest correlation to the subjective scores, followed by the modified L3PM, with the L3PM clarity model showing the weakest relationship. Although the L3PM is most efficient in coding applications, the stronger correlation achieved by the modified L3PM clarity model showed that the linearly grouped scale-factor bands of the L2PM were advantageous to performance in this case. All variations of the model responded similarly to the degradation introduced in the independent dataset; the L3PM clarity model's response was less extreme to that of the modified L3PM and L2PM clarity models, whose responses were very similar. Addition of pink noise and reverberation caused an increase in masking score, reflecting a decrease in clarity. Clipping caused a somewhat more complex response, where stimuli which were noise-like and received high masking scores at their reference level gained higher masking scores when clipped. Stimuli that had a low level of residual energy and a low masking score at their reference level received even lower masking scores when clipped.

Further work is ongoing to validate the proposed model's performance against a larger subjective data set.

Author Contributions: Conceptualisation, A.P. and S.F.; methodology, A.P. and S.F.; software, A.P.; validation, A.P.; formal analysis, A.P.; investigation, A.P.; resources, A.P.; data curation, A.P.; writing—original draft preparation, A.P.; writing—review and editing, S.F.; visualisation, A.P.; supervision, S.F.; project administration, S.F.; funding acquisition, S.F. All authors have read and agreed to the published version of the manuscript.

Funding: The article processing charge (APC) was funded by The University of Huddersfield.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Acknowledgments: The authors would like to thank participants of the subjective listening test. Moreover, the authors would like to thank the reviewers, whose insightful comments contributed to the quality of the work presented.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Fenton, S.; Lee, H. A Perceptual Model of ‘Punch’ Based on Weighted Transient Loudness. *J. Audio Eng. Soc.* **2019**, *67*, 429–439. [\[CrossRef\]](#)
2. Parker, A.; Fenton, S.; Lee, H. A Real-Time System for the Measurement of Perceived Punch. In Proceedings of the 145th Audio Engineering Society Convention, New York, NY, USA, 17–20 October 2018; pp. 1–11.
3. Williams, D.; Brookes, T. Perceptually-Motivated Audio Morphing: Brightness. In Proceedings of the 122nd Audio Engineering Society Convention, Vienna, Austria, 5–8 May 2007; pp. 1–9.
4. McAdams, S.; Winsberg, S.; Donnadieu, S.; de Soete, G.; Krimphoff, J. Perceptual Scaling of Synthesized Musical Timbres: Common Dimensions, Specificities, and Latent Subject Classes. *Psychol. Res.* **1995**, *58*, 177–192. [\[CrossRef\]](#) [\[PubMed\]](#)
5. ITU-R, BS.1770-4 Algorithms to Measure Audio Programme Loudness and True-Peak Audio Level BS Series Broadcasting Service (Sound); ITU-T/ITU-R/ISO/IEC 1770-3:2015; ISO: Geneva, Switzerland, 2015.
6. Pedersen, T.H.; Zacharov, N. The Development of a Sound Wheel for Reproduced Sound. In Proceedings of the 138th Audio Engineering Society Convention, Warsaw, Poland, 7–10 May 2015; pp. 1–13.
7. Disley, A.C.; Howard, D.M. Spectral Correlates of Timbral Semantics Relating to the Pipe Organ. *TMH-QPSR* **2004**, *46*, 25–40.
8. Fenton, S.; Wakefield, J. Objective Profiling of Perceived Punch and Clarity in Produced Music. In Proceedings of the 132nd Audio Engineering Society Convention, Budapest, Hungary, 26–29 April 2012; pp. 1–15.
9. Hermes, K. *Towards Measuring Music Mix Quality: The Factors Contributing to the Spectral Clarity of Single Sounds*; University of Surrey: Guildford, UK, 2017.
10. Ronan, D.; Ma, Z.; Namara, P.M.; Gunes, H.; Reiss, J.D. Automatic Minimisation of Masking in Multitrack Audio using Subgroups. *arXiv* **2018**, arXiv:1803.09960.
11. Toole, F.E. Subjective Measurements of Loudspeaker Sound Quality and Listener Performance. *J. Audio Eng. Soc.* **1985**, *33*, 2–23.
12. Hermes, K.; Brookes, T.; Hummersone, C. The Harmonic Centroid as a Predictor of String Instrument Timbral Clarity. In Proceedings of the 140th Audio Engineering Society Convention, Paris, France, 4–7 June 2016; pp. 1–10.
13. de Man, B.; Reiss, J.D. A knowledge-engineered autonomous mixing system. In Proceedings of the 135th Audio Engineering Society Convention, New York, NY, USA, 17–20 October 2013; pp. 281–291.
14. Tom, A.; Reiss, J.; Depalle, P. An Automatic Mixing System for Multitrack Spatialization for Stereo Based on Unmasking and Best Panning Practices. In Proceedings of the 146th Audio Engineering Society Convention, Dublin, Ireland, 20–23 March 2019.
15. Perez-Gonzalez, E.; Reiss, J. Automatic Equalization of Multi-Channel Audio Using Cross-Adaptive Methods. In Proceedings of the 127th Audio Engineering Society Convention, New York, NY, USA, 9–10 October 2009; pp. 1–6.
16. Griesinger, D. The Importance of the Direct to Reverberant Ratio in the Perception of Distance, Localization, Clarity, and Envelopment. In Proceedings of the 126th Audio Engineering Society Convention, Munich, Germany, 7–10 May 2009; pp. 1–13.
17. Beerends, J.G.; van Nieuwenhuizen, K.; van den Broek, E.L. Quantifying sound quality in loudspeaker reproduction. *J. Audio Eng. Soc.* **2016**, *64*, 784–799. [\[CrossRef\]](#)
18. Rix, A.W.; Beerends, J.G.; Hollier, M.P.; Hekstra, A.P. Perceptual evaluation of speech quality (PESQ)—A new method for speech quality assessment of telephone networks and codecs. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221), Salt Lake City, UT, USA, 7–11 May 2001; Volume 2, pp. 749–752.
19. Thiede, T.; Treurniet, W.C.; Bitto, R.; Schmidmer, C.; Sporer, T.; Beerends, J.G.; Colomes, C. PEAQ: The ITU Standard for Objective Measurement of Perceived Audio Quality. *J. Audio Eng. Soc.* **2000**, *48*, 29.
20. Parker, A.; Fenton, S. Mix Clarity Prediction using Multiresolution Inter-Band Relationship Analysis. In Proceedings of the 148th Audio Engineering Society Convention, Online, 2–5 June 2020; pp. 1–11.
21. Fenton, S.; Fazenda, B.; Wakefield, J. Objective Measurement of Music Quality Using Inter-Band Relationship Analysis. In Proceedings of the 130th Audio Engineering Society Convention, London, UK, 13–16 May 2011; pp. 1–10.
22. Zwicker, E.; Fastl, H. *Psychoacoustics*; Springer: Berlin, Germany, 1990.
23. *Information Technology-Coding of Moving Pictures and Associated Audio for Digital Storage Media at up to about 1, 5 mbit/s-part 3: Audio*; ISO/IEC 11172-3:1995; ISO: Geneva, Switzerland, 1995.
24. Thiagarajan, J.J.; Spanias, A. *Analysis of the MPEG-1 Layer III (MP3) Algorithm Using MATLAB*; Morgan & Claypool: San Rafael, CA, USA, 2011; Volume 3.
25. Aichinger, P.; Sontacchi, A.; Schneider-Stickler, B. Describing the Transparency of Mixdowns: The Masked-to-Unmasked-Ratio. In Proceedings of the 130th Audio Engineering Society Convention, London, UK, 13–16 May 2011; pp. 1–10.
26. Tsilfidis, A.; Papadakos, C.; Mourjopoulos, J. Hierarchical Perceptual Mixing. In Proceedings of the 126th Audio Engineering Society Convention, Munich, Germany, 7–10 May 2009; pp. 1–7.
27. Hennequin, R.; Khelif, A.; Voituret, F.; Moussallam, M. Spleeter: A Fast and Efficient Music Source Separation Tool with Pre-Trained Models. *J. Open Source Softw.* **2020**, *5*, 2154. [\[CrossRef\]](#)
28. Tachibana, H.; Ono, N.; Kameoka, H.; Sagayama, S. Harmonic/Percussive Sound Separation Based on Anisotropic Smoothness of Spectrograms. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2014**, *22*, 2059–2073. [\[CrossRef\]](#)
29. Bello, J.P.; Sandler, M. Phase-based note onset detection for music signals. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP ’03), New Paltz, NY, USA, 19–22 October 2003; Volume 5, pp. 441–444.
30. Driedger, J.; Müller, M.; Disch, S. Extending Harmonic-Percussive Separation of Audio Signals. In Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR 2014), Taipei, Taiwan, 27–31 October 2014; pp. 611–616.

31. Lagrange, M.; Raspaud, M.; Badeau, R.; Richard, G. Explicit Modeling of Temporal Dynamics within Musical Signals for Acoustical Unit Similarity. *Pattern Recognit. Lett.* **2010**, *31*, 1498–1506. [[CrossRef](#)]
32. Fenton, S.; Lee, H.; Wakefield, J. Hybrid Multiresolution Analysis Of ‘Punch’ In Musical Signals. In Proceedings of the 138th Audio Engineering Society Convention, Warsaw, Poland, 7–10 May 2015; pp. 1–10.
33. Defferrard, M.; Benzi, K.; Vandergheynst, P.; Bresson, X. FMA: A Dataset For Music Analysis. In Proceedings of the 18th International Society for Music Information Retrieval Conference, Suzhou, China, 23–27 October 2017; pp. 1–8.
34. Cycling '74. MAX. Available online: <https://cycling74.com/> (accessed on 10 March 2018).
35. Gribben, C.; Lee, H. Towards the Development of a Universal Listening Test Interface Generator in Max. In Proceedings of the 138th Audio Engineering Society Convention, Warsaw, Poland, 7–10 May 2015; pp. 1–6.
36. ITU-R. *R BS.1284-2 General Methods for the Subjective Assessment of Sound Quality*; ITU-R: Geneva, Switzerland, 2019; Volume 2.
37. IBM Corp. *IBM SPSS Statistics for Windows*; IBM Corp.: Armonk, NY, USA, 2016.
38. Koo, T.K.; Li, M.Y. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *J. Chiropr. Med.* **2016**, *15*, 155–163. [[CrossRef](#)] [[PubMed](#)]