Human-Made Rock Mixes Feature Tight Relations Between Spectrum and Loudness

EMMANUEL DERUTY, FRANÇOIS PACHET, AND PIERRE ROY
(emmanuel.deruty@gmail.com)

Sony Computer Science Laboratory, Paris, France

The tremendous success of rock music in the second half of the 20th century has boosted the sophistication of production and mixing techniques for this music genre. However, there is no unified theory of mixing from the viewpoint of sound engineering. In this paper, we highlight relationships between loudness and spectrum in individual tracks, established during the process of mixing. To do so, we introduce an ad hoc, three-dimensional model of the spectrum of a track. These dimensions are derived from an optimal monitoring level, that is, the level that optimizes the number of frequency bands at the same, maximum loudness. We study a corpus of 55 rock multi-tracks and correlate the model with the loudness of the tracks. We suggest that (1) at high monitoring levels and/or on high-end monitors, track loudness is a linear function of its spectral centroid, and (2) at low monitoring levels and/or on budget monitors, a track’s optimal monitoring level is a linear function of its loudness. This indicates that under good listening conditions, human mixers tend to focus on spectral balance, whereas under bad conditions, they favor individual track comprehension. We discuss the implication of our results for automatic mixing.

0 INTRODUCTION

Mixing is a crucial step in popular music production. However, the human mixing process, viewed from a data flow perspective, is still poorly understood. Mixing is mostly considered a craftsmanship rather than a science, on the grounds that it “is ‘highly nonlinear’ [1] and ‘unpredictable’ [2], and that there are ‘no hard and fast rules to follow’ [1]” [3]. In this paper, we contribute to the understanding of the human mixing process by exhibiting invariant relations in tracks produced by human mixers. Whether these relations are produced consciously or not lies out of the scope of this study. Our primary goal is to identify these relations from the analysis of a corpus in the mainstream rock genre.

A fundamental concern regarding the mixing process is the extent to which the listener can hear each track making up a mix individually. Indeed, several automatic mixing frameworks are based on the sole hypothesis according to which each track in the mix should be as audible as possible [4]–[9]. Of crucial importance in regard to track audibility is the gain applied to each track during the mixing process [10]. However, relevant literature shows that no consensus is reached as to the settings of the tracks’ relative gains. Gains may be “subjective” and “influenced by taste” [4], they may result in equal track loudness [5],[6],[9], or they may favor soloing instruments [5],[11]. In this paper, we examine the possibility of individual track gain being set by human mixers so that track audibility is optimal, in the sense that the number of frequency bands that can be heard is optimized for each track.

Another fundamental concern regarding the mixing process concerns the spectral balance of the result [12]–[14]. Many mixing engineers mix towards a subconscious target frequency response curve, which may be approximated by the average spectrum of a large commercial recording dataset [12],[13],[15]. This leads us to also examine the possibility of individual track gain being set by human mixers so that the spectral balance of the mix approaches a typical spectral envelope. Under this point of view, the overall spectral balance of a mix would be the result of both track gains and individual track equalizations.

Following these concerns, we propose a signal descriptor that provides an approximation of track audibility. We then compare the descriptor’s values with individual track loudness. We find that there exists a significant correlation between individual track audibility and loudness, which can be observed at low monitoring level and/or on budget monitors. We also consider a variant of the spectral centroid that derives from the audibility descriptor, which we again compare with individual track loudness. Using this variant, we exhibit a significant correlation between individual
track brightness and loudness, which can be observed at high monitoring levels and/or on high-end monitors, and which points to a specific overall spectral profile.

These findings suggest that the mixing engineers’ work is guided by two implicit directives. Under bad listening conditions, track audibility stands out as a priority. Under good listening conditions, priority shifts to fitting a specific spectral profile. The success of a mix may lie in the compliance to these two directives.

Since the results of both directives are observed as correlations between spectral and loudness-based descriptors, they indicate the existence of tight relations between spectrum and loudness in commercial rock mixes. Such relations may be of interest in the field of automatic mixing, by providing a basis for automatic track gain adjustment based solely on the track’s spectrum.

1 MATERIAL AND METHODS

1.1 Corpus

Following [16], the music corpus we rely on consists of 55 multi-track songs from the Rock Band video game1. Song selection was the result of a compromise between the following constraints:

- The corpus should focus on the commercial rock genre.
- Instrumentation should focus on the standard drums/bass/guitars/vocals setup.
- The drum section should be available as separate tracks and not bounced into a stereo mix.
- To ensure representativeness, not more than three songs from the same band should be chosen.
- The release date for tracks should be as evenly split as possible.
- Lead guitar parts should be as numerous as possible, even though all songs do not feature such parts.

Each song from the game is typically split into seven classes of tracks. “kick drum,” “snare drum,” “overheads,” “bass guitar,” “guitar,” “vocals,” and “miscellaneous.” The “overheads” are a pair of microphones placed above the drum kit. They capture a global image of the instrument, with an emphasis on cymbals [17].

All tracks from the corpus are produced. They are not raw audio tracks captured directly from the inputs to a mixing console. We can hear obvious equalization, compression, chorusing, reverberation, and distortion. Summing the track results in a mix that’s close the commercial version of the songs. According to the mixing engineer who adapted the songs to the Rock Band game, “if everyone’s playing the game perfectly, [the song] sounds just like the record” [18].

A manual check on all songs generated very few exceptions, with 5 tracks being either much louder or much softer than they appear in the commercial mix. The levels on these tracks were manually corrected. Tracks from the resulting corpus can simply be summed in order to get a very good approximation of the final song, even to professional mixing engineers’ standards.

The original guitar tracks often contained a sequential mix of different guitar parts. We manually split such tracks so that one audio file corresponds to a single instrumental part. As illustrated in Fig. 1, this results in a number of instrumental parts that may be different from the number of songs.

The class “miscellaneous” designates audio files that contain keyboard, backing vocals, additional sound effects or extra guitar parts. Selection of the songs ensures that keyboard, backing vocals, and sound effects are minimal and can therefore be considered as insignificant. Guitar parts from the “miscellaneous” tracks were manually extracted and treated as guitar tracks. Finally, we distinguish between lead guitars (solos) and nonlead guitars (accompanyment/rhythmic guitars). As a result, there are seven classes of tracks: “kick drum,” “snare drum,” “overheads,” “bass guitar,” “nonlead guitar,” “lead guitar,” and “vocals.”

The corpus does not contain the original songs, but premixes in which all the tracks pertaining to each instrument type have been mixed together. Therefore, conclusions reached during the article may only apply to premixes, not to individual tracks. Fig. 1 summarizes the final corpus content. 392 tracks were extracted from 55 unique songs. During the mixing stage, “comparative checks against stylistically similar releases [should be performed]” [19], which indicates that principles involved in mixing may be specific to a music genre. Observations made in the present paper apply to the mainstream rock genre, even though the method involved may be applicable to other music genres.

1.2 Loudness descriptors

Loudness as a subjective measure is a widely studied field [20]–[24]. The basis for most loudness algorithms lies on the frequency filtering of the signal by the ear, which is

---

1http://www.rockband.com/
referred to as frequency weighting. Fig. 2 shows a number of frequency weighting contours (FWCs) that are routinely used in loudness evaluation [23]. One particular class of loudness models we will rely on, the measure of the equivalent continuous sound level, or $L_{eq}$, evaluates loudness by first filtering the signal with a FWC, and then by computing the RMS of the result. In case of stereo tracks, following [15], power summation was used.

The contours shown in Fig. 2 are not level-dependent, which makes them unusable for our purpose. In this paper, we use $L_{eq}$ measures based on level-dependent frequency weighting contours drawn from the Fletcher–Munson [25] and ISO226:2003 [26] standards, the latter being detailed in Fig. 3. While there exists many such standards, we select ISO226:2003 on the grounds that it is an up-to-date international norm, and Fletcher–Munson because it is a reference to which more modern standards are compared [27]. Mixing on headphones is often considered as very different to mixing on loudspeakers [28]. Reference to both Fletcher–Munson and ISO226:2003, respectively obtained using headphones and loudspeakers [25]–[26], may address some differences between headphone and loudspeaker mixing. We will write as $L_{eq}$ (Standard-phon value) the corresponding loudness approximation.

As a sanity check, we compare level-dependent models against other loudness models. We focus on three FWCs from each standard: 10 phon (low loudness), 50 phon (medium loudness), and 90 phon (high loudness). This results in six $L_{eq}$-based loudness models, which we compare with each other and with other loudness models by evaluating the linear correlations between loudness measures made on our corpus using each model. The other models are $L_{eq}(A)$, $L_{eq}(C)$ [23], EBU3341 [29], Zwicker and Fastl’s model for nonstationary sounds [30], and Glasberg & Moore’s models for nonstationary sounds [31]. Results are shown in Fig. 4. Similar-level measures from different standards are generally better correlated to each other than they are to other loudness models, thus indicating a consensus between the two standards at similar monitoring levels.

1.3 Audibility and brightness descriptors

In pop and rock music, equalization is omnipresent [32] and is frequently applied liberally [15]. The two major corrective purposes of equalization are “the unmasking of sound sources” and “the avoidance of spurious resonances” [15].

Masking is a phenomenon that has been deemed as undesirable [5],[9], and mixing has been shown to generally lower the amount of masking between tracks [15]. However, in the context of our corpus, and as shown in Fig. 5, all instruments are liable to mask other instruments at particular frequencies. Equalization may be used to minimize masking, but only in favor of a specific instrument at a particular frequency - a point of view shared by [32]. Each track is assigned a frequency region in which it is allowed to mask the other tracks.
perceived. The greater the weighted bandwidth, the more elements from the track can be heard.

As for resonance attenuation, it is a classic equalization technique that can be used to lessen the individuality of each instrument in favor of a better blend [33]. With less resonances, the resulting spectral envelope is flatter than the original.

As far as the corpus is concerned, equalization therefore shapes a track in two aspects: creation of a privileged frequency region and flattening of the spectrum. We proceed to design spectrum-related descriptors that account for both aspects.

As shown in Fig. 6, if we weigh the corpus tracks’ power spectrum with FWCs, thus producing an approximation of the tracks’ loudness depending on the frequency, we can generally identify three zones: a central, flatter section, surrounded by two roll-offs. The width of the central zone accounts for the frequencies for which the resonances have been equalized and attenuated. The privileged frequency region as previously illustrated in Fig. 5 is accounted for by the central zone’s center frequency.

We propose three descriptors. The overall width of the part of the perceived spectrum that corresponds to the central zone we call weighted bandwidth. Loudness values from bands inside the central zone are considered equal, and greater than loudness values from the bands outside the central zone. Therefore, masking effects between tracks notwithstanding, the frequencies inside the central zone correspond to the track elements that can be the most easily
There are 91 weighted spectra, and one middle zone for each weighted spectrum. As illustrated in Fig. 9, bottom, there exists an FWC for which the “flat” zone is widest. We select this FWC, and consider, as previously illustrated in Fig. 7, that it can be approximated as a set of three line segments, respectively, corresponding to the low frequency roll-off, the high frequency roll-off, and the middle “flat” zone. The roll-off sections are considered as negligible, for the reason that they correspond to lower loudness values that are more difficult to hear. This leaves us with only one horizontal line segment, which can be entirely described using only two parameters, its center and its width.

The methodology we propose is motivated by a cognitive interpretation: for each track, there exists a monitoring level (FWC level in phon) for which a maximum of spectrum bands are equally loud, and louder than the other bands (largest “flat” zone). The FWC phon value corresponding to the largest “flat” zone is the optimal monitoring level (OML). The middle zone center is the optimal weighted spectral centroid (OWSC). The “flat” zone width is the OWB. The OWSC provides an approximation of the sound’s cognitive brightness at the OML, as does the original spectral centroid [46].

2 RESULTS

2.1 Relation between Loudness and OWSC

We first evaluate the relation between loudness and OWSC. Starting from the Corpus described in Section 1.3, we evaluate the loudness for each track. The FWC level values we consider for the experiment range from 10 to 90 phon, with a 5-by-5 phon increment. Each track is therefore measured using the two different standards with seventeen FWC values for each standard (Fletcher–Munson and ISO226:2003). Simultaneously, we evaluate the OWSC for each track. Since the OWSC is dependent on the standard from which the FWCs that are used for its evaluation are extracted, each track corresponds to two OWSC values.

For each loudness model, each FWC level and each OWSC standard, we evaluate the linear correlation between loudness and OWSC values. To do so, we use two methods. The first method consists of evaluating the correlation based on the 392 couples of values. The second method consists of first grouping each track into its class (kick drum, snare drum, overheads, bass guitar, nonlead guitar, lead guitar, and vocals), calculating the median values for each descriptor inside the class, and then evaluating the correlation based on the resulting 7 couples of values. The results are compiled in Fig. 10. For low FWC levels, the correlation is positive, with OWSC values getting higher as loudness increases. For high FWC levels, the correlation is negative, with, on the contrary, OWSC values getting lower as loudness increases.

We expand the experiment towards a larger variety of models, using smoothed versions of FWCs and allowing combinations between models. Fig. 11 illustrates the result, by illustrating OWSC and loudness based on $L_{eq}$ models relying on the particular 10, 50, and 90 phon smoothed
FWCs that provide the highest correlation spans and therefore the strongest relations between loudness and OWSC. Correlations based on track classes decrease from 0.56 (p value 0.19) to −0.76 (p value 0.05) when monitoring level increases, while correlations based on all single tracks decrease from 0.48 (p value 0.00) to −0.53 (p value 0.00). Significance is discussed in Section 2.3. Representations of the 25th and 75th percentiles for the distributions show that most songs follow an archetypal OWSC/loudness pattern. To get more information about the consensus, we cluster the 55 songs into 10 clusters. Over all models and standards, a mean of 82% songs are sorted in one single cluster, with the other 18% being evenly distributed between the nine remaining clusters. This indicates a strong consensus towards a typical arrangement, with a number of singular exceptions.

2.2 Relation between Loudness and OML

Using the same protocol as in Section 2.1, we evaluate the relation between loudness and OWB. Results are shown in Fig. 12. For low FWC levels, the correlation is positive,
with OML values getting higher as loudness does. For high FWC levels, the correlation is negative.

Again, we expand the experiment towards a larger variety of models, using smoothed versions of FWCs and allowing combinations between models. Fig. 13 illustrates the highest correlation span found. Use of the 50- and 90-phon FWCs provides no obvious arrangement and should therefore be discarded. Correlation based on track classes is 0.72 (p value 0.07), while correlation based on all single tracks is 0.50 (p value 0.00). Significance is discussed in Section 2.3. Use of the 10-phon FWCs results in a very good alignment, with OML clearly increasing with loudness, the only exception concerning the bass class. Bass class excluded, the correlation based on track classes reaches 0.98 (p value 0.00). Consensus is still important, with 72% of individual songs being sorted into the main cluster.

### 2.3 Significance

We examine the p values found in relation to the experiments conducted in Sections 2.1 and 2.2. As far as correlations evaluated on all tracks are concerned, p values corresponding to high and low FWC levels are always less than 0.01, indicating a significant correlation. p values evaluated on track classes, however, are higher than 0.1, which is to be expected given the low number of classes. That said, the two sets of correlations follow a similar behavior, suggesting that correlations evaluated on track classes are indeed significant despite high p values.

To further confirm the results’ significance, we now perform an additional experiment, in which we look for similar relations in a corpus of badly produced tracks - and fail, which shows that the results obtained in Sections 2.1 and 2.2 are neither random nor trivial.

What is bad production is not easy to characterize. In the studio, highly unorthodox techniques, such as processing snare drums using the engineer’s talkback [47], splitting the output of a low-range drum machine into different guitar amps [48], or extracting the dynamic envelope of a single track to regulate the whole mix [49] are commonplace, even in the context of mainstream music. Since we’re dealing with a rock oriented corpus, we will consider as bad production spectral modifications that are seldom heard in the context of this music genre. To that purpose, we use narrow-band EQs, which we devise so that the processed tracks sound highly unnatural, tinny, and “electronic,” a style of sound that’s not common to rock music.

We process each track in the corpus ten times using a 25-band filter, the gain for each band being a random value between −20 and +20 dB (uniform probability). The resulting tracks sound completely out of place in the context of
mainstream rock music. We then follow the same protocol as in Sections 2.1 and 2.2. Results are shown in Figs. 14 and 15. For comparison purposes, the scale for the vertical axis is the same as in Figs. 10 and 12.

Results drawn from the degraded corpus are clearly different and do not indicate any clear relation between the different descriptors. In particular, correlations evaluated on all tracks are close to zero, and, unlike what was observed in Sections 2.1 and 2.2, the two sets of correlations don’t converge. This indicates that the previously observed relations between loudness, OWSC and OML do not apply to a badly produced song. It confirms that these relations cannot be observed in the context of any ensemble of tracks, and therefore, cannot be considered as trivial.

### 2.4 Interpretation

In Section 2.1, we found significant correlations between loudness and OWSC. Assuming OWSC provides an approximation for cognitive brightness, and taking into account observations made in Section 1.3 pertaining to the monitoring level/monitor range equivalence, the correlations suggest that at lower monitoring levels and/or on consumer monitoring systems, comparatively softer tracks sound darker, and comparatively louder tracks sound brighter. Conversely, at higher monitoring levels and/or on high-range monitoring systems, comparatively softer tracks sound brighter, and comparatively louder tracks sound darker.

In Section 2.2, we found significant correlations between loudness and OML. We focus on the results that concern the use of 10-phon FWCs during loudness evaluation. The unit for the $x$-axis is the generic “LU” (Loudness Unit), which is roughly equivalent to the dB. The unit for the $y$-axis is the phon, which is also roughly equivalent to the dB. The slope of the regression line is 2.5, which means that an increase of 1 LU of loudness results in an increase of 2.5 phon of OML. In other words, the optimal monitoring level increases faster than the actual level. However, in the neighborhood of 30-phon FWCs, the slope is close to 1. The optimal monitoring level increases correspondingly to the actual level. Under the assumptions made in the previous Sections concerning the cognitive and practical meaning of OML, we can conclude that at relatively low monitoring levels and/or on relatively low-range monitors, track spectrum and loudness are conjointly adjusted so that each track, whatever its relative loudness, remains optimally perceived and understood.

### 3 CONCLUSION

Mixing is often considered a mysterious activity. In the like of old-fashioned guild artisans, engineers and producers reputedly learn “tricks of the trade” from the “Greatest Teachers” or “mentors” [50]–[51], masters who share the mysteries of their craft with their disciples. This attitude has complicated the task of researchers who want to rationalize mixing. Potential myths have to be debunked, and disagreement settled [15].

This paper suggests that as far as this particular corpus is concerned, there exist constant underlying trends enforced by human mixers that have not been previously highlighted:

1. At higher monitoring levels and/or on full-range monitors, comparatively brighter tracks are mixed softer, and comparatively darker tracks are mixed louder. Given such monitoring conditions, audio engineers appear to be concerned by the perceived spectral balance.

2. At lower monitoring levels and/or on budget monitors, with the exception of the bass guitar, track spectrum and loudness are set conjointly so that each track is optimally understandable. Given such
monitoring conditions, audio engineers appear to be concerned by the comprehension of each individual track.

Even though the corpus is particular in the sense that it is based on sub-groups rather than on individual tracks, this would imply that under reliable listening conditions, comprehension is not an issue, and the music is mixed to sound good – whatever the actual meaning of “good.” When checking the mix on consumer monitors, such as on laptop computer loudspeakers, sound is bad, and the main concern switches to how much of each track can be heard properly.

Comprehension of the act of mixing can only be beneficial to the field of automatic mixing. Current literature shows that authors tend to resort to hypotheses such as:

Hypothesis 1. A mix exhibits equal loudness between tracks [7],[9], except for soloing instruments [11].

Hypothesis 2. A mix exhibits equal average perceptual loudness on all frequencies amongst all multi-track channels [6]. Loudness differences on particular frequencies are in most cases an undesired artifact because they induce masking between tracks [11].

We show in this article that hypothesis 1 may not be accurate. Soloing instruments such as lead vocals and guitars may be louder, but equal loudness between the other tracks cannot be observed. This confirms a similar conclusion previously reached by [15] using a completely different methodology. The stronger hypothesis 2 has also been shown as inaccurate, with loudness differences at particular frequencies being common.

We suggest instead that track balance and spectral profile may be governed by principles that are distinct from level equality and the minimization of masking effects between tracks. We hope that the present article will contribute to a better understanding of the mixing process and therefore to the field of automatic mixing.

4 ACKNOWLEDGMENTS

This research is conducted within the Flow Machines project, which received funding from the European Research Council under the European Union’s Seventh Framework Programme (FP/2007-2013)/ERC Grant Agreement n. 291156.

5 REFERENCES


Emmanuel Deruty graduated from the Conservatoire de Paris (CNSMDP), Paris, France. He has worked as a sound designer for IRCAM, Paris, France and Soundwalk, New York, NY, as a film music composer in Europe and in the US, as a writer for the Sound on Sound magazine, Cambridge, UK, and as a consultant in musicology applied to M.I.R. for INRIA, Rennes, France, as well as for IRCAM, Akoustic Arts and Sony Computer Science Laboratory, Paris, France.

Francois Pachet received his Ph.D. and Habilitation degrees from Paris 6 University (UPMC). He is a Civil Engineer (Ecole des Ponts and Chaussées) and was Assistant Professor in Artificial Intelligence and Computer Science, at Paris 6 University, until 1997. He is now director of the Sony Computer Science Laboratory in Paris, where he conducts research in interactive music listening and performance and musical metadata and developed several innovative technologies and award winning systems.

Pierre Roy was born in France in 1970. He studied mathematics and computer science and received the Ph.D. degree from the University Paris 6, Paris, France, in 1998. He then spent a year in Santa Barbara, CA, as a Postdoctoral Fellow in the Center for Research in Electronic Arts and Technology. He came back to France and worked a few years for a multimedia company in Paris. He joined Sony CSL, Paris, in 2005 as a Research Assistant in the music team where he specializes in intelligent digital signal processing techniques with applications in domains such as music delivery, music interaction, environmental sound analysis, and animal behavior.