Audio Engineering Society

Convention Paper 9762

Presented at the 142nd Convention
2017 May 20–23 Berlin, Germany

Long-term Average Spectrum in Popular Music and its Relation to the Level of the Percussion

Anders Elowsson[1], Anders Friberg[1]

[1] *KTH Royal Institute of Technology, School of Computer Science and Communication, Speech, Music and Hearing*

Correspondence should be addressed to Anders Elowsson (elov@kth.se)

ABSTRACT

The spectral distribution of music audio has an important influence on listener perception, but large-scale characterizations are lacking. Therefore, the long-term average spectrum (LTAS) was analyzed for a large dataset of popular music. The mean LTAS was computed, visualized, and then approximated with two quadratic fittings. The fittings were subsequently used to derive the spectrum slope. By applying harmonic/percussive source separation, the relationship between LTAS and percussive prominence was investigated. A clear relationship was found; tracks with more percussion have a relatively higher LTAS in the bass and high frequencies. We show how this relationship can be used to improve targets in automatic equalization. Furthermore, we assert that variations in LTAS between genres is mainly a side-effect of percussive prominence.

1 Introduction

1.1 Long-term average spectrum in music

Musical instruments produce a wide range of spectra [1]. When instrumental performances are mixed together in a music recording, the frequency distribution of that recording will be determined by the included instruments, the performance (e.g. the dynamics) of the musicians [2], and the technical choices, such as the equalization made by e.g. the recording and mixing engineer. The spectral distribution affects listener perception. As an example, mixing engineers generally associate descriptive words with different frequencies, e.g., *presence* = 4-6 kHz, *brilliance* = 6-16 kHz [3].

By measuring the long-term average spectrum (LTAS) of a musical signal, it is possible to get a compact representation that mediates some of these perceptual qualities of the mix. Due to variations in e.g. instrumentation, the LTAS can be expected to vary somewhat between songs. The average LTAS across different recordings, the extent of the variations in LTAS, and the factors behind the variations, should all be useful knowledge in audio engineering and to applications in automatic equalization.

1.2 Methodology of earlier studies

The frequency spectrum of music has been studied previously, however, often for so small datasets and with such wide-ranging methodologies that comparisons are difficult to make. An early study that analyzed the effect of the bandwidth of the filters used a small dataset filtered with octave-spaced bands [4]. Another study used 51 mel scale filter bands [5]. Benjamin [6], studied peak and RMS spectra for a dataset of 22 tracks with one-third octave bands. A few excerpts of popular music were also analyzed with one-third octave bands by the British Broadcasting Corporation [7]. The study closest to ours is one by Pestana et al. [8], who analyzed the LTAS of 772 commercial recordings, and connected the results to production year and genre.

One type of application that could benefit from LTAS analysis of big datasets is automatic equalization in mixing and mastering. The field is fairly new, with researchers trying out different techniques and focusing on different aspects. One example is to equalize

with the goal of achieving equal perceptual loudness for every frequency band across every multi-track [9]. In another implementation, a simple machine learning model was trained, based on listener input, to equalize isolated sounds [10]. In [11], the mean LTAS computed in [8] was smoothed with a moving average filter, and an IIR filter developed to match this curve during equalization. A good overview of methods and challenges for automatic mixing is given in [12].

1.3 Relation between LTAS and the level of the percussion

In a previous study of LTAS in commercial recordings, differences between genres were analyzed [8]. One of the main findings was that genres such as hip-hop, rock, pop and electronic music had louder low frequencies (up to 150 Hz) than genres such as jazz and folk music. The same relationship was evident also for the high frequencies (5 kHz and above), with hip-hop, rock, pop and electronic music being the loudest. The differences in the mean LTAS were clear: some genres have a (relatively) louder low-end and high-end of the spectrum, whereas other genres such as jazz and folk music generally have a (relatively) higher sound level in the mid frequencies. Similar differences were found between popular music and opera music in a smaller study [13].

Why is this? Although certain genres have somewhat stylistic preferences with regards to the LTAS (a prime example being the heavy bass in reggae), mastering engineers generally try to keep the "symphonic tonal balance" as a basic reference for most pop, rock, jazz and folk music [14]. Could there then be something else than the general genre that give rise to the differences in LTAS? Given that genres that were found to have a higher sound level in the low end and high end of the spectrum have more emphasis on rhythm, the relative level of the rhythm instruments seems to be a relevant factor. In this study, we will explore the relationship between the sound level of the percussive instruments in a musical mixture and the LTAS.

1.4 Applications of LTAS analysis

A well-balanced frequency spectrum is an important aspect of a mix [15] and something that both mixing engineers and mastering engineers try to achieve. There are some guidelines that can be used as a starting point: music generally exhibits a dip in energy toward higher frequencies. This *spectrum slope* has been estimated across genre to be approximately 5 dB/octave on average [8]. However, the slope steepens for higher frequencies. It would therefore be interesting to study the characteristics of the slope in finer detail. There are many factors (such as the instrumentation previously mentioned) that must be considered to achieve an appropriate equalization. As acknowledged by audio engineer Neil Dorfsman, it is common to use spectral analysis and comparisons with *reference mixes* during mixing [8]. Reference mixes are mixed tracks (often successful commercial recordings) that represent a target, *e.g.*, for a desirable spectral balance. The choice of reference tracks is often adapted to the source material being mixed [15]. For example, when mixing a track with no drums, it is beneficial to use reference tracks without drums.

Mixing engineers can use a few tracks as reference points when adjusting the frequency balance of the mix, but with the availability of large datasets of music recordings, it is possible to analyze the frequency spectrum of thousands of songs. The mean LTAS of these songs can be used as a *target spectrum* [11], and the processed mix can be altered to better coincide with the LTAS of the target. The usefulness of the target spectrum should increase if it is based on recordings with a similar instrumentation as the processed mix, a strategy that would mimic how engineers chooses their reference tracks. A successful system will therefore need to disentangle the factors (such as the instrumentation) that has an effect on LTAS, and then take these factors into account when establishing an appropriate target LTAS. In this study, we will try to use information about the level of the percussion in the tracks to achieve better targets.

1.5 Outline of the article

In Section 2 we give a summary of the dataset (Section 2.1), describe the signal processing used to calculate the LTAS of each track (Section 2.2) and compute the amount of percussion (L_{perc}) in the tracks (Section 2.3). The mean and variance in LTAS across the dataset is explored in Section 3.1, and in Section 3.2 an equation for the mean LTAS is computed and

used to derive the spectrum slope, including its variation across frequency. In Section 4 we relate L_{perc} to LTAS, and find a frequency range where the spectrum slope is the same regardless of the amount of percussion. In Section 5 we show that LTAS targets for automatic equalization can be improved for low and high frequencies by incorporating information about percussive prominence. In Section 6, the different findings of the study are discussed.

2 Data and Signal Processing

2.1 Dataset

We used 12345 tracks of popular music in the study. These were selected from a slightly bigger dataset of 12792 tracks by excluding monophonic songs and songs with a playing time longer than 10 minutes. The tracks were chosen mainly from artists that have made a significant impact in popular music. Furthermore, artists that use a wide variation in the amount of percussion in their recordings were prioritized. The latter condition resulted in a somewhat larger focus on folk pop music. The artists with the largest number of tracks in the whole dataset were Bob Dylan, Neil Young, Bonnie 'Prince' Billy, The Beatles and Bright Eyes. Tracks had been mastered or remastered for the compact disc (CD) format.

Some of the analyzed tracks had previously been converted to the MP3 format. This was convenient due to the large size of the dataset. MP3 files have a high-frequency cut-off adapted to the limits of human hearing. Between 30 Hz and 15.7 kHz the LTAS of a PCM-encoded track and an MP3 version of that track are fairly similar. We therefore performed our analysis in this frequency range. To verify the similarity, we calculated the absolute difference in LTAS of 30 PCM-encoded tracks and the same tracks converted to MP3 files (192 kHz). The same settings were used for the computation of LTAS and conversion to a log-frequency spectrum as outlined in Section 2.2 and 3.2. The mean absolute change in LTAS over the frequency bins was just below 0.06 dB on average for the 30 tracks, which we regard as an acceptably small range of deviation.

2.2 Calculating long-term average spectrum

The LTAS was calculated for each mixture with *ioSR Matlab Toolbox* [16]. First, the short-term frequency transform (STFT) was calculated for both of the stereo channels, using a window size of 4096 samples (approx. 93 ms at the 44.1 kHz sample rate used), and a hop size of 2048 samples (approx. 46 ms). The mean power spectral density (PSD) of each channel was subsequently calculated from each spectrogram. The loudness standard specification ITU-R BS.1770-4 [17] contains a second stage weighting curve, which resembles C-weighting for loudness measurements. A filter, F_{ITU}, was computed as the element-wise power of the frequency response derived from the filter coefficients of this weighting curve. The normalized PSD (PSD_N) was then computed using the filter F_{ITU}

$$\mathrm{PSD}_N = \frac{\mathrm{PSD}}{\mathrm{PSD} \cdot F_{ITU}}, \qquad (1)$$

where · represents the dot product that results in a scalar used as the normalization factor. The filtering during loudness normalization does not affect the relative sound level of the different frequency bins in each track, so the *mean* LTAS computed in Section 3.1 is unaffected in this regard. It will however have a small effect on the *standard deviation* and *order statistics*, as the LTAS of songs with a lot of sub-bass, for example, will be normalized to a different level.

At this point, each song was represented by a vector of 2049 frequency bins (PSD_N), covering the range of 0-22050 Hz. We combined the vectors of all 12345 mixtures to form the 12345×2049 matrix *M*. In earlier works, it was shown that the average spectrum will be influenced by tonal harmonics in the upper frequencies [8]. This effect is a reflection of common keys (and/or pitches) in western music. To remove these peaks and make the average spectrum less affected by tonality, we included an adjunct processing step where *M* was smoothed across frequency. The power spectrum was in this step smoothed by a Gaussian filter with a -3 dB-bandwidth of $1/6^{th}$ of an octave (one-sixth octave bands). This bandwidth is somewhat narrower than that of previous studies (one-third octave bands), a deliberate choice to ensure that subtle variations in the LTAS are retained. The standard devia-

tion σ for the Gaussian at each frequency F was obtained from the bandwidth b = 1/6 according to the equation used in the *ioSR Matlab Toolbox* [16]

$$\sigma = \frac{Fb}{\pi}. \quad (2)$$

The filtered and unfiltered versions of M were both converted to signal level by computing the log-spectrum, LTAS = $10 \log_{10} \mathrm{PSD_N}$. By converting to signal level before averaging over songs, the influence from spectral bins of the different tracks will be more directly connected to their perceived loudness.

2.3 Harmonic/percussive source separation

In order to measure the effect of the amount of percussion in the audio, we first performed source separation on both stereo channels of the tracks, and then calculated the RMS of the percussive waveform in relation to the unfiltered waveform. The harmonic/percussive source separation (HPSS) is a two-step procedure that was initially developed for tracking the rhythmical structure of music, and it has been described in more detail previously [18].

In the first step, median filtering is applied to a spectrogram of the audio computed from the STFT, as suggested by FitzGerald [19]. Harmonic sounds are detected as outliers when filtering in the frequency direction, and percussive sounds are detected as outliers when filtering in the time-direction. The harmonic and percussive parts of the spectrogram (H and P) are then used to create soft masks with Wiener filtering. For the harmonic mask (M_H), the ith frequency of the nth frame is

$$M_{H_{i,n}} = \frac{H_{i,n}^2}{H_{i,n}^2 + P_{i,n}^2}. \quad (3)$$

The element-wise (Hadamard) product between the masks and the complex original spectrogram $\hat{S}$ is then computed, resulting in the complex spectrograms $\hat{H}$ and $\hat{P}$ [19]. These spectrograms are then inverted back to the time domain with the inverse STFT, producing a harmonic and a percussive waveform.

In the second step, the percussive waveform is filtered again with a similar procedure, to remove any traces of harmonic content such as note starts in the bass guitar or vibratos from the vocals. Here, the constant-Q transform (CQT) is applied to compute the spectrogram [20]. With the log-frequency resolution of the CQT it becomes possible to remove the harmonic traces, as the frequency resolutions is high enough to discern between, for instance, the bass guitar and the kick drum in the lower frequencies, and low enough in the higher frequencies to accurately detect the higher harmonics of vocal vibrato. We used the same settings for the frequency resolution (60 bins per octave) and the length of the median filter used for filtering across the frequency direction (40 bins) as proposed in the earlier study [18]. These settings were established as preferable for filtering out harmonic traces. A clean percussive waveform is finally achieved by applying the inverse CQT (ICQT) to the filtered complex spectrogram.

After applying source separation, we estimated how much percussion each audio file contained, the *percussive level* (L_{perc}), by computing the RMS of both the clean percussive waveform (CP) and the original waveform (O), and then computing the difference in signal level

$$L_{perc} = 20 \log_{10} \frac{CP_{RMS}}{O_{RMS}}. \quad (4)$$

For tracks where L_{perc} is higher (closer to 0 dB), the audio contains more percussion, whereas a lower value corresponds to less percussion. In Figure 1 we plot a histogram of L_{perc}-values for the dataset. Values range from around -30 dB to -5 dB, with a mean of about -15 dB.

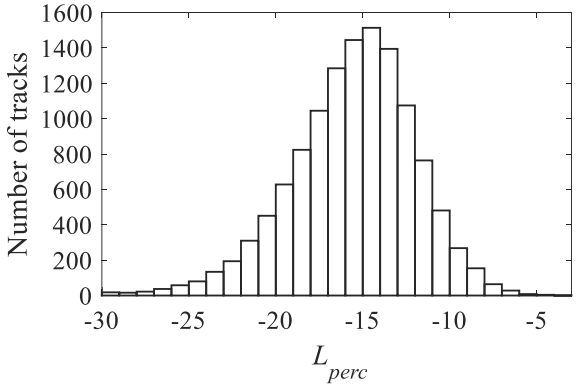

Figure 1. A histogram of L_{perc}-values for the dataset. Edge bins include all tracks beyond the edges.

3 Analysis of LTAS in the Dataset

3.1 Mean LTAS and variations in LTAS

To what extent does LTAS vary in the dataset? To explore this, we computed both the mean LTAS as well as the standard deviation and order statistics across all songs of the whole dataset from *M* (the matrix previously computed in Section 2.2). Figures 2-3 show mean and standard deviation of the LTAS, with and without one-sixth octave smoothing.

The raggedness in the frequency response stems from the partials of harmonic instruments. These congregate at certain frequencies, partly due to certain keys and tones being more commonly used than others in the music. By smoothing the LTAS, the influence of key/tonality is removed. Note that in Figure 2 of mean LTAS, the sound level increases up to about 100 Hz, from where the spectrum exhibits a slope which increases in steepness for higher frequencies.

Figure 3 shows the standard deviation of the dataset, which is the highest in the low and high frequencies. It is the lowest in low-mid frequencies (200-1000 Hz) and also rather low in the high-mid frequencies (1-4 kHz). These frequencies (200 Hz – 4 kHz) are generally used as the range of telephony, since they are the most important frequencies for speech intelligibility [21]. The fundamental frequencies of most instruments that perform the lead melody or provide the chord accompaniment belong to the low-mid frequencies, and these instruments (including the voice) will

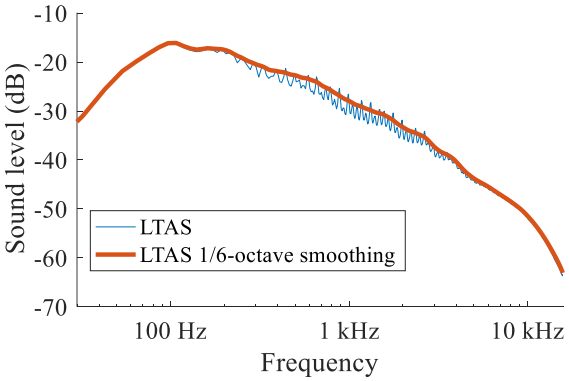

Figure 2. The mean LTAS (blue) and the smoothed mean LTAS (red) of the whole dataset. The smoothed LTAS was computed for each track of the dataset was with a 1/6-octave Gaussian filter.

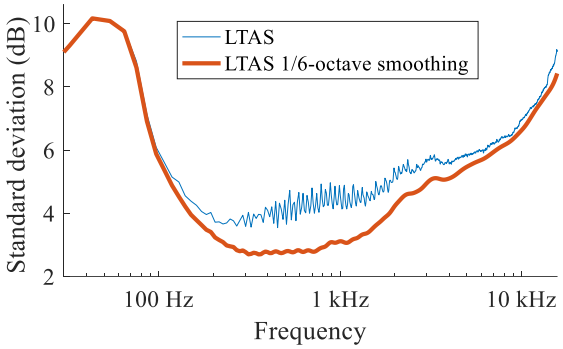

Figure 3. The standard deviation for the LTAS of the whole dataset (blue), and the standard deviation for the LTAS of the smoothed tracks (1/6-octave Gaussian filter) of the whole dataset (red).

usually have harmonics in the high-mid frequencies. A simple interpretation is thus that these frequencies contain the most important information in popular music, just as they do in speech, while the presence of instruments that cover the low- and high-end of the spectrum is not as critical. We observe that smoothing reduces the standard deviation the most in the mid frequencies where the influence of harmonic partials is the strongest. One implication is that comparisons between songs will be much more relevant in the mid frequencies if the LTAS is smoothed.

To study variations across frequency more closely, we computed order statistics for each frequency bin of *M*. This was done by sorting each column of *M* (across songs) and computing percentiles. The LTAS between the 3rd and the 97th percentile for each frequency bin is shown in Figure 4.

The larger variance at low frequencies (up to 200 Hz) that was shown in Figure 3 is also visible in Figure 4 of order statistics. It is interesting to note that the higher variance in these frequencies are to a larger extent due to a lower sound level in the percentiles below 40. The reason is probably that many songs lack instruments that have substantial energy in the low frequencies. One implication for automatic equalization is that when there is a big difference between the mean LTAS of the dataset and the LTAS of a specific track in the bass, deviations below the mean are generally less alarming. For these frequencies, deviations of up to around 10 dB are not uncommon.

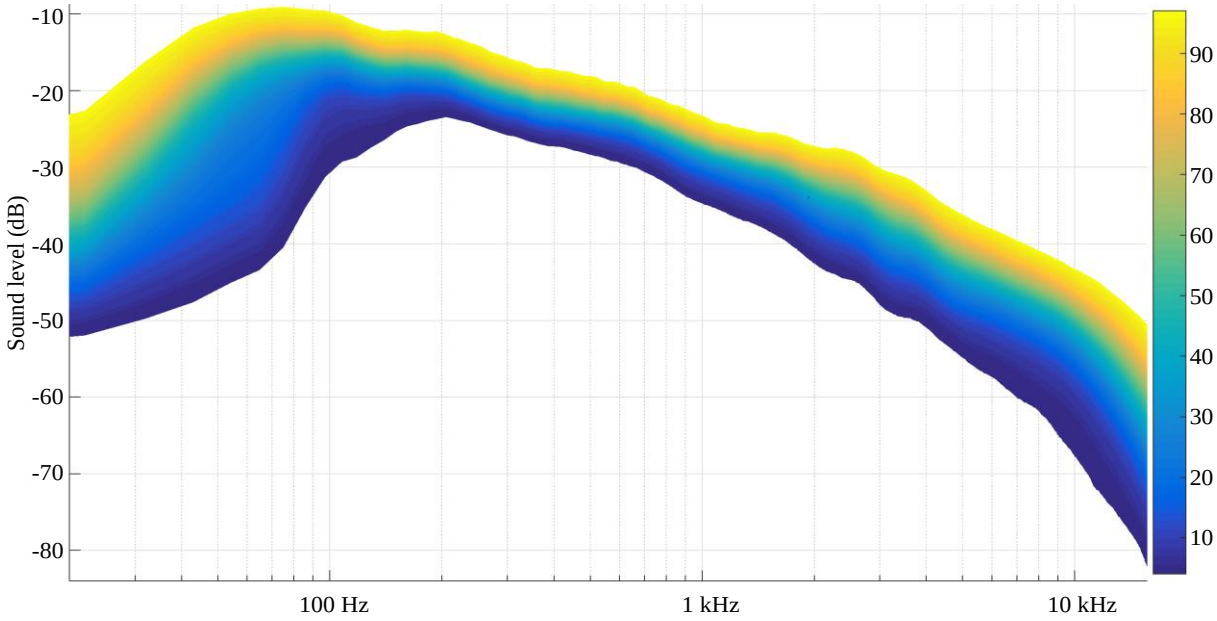

Figure 4. The variation in sound level between different songs, shown as percentiles of the LTAS of all the tracks in the dataset.

3.2 An equation for LTAS in popular music

To be able to describe the mean LTAS of popular music with a compact representation, an equation was fitted to the smoothed mean LTAS of the dataset. If the fitting is done directly on the LTAS-vector, the bass frequencies will be represented by only a few bins, and the fitting will as a consequence over-emphasize the importance of the high frequencies. We therefore converted the vector to log-frequency by sampling the mean LTAS from logarithmically spaced frequency bins in the range of 30 Hz – 15.7 kHz, using 60 bins per octave. The logarithmically spaced frequency vector x had 543 bins. Noting that the mean LTAS in Figure 2 rises until bin x = 100 (94 Hz), and then falls, we decided to do one quadratic fitting for the bass frequencies (x = 1-100) and one quadratic fitting for the rest of the spectra (x = 100-543). The two fittings were then adjusted to intersect (to have identical sound level) at bin 100. The result is shown in Figure 5.

The quadratic fitting in the bass frequencies was

$$y_1 = p_1 x^2 + p_2 x + p_3,$$
$$p_1 = 0.000907,\ p_2 = 0.256,\ p_3 = -32.942, \quad (5)$$

and the quadratic fitting for the mid and high frequencies was

$$y_2 = p_1 x^2 + p_2 x + p_3,$$
$$p_1 = -0.000183,\ p_2 = 0.0213,\ p_3 = -16.735. \quad (6)$$

A linear fitting was also done for the two frequency ranges, primarily to calculate the spectrum slope of

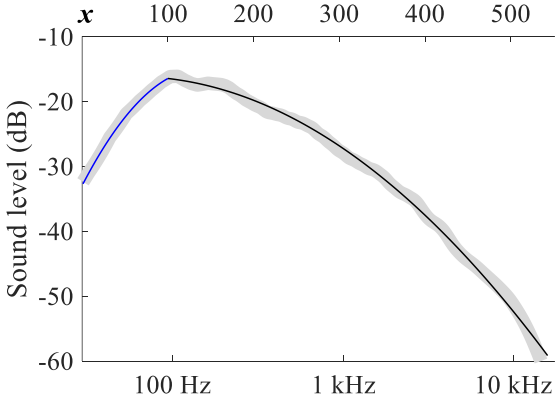

Figure 5. Two quadratic fittings (blue and black) overlaying the mean LTAS (grey).

the high frequencies, and the result was a slope of 5.79 dB/octave from 94 Hz to 15.7 kHz. It is however evident from Figure 5 that the quadratic fittings are a better fit for the mean LTAS curve. It reduced the norm of residuals in y_2 by a factor of 4.1 in relation to the linear fitting. Computing the derivative of Eq. 6 produces an equation from which the spectrum slope can be extracted directly

$$y'_2 = 2\,p_1 x + p_2,$$
$$p_1 = -0.000183, \quad p_2 = 0.0213. \qquad (7)$$

The spectrum slope at different octaves is shown in Table 1.

Center freq.	*x (log bin)*	Slope - dB/Oct
200 Hz	165.22	-2.350
400 Hz	225.22	-3.668
800 Hz	285.22	-4.985
1.6 kHz	345.22	-6.303
3.2 kHz	405.22	-7.621
6.4 kHz	465.22	-8.938

Table 1. The slope of the mean LTAS of popular music expressed in dB/octave at different center frequencies.

The slope steepens with frequency, with more negative slopes toward the higher frequencies. The results when the center frequency is 800 Hz are in line with the slope of 5 dB/octave observed by Pestana et al. [8] (in that study for the range of 100 Hz to 4 kHz).

4 Relationship Between LTAS and the Percussiveness of Music Audio

4.1 Variations in frequency response

As noted in the introduction, the mean LTAS of different genres varies. For example, hip-hop and electronic music have a higher sound level at both low- and high frequencies than jazz and folk music, which are louder (relatively) in the mid frequencies [8]. In the Introduction, our hypothesis was that these differences are not directly related to genre *per se*, but rather reflect the relative loudness of the percussive instruments in the musical mixture. In this Section, the relationship between the computed L_{perc} from Section 2.3 and the LTAS computed in Section 3 is explored.

First, we will show how the L_{perc} relates to LTAS. To visualize the relationship, the smoothed LTAS vectors of the tracks were sorted based on their corresponding L_{perc} to form the 12345×2049 matrix M_S. Then, M_S was divided into 11 groups so that each group consisted of about 1122 tracks with neighboring L_{perc}-values. The mean LTAS for the tracks of each group was finally computed. The mean L_{perc} of each group is shown in Table 2, and the relationship in LTAS between groups is plotted in Figure 6.

Group	**1**	**2**	**3**	**4**	**5**	**6**
Mean L_{perc}	-23.1	-19.5	-18.0	-16.9	-16.1	-15.3
Group	**7**	**8**	**9**	**10**	**11**	
Mean L_{perc}	-14.5	-13.8	-12.9	-11.9	-9.8	

Table 2. Mean L_{perc} for the 11 groups that we calculate the average LTAS from.

As seen in Figure 6, tracks with higher L_{perc} on average have a higher loudness in the low and high frequencies. The variation in LTAS between songs with a different percussive prominence has a similar characteristic as the variation in LTAS between genres with a different percussive prominence found in [8]. To study this relationship closer, the LTAS-curves were normalized relative to the sound level in group 6. This is shown in Figure 7. As seen in the Figure, the relationship seems to be consistent across all groups, i.e. a group with a higher L_{perc} than another group also always had a higher energy in the bass and high frequencies.

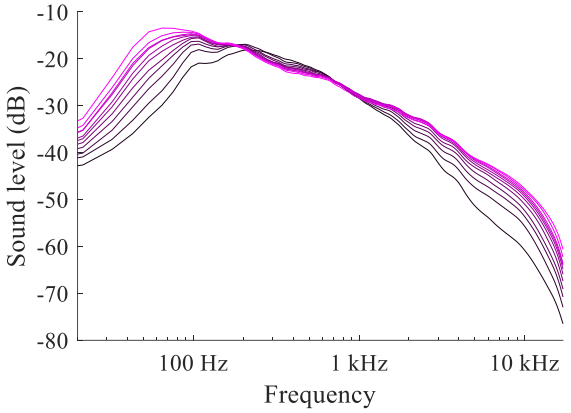

Figure 6. The mean LTAS of the 11 groups, consisting of tracks with neighboring L_{perc}-values (see Table 2). Brighter (magenta) means more percussion.

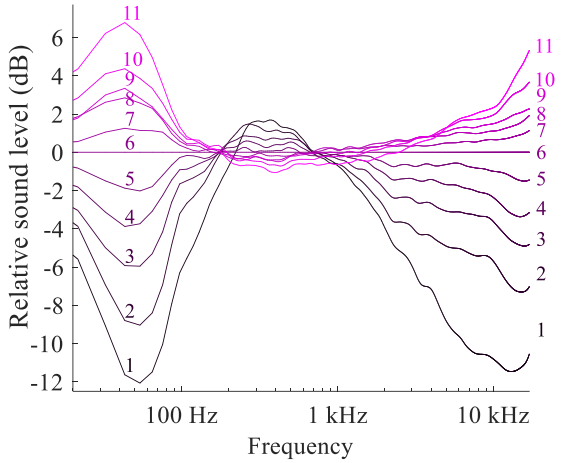

Figure 7. The variation in mean LTAS for the 11 groups of tracks with similar L_{perc}-values . The mean LTAS-curves were normalized relative to the mean LTAS of group 6.

4.2 An L_{perc}-invariant range for measurements of spectrum slope

Observing the LTAS-curves in Figures 6-7, it seems like there exists a frequency in the bass for which the interrelationship between LTAS-levels of the different groups is the same as in a frequency in the treble. This would then translate to an identical spectrum slope for the groups, or in other words, a range for which the spectrum slope is independent of percussive prominence. Such frequency range-estimates could be useful for establishing if a track conforms to a desirable spectrum slope, without having to consider the amount of percussion in the track. We therefore tried to compute the optimal range in the frequency spectrum of the groups to get as similar a slope as possible.

The spectrum slope of the mean LTAS of the 11 groups from Section 4.1 was computed, while iterating over different frequency pairs x_1 and x_2, where x_1 was set to the range of 60-240 Hz and x_2 was set to the range of 1.25-10 kHz. To get a higher resolution for x_1, the log-frequency LTAS-curves described in Section 3.2 were used for both x_1 and x_2. In the analysis, each frequency pair generated a vector of 11 slopes (dB/octave), and we computed the standard deviation of this vector. The x_1 and x_2-pair that minimized the standard deviation was then selected. The

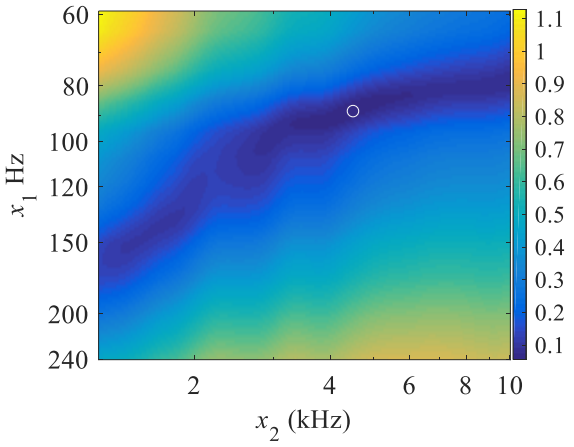

Figure 8. The variation in standard deviation between the spectrum slopes of the 11 L_{perc}-groups for the frequencies x_1 and x_2. The standard deviation is minimized at x_1 = 89 Hz and x_2 = 4.5 kHz (white circle in the plot).

standard deviation of the different frequency pairs is shown in Figure 8. The minimum standard deviation (0.055) was found for x_1 = 89 Hz and x_2 = 4.5 kHz, with a spectrum slope of 4.53 dB/octave.

The implication of the analysis is that although the LTAS of popular music varies significantly for tracks with different percussive prominence, the average spectrum slope between around 90 Hz and 4.5 kHz is rather independent of the amount of percussion.

5 Applications - Automatic Equalization

The findings of this study can be utilized to perform automatic equalization on e.g. previously mixed audio recordings. As outlined in Section 1.4, the dataset can be used as a collection of reference tracks that guides the equalization, similarly to how an audio engineer uses reference tracks in the equalization process. The reference tracks define a *target* spectrum, and the mix is altered to better coincide with the LTAS of the target. This type of frequency matching has been described in e.g. [11]. The characteristics of the reference tracks are however important; ideally, the tracks should be well balanced and have a similar instrumentation as the processed track. A clear relationship between LTAS and L_{perc} was shown in Section 4. In this Section we will first use the smoothed mean LTAS as a target in automatic equalization, and

then utilize the computed L_{perc} of the dataset to automatically refine the target individually for each track.

The smoothed mean LTAS, defined as the target T, was used to compute the error T_e for each frequency bin b across all N songs x of the dataset as

$$T_{eb} = \sqrt{\frac{1}{N-1}} \sum_{x=1}^{N} (\mathrm{LTAS}_{xb} - T_b)^2. \quad (8)$$

In this equation, $\mathrm{LTAS}_{xb} - T_b$ is the difference in sound level for a frequency bin between a track x and the mean of all tracks, the target T. The vector T_e is thus identical to the standard deviation of the LTAS for the dataset, previously shown in Figure 3. The error T_e increases when T diverges from the LTAS of a track. Implicitly, each track is therefore assumed to have been previously equalized to the optimal LTAS for the mix, a sort of "ground truth".

The mean error across all frequencies $\overline{T_e}$, was 3.94 dB on average for the dataset. How can the findings for the relationship between L_{perc} and LTAS from Section 4 be used to improve T, and thus reduce T_e? In other words, to what extent can we find a better target LTAS for each song by incorporating information about the amount of percussion in the music? This was investigated by using L_{perc} to compute the adjusted target LTAS T', and then track the reduction in error. For each track x, a target T_x was derived by computing the weighted mean LTAS of all tracks with a similar L_{perc}. The weight w_j that defined the contribution to the target LTAS from each track j was computed from the absolute distance in L_{perc} as

$$w_j = 1 - \frac{|L_{perc_j} - L_{perc_x}|}{L_{perc}^{lim}}. \quad (9)$$

The constant L_{perc}^{lim} that defined the limit at which a track will contribute with a non-zero weight, was set to 1.5 dB. For $L_{perc}^{lim} < 1.5$, the targets depended too much on random variations in the songs. For $L_{perc}^{lim} > 1.5$, the algorithm could not to the same extent pick up local variations in LTAS due to L_{perc}. If there were less than 200 tracks with a weight w_j above 0, L_{perc}^{lim} was set to the absolute difference of the 200th closest weight before the computations of Eq. 9 was repeated. This ensured that enough examples were used for the tracks with a larger spread of L_{perc}.

We used a leave-one-out cross-validation by setting the weight of track x to 0. Furthermore, negative values for w_j were set to 0. The new target T_x' for each track was given as the weighted mean

$$T_x' = \frac{\Sigma_j^{12345} w_j \times \mathrm{LTAS}_j}{\Sigma_j^{12345} w_j}. \quad (10)$$

The new error T_e' was computed as in in Eq. 8, but with T_b replaced by T_{xb}'. The resulting reduction of the mean error $\overline{T_e} - \overline{T_e'}$ became 0.61 dB. The reduction $T_e - T_e'$ across all frequencies is shown in Figure 9.

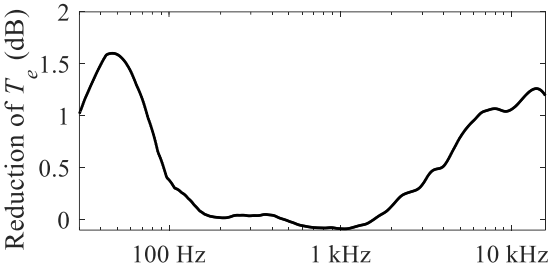

Figure 9. The reduction of the error (T_e) when the target T' was based only on tracks with a similar L_{perc}. Deriving a target from these tracks reduced the error for the low frequencies (below 100 Hz) and for frequencies above 2 kHz.

Figure 9 shows that the error is reduced for frequencies below around 100 Hz and frequencies above around 2 kHz. In the mid frequencies, it is almost as accurate to use T as T' for the target. When comparing the errors across tracks instead of across frequency, we found, unsurprisingly, that the reduction in error was the greatest for the tracks with an L_{perc} that differed the most from the mean L_{perc}. We also conclude that tracks with a high L_{perc} generally had a lower error than tracks with a low L_{perc}.

6 Discussion

6.1 Analysis of the computed LTAS and its applications

The smoothed mean LTAS is interesting to study in more detail. One reason for smoothing with one-sixth octave bands instead of the more common one-third octave bands is the potential to detect finer-grained variation in the mean LTAS curves. This narrower bandwidth should still give reliable results because of

the large size of the analyzed dataset. When looking closely at the curve in Figure 2, it is evident that the spectrum falls relatively sharply at around 4.5 kHz. This could be due to a decrease in sound level of the voice spectrum, caused by a pair of antiresonances from cavities in the piriform fossa [21, 22]. Another possible factor is that electric guitars without heavy distortion tend to fall in energy relatively sharply around this frequency. These guitars are common in popular music. The small dip at around 150 Hz in Figure 2 may also be explained by the voice spectrum in combination with the spectrum of other instruments. This region is slightly below where most energy of the vocals in popular music recede. Furthermore, it is slightly above where most of the energy of the bass and the kick drum is located. The absence of vocal energy may also be the reason for a sharp increase in variance below 150 Hz in Figures 3-4.

By computing the percentiles of the LTAS of a large dataset of popular music, we have provided a comprehensive visualization of spectral distribution for the genre. This kind of representation can be very useful for numerous tasks involving (primarily) the manipulation of the frequency spectrum during mixing or mastering. For example, when the spectrum of a track contains outliers in relation to the percentile matrix, it would be wise to focus attention to the deviating frequencies. In this case, it is important to analyze the reason for the deviations. Can the spectral outliers be motivated by the musical arrangement, e.g. are the outliers unavoidable to be able to maintain a desirable spectrum for the instruments in the mix? If not, the outliers may be alleviated by adjusting the frequency spectra of the most relevant instruments in the mix, or by adjusting the equalization curve of the master track directly. In the latter case, the spectrum should just be adjusted partly towards the target, and only so for the frequencies that deviates significantly. If the outliers however are unavoidable given the arrangement, attention may instead be directed to the instrumentation; is it possible to add or remove any instruments in the arrangement? An extension would be to perform the analysis at different levels of smoothings of the LTAS. This would make it possible to identify different types of discrepancies, such as energy build-ups in narrow bands, or a general imbalance between high and low frequencies.

It is important to note that the ideas above are not meant to be limited to manual adjustment by e.g. the producer or mastering engineer. Instead, they should be regarded also as suggestions to establish new and more elaborate algorithms in automatic music production. The idea of a balanced frequency response is one that pertain to many areas of music production. The success of LTAS-targets depends on an understanding of how the arrangement of a song calls for a specific target. By using the L_{perc} of the tracks, this was accounted for by relating percussive prominence to LTAS. Further progress could be made in this area by connecting additional features of the music to LTAS. This would allow an algorithm (or a human) to make more informed decisions with regard to equalization and/or arrangement. A key point though, is that the extracted features can only be linked to the LTAS of a track indirectly. If a feature is extracted which is directly related to LTAS, such as e.g. the spectrum slope, we would not, in any meaningful way, be able to distinguish the tracks that have a deviant but desirable LTAS from the tracks that have a deviant LTAS that should be adjusted. Reductions in T_e would in this case not imply an improved equalization. However, for this study, HPSS is not directly related to LTAS. Instead, the improvement in target LTAS stems from the fact that percussive instruments have a tendency to cover a broader frequency range than do harmonic instruments. The validity of using the L_{perc} will depend on the ability of the presented filtering technique (Section 2.3) to separate harmonic and percussive instruments. Here we conclude that the first stage of the filtering is a well-established method for audio source separation, and that the second stage has been used successfully for varying tasks such as beat tracking [23] and the modeling of the perception of speed in music audio [24]. There may however be a slight propensity for the algorithm to overstate the amount of percussion for high frequencies, as harmonic horizontal ridges in the spectrogram are harder to identify for these frequencies.

6.2 Relationship between LTAS and L_{perc}

In Section 5 it was noted that the songs with the lowest L_{perc} generally had a higher error in the computed targets than songs with a high L_{perc}. The implication is that it is harder to establish a target LTAS for songs

without much percussion. This is perhaps not surprising, as sounds of non-percussive instrument will appear as horizontal ridges in a spectrogram, which translates to narrow peaks in the LTAS if they occur for longer periods of time in the track. One important factor related to this is the width of the Gaussian smoothing of the LTAS. If a broader smoothing had been used, the errors in the targets for tracks with a low L_{perc} would have been significantly reduced (the curve of mean LTAS of the dataset would however be less precise). As shown in Figure 3, the smoothing affects the standard deviation in the mid frequencies the most. This is due to a smoothening of the narrow peaks from non-percussive instruments.

The wide variation in LTAS of the 11 L_{perc}-groups is rather remarkable. To put the variation into perspective, it can be compared with the differences in LTAS between genres from the earlier study by Pestana et al. [8]. When accounting for an apparent lack of loudness normalization of the tracks in that study, it is evident that the amount of percussion is a stronger factor for variation in LTAS than genre, even for such varying genres as jazz and hip-hop (these were the two genres that varied the most in that dataset). In the Introduction we formulated a hypothesis, that LTAS is not directly related to genre; rather, it is an effect of variations in percussive prominence between genres. The analysis in section 4.1 validates the hypothesis, since a strong relationship between LTAS and L_{perc} was found, and since that relationship had the same characteristics as that between genres that are known to vary in the amount of percussion. Our conclusion then is that *the variation in LTAS between genres is not directly related to inherent stylistic variations. Instead it is primarily a side-effect of variations in the amount of percussion between genres.* Thus, for a dataset of hip-hop songs with very little percussion (to the extent that it still belongs to the genre) and a dataset of folk songs with a lot of percussion, a reversed relationship concerning LTAS would likely be observed.

6.3 Simple fittings of mean LTAS

We have shown that the spectrum slope of popular music can be accurately approximated with a simple fitting. Generally, spectrum slope has previously been referred to as a linear function. But as shown, for a higher accuracy, a quadratic term should be introduced to specify the increasing steepness of the slope with increasing frequency. The effect is rather significant; at 800 Hz the spectrum slope is around 5 dB/octave, but at 3.2 kHz the slope is 7.6 dB/octave. Furthermore, in Section 4.2, we have calculated a frequency range where the spectrum slope does not depend on the amount of percussion in the tracks. It was found that between 89 Hz and 4.5 kHz, the spectrum slope was close to 4.53 dB/octave for all 11 L_{perc}-groups. For future studies, the idea could be generalized to calculate a frequency range that gives the lowest standard deviation in spectrum slope without trying to account for any specific feature.

7 Acknowledgement

We would like to thank Sten Ternström for proofreading the manuscript before the peer-review process.

References

[1] L. J. Sivian, H. K. Dunn, & S. D. White, "Absolute amplitudes and spectra of certain musical instruments and orchestras," *The Journal of the Acoustical Society of America*, vol. 2, no. 3, pp. 330-371, (1931).

[2] A. Elowsson, & A. Friberg, "Predicting the perception of performed dynamics in music audio with ensemble learning," *The Journal of the Acoustical Society of America*, vol. 141, no. 3, pp. 2224-2242, (2017).

[3] B. Owsinski, *The Mixing Engineer's Handbook,* Vallejo, CA: Mix Books, first ed., (1999).

[4] B. B. Bauer, "Octave-band spectral distribution of recorded music," *Journal of the Audio Engineering Society*, vol. 18, no. 2, pp. 165-172, (1970).

[5] E. V. Jansson, & J. Sundberg, "Long-time-average-spectra applied to analysis of music. Part I: Method and general applications," *Acta Acustica united with Acustica*, vol. 34, no. 1, pp. 15-19, (1975).

[6] E. Benjamin, "Characteristics of musical signals," *In Audio Engineering Society Convention 97*, (1994).

[7] K. Randall, “BBC RD 1981/2: An Investigation into the Spectral Content of Pop Music," *tech. rep.*, BBC, (1981).

[8] P. D. Pestana, Z. Ma, J. D. Reiss, A. Barbosa, & D. A. Black, “Spectral characteristics of popular commercial recordings 1950-2010,” *In Audio Engineering Society Convention 135*, (2013).

[9] E. Perez-Gonzalez, & J. Reiss, "Automatic equalization of multichannel audio using cross-adaptive methods," *Audio Engineering Society Convention 127,* (2009).

[10] D. Reed, “A perceptual assistant to do sound equalization,” *In Proceedings of the 5th international conference on Intelligent user interfaces,* pp. 212-218, ACM, (2000).

[11] Z. Ma, J. D. Reiss, & D. A. A. Black, "Implementation of an intelligent equalization tool using Yule-Walker for music mixing and mastering" *Audio Engineering Society Convention 134*, (2013).

[12] J. D. Reiss, “Intelligent systems for mixing multichannel audio,” *In 17th International Conference on Digital Signal Processing (DSP).* pp. 1-6, IEEE, (2011).

[13] D. Z. Borch, & J. Sundberg, “Spectral distribution of solo voice and accompaniment in pop music,” *Logopedics Phoniatrics Vocology*, vol. 27, no. 1, pp. 37-41, (2002)

[14] B. Katz, *Mastering Audio, the Art and the Science,* Oxford: Focal Press, (2002).

[15] R. Izhaki, *Mixing audio: concepts, practices and tool*, Taylor & Francis, (2012).

[16] C. Hummersone, “ioSR Matlab Toolbox,” *Computer software*, Available from https://github.com/iosr-surrey/matlabtoolbox, (2016).

[17] ITU, “ITU-R BS.1770-4: “Algorithms to measure audio programme loudness and true-peak audio level; BS Series," *Tech. rep*., Int. Telecommunications Union, (2015).

[18] A. Elowsson, & A. Friberg, ”Modeling the perception of tempo,” *Journal of the Acoustic Society of America*, vol. 137, no. 6, pp. 3163-3177, (2015).

[19] D. FitzGerald, “Harmonic/Percussive Separation Using Median Filtering,” *Proc. of DAFx*, 4 pages, (2010).

[20] C. Schörkhuber, & A. Klapuri, “Constant-Q transform toolbox for music processing,” *Proc. of the 7th SMC*, pp. 322-330, (2010).

[21] S. O. Ternström, “Hi-Fi voice: observations on the distribution of energy in the singing voice spectrum above 5 kHz,” *Journal of the Acoustical Society of America*, vol. 123, no. 5, pp. 3379-3379, (2008).

[22] J. Dang, & K. Honda, “Acoustic characteristics of the piriform fossa in models and humans,” *The Journal of the Acoustical Society of America*, vol. 101, no. 1, pp. 456-465, (1997).

[23] A. Elowsson, "Beat Tracking with a Cepstroid Invariant Neural Network," *in 17th International Society for Music Information Retrieval Conference*, pp. 351-357, (2016).

[24] A. Elowsson, A. Friberg, G. Madison, & J. Paulin, “Modelling the speed of music using features from harmonic/percussive separated audio,” *Proc. of ISMIR*, pp. 481-486, (2013).