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 game[1]. 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.

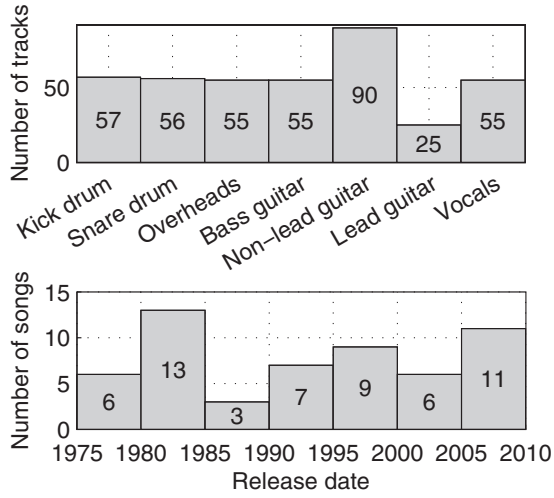

Fig. 1. Overview of the multi-track corpus. Distribution of tracks by instrument (top), and distribution of songs by release date (bottom).

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 (accompaniment/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

[1] http://www.rockband.com/

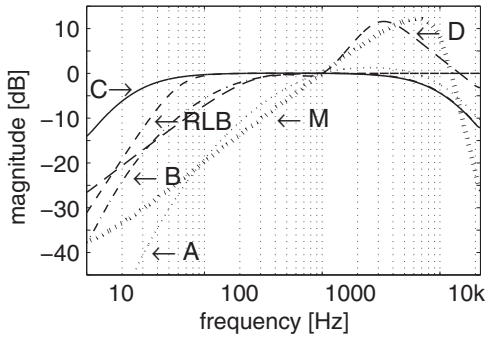

Fig. 2. Common L_{eq} frequency weightings. Adapted from [23].

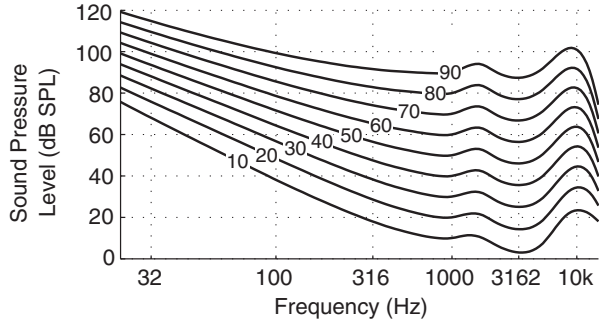

Fig. 3. FWCs corresponding to ISO226:2003. The numbers over each curve correspond to the perceived loudness in phon.

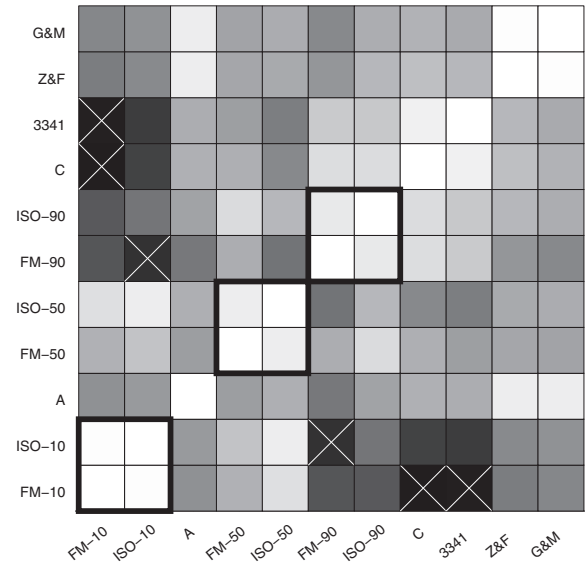

Fig. 4. Comparison between different loudness models, including the custom models presented in this paper. All models are L_{eq} measurements, except for *Z&F* (Zwicker and Fastl) and *G&M* (Glasberg and Moore). Lighter shades of gray indicate higher linear correlations. White crosses denote p values higher than 0.05. The black squares along the diagonal gather models using FWCs corresponding to similar levels.

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[2]. 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, which were 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.

[2]Matlab implementation by Jeff Tackett, http://www.mathworks.com/matlabcentral/fileexchange/authors/17361

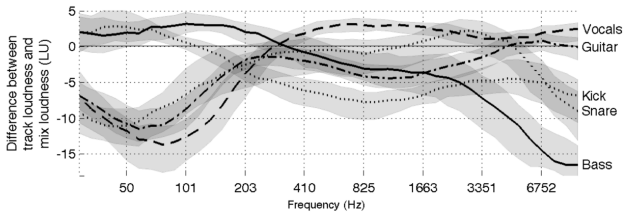

Fig. 5. Median L_{eq} differences between individual tracks from the corpus and the corresponding mix. The gray areas around the median represent the 25th and 75th percentiles. Due to the nature of the L_{eq} models, the differences are model-independent and also correspond to the RMS differences between individual tracks and the mix.

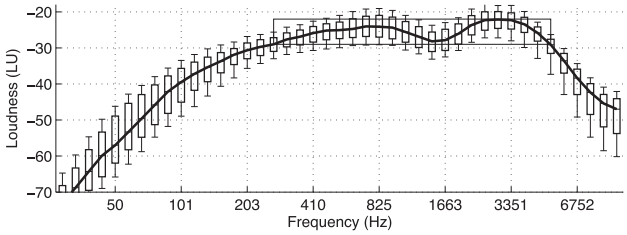

Fig. 6. Power spectrum for the corpus tracks, weighted using the ISO FWC corresponding to 50 phon. The solid line represents the median, the vertical rectangles the 25th and 75th percentiles, and the whiskers the 10th and 90th percentiles. The horizontal rectangle shows the central, flatter zone. The central zone is defined as the largest contiguous zone inside a 6 dB range. This average weighted spectrum is reminiscent of the average spectrum found in [12].

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 envevelope 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

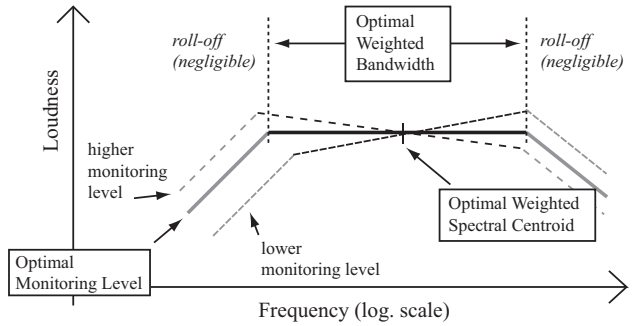

Fig. 7. Power spectrum model using three descriptors: optimal weighted spectral centroid, optimal weighted bandwidth, and optimal monitoring level. At higher and lower monitoring levels, weighted bandwidth is not maximal.

perceived. The greater the weighted bandwidth, the more elements from the track can be heard.

The perceived spectrum depends on the level at which the track is played [25],[26]. Therefore, the weighted bandwidth is dependent on the monitoring level and for each track, there exists a monitoring level for which the weighted bandwidth is the largest. This level we call *optimal monitoring level* (OML), and the corresponding weighted bandwidth the *optimal weighted bandwidth* (OWB). Given that all the loudest frequency bands are concomitant, which we find to be almost always the case in a mainstream rock context, then a third descriptor is the *optimal weighted spectral centroid* (OWSC). It is evaluated as the central frequency for the loudest bands. As illustrated in Fig. 7, these three descriptors, OML, OWB, and OWSC, form a three-dimensional approximation of a track's power spectrum.

In the course of rock music mixing, it is universal practice to listen to the work in progress on high-end and control monitors alternatively [34]. While high-end monitors provide transparency, control monitors simulate band-limited, cheap consumer systems. A comparison of frequency responses from professional midfield and control monitors is shown in Fig. 8, top. The main difference lies in low frequency restitution, which is confirmed by the literature [44],[45]. According to psychoacoustic models, the main difference between frequency perception at low and high monitoring levels also lies in low frequency restitution [27]. As shown in Fig. 8, bottom, both differences are similar. Monitoring level and monitor range have the same influence on frequency perception. Both dimensions can be considered as one. This makes OML, OWB, and OWSC not only pertinent to monitoring level, but also to a scale that ranges from low-level and/or low-end monitoring to high-level and/or high-end monitoring.

We now proceed through the methodology leading to the three spectrum-based descriptors. From either one of the standards, we extract 91 FWCs belonging to the common loudness range for the three standards, 10 to 90 phon. Given any of the three standards, we use as frequency weighting each one of the 91 FWCs. This amounts to performing a term-by-term multiplication of the spectrum by each FWC,

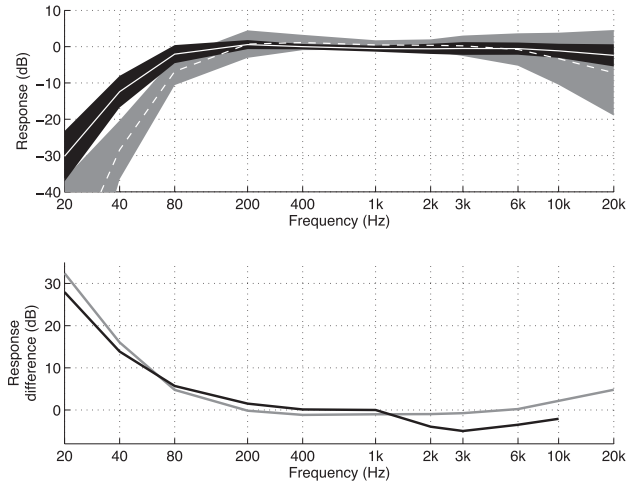

Fig. 8. Top, frequency responses from midfield (solid line) and control monitors (dashed line). Responses are set at 0 dB at 1000 Hz. Solid gray and black areas indicate the standard deviation. Midfield monitors are Adam S3XV [35], Dynaudio BM5 [36], KRK K-RO [36], Genelec 1032B [37], and Neumann O-410 [38]. Control monitors are Auratone 5C Super Sound Cube [39], Avantone Active MixCubes [40], Equator D5 [41], Genelec 6010A [42], and Yamaha NS10M [43]. Bottom, difference between midfield and control monitor response (gray line), and difference between ISO226:2003 equal loudness curves at 61 and 1 phon (black) [26].

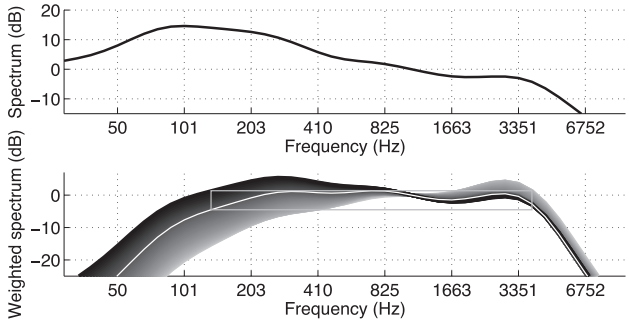

Fig. 9. Top, the power spectrum for the bass guitar track in 30 Seconds To Mars, "Attack", 2005. The data is smoothed for readability purposes, and the 1000 Hz band is set to 0 dB. Bottom, the 91 corresponding weighted power spectra using the Fletcher–Munson model. Darker shades correspond to high levels, and lighter shades to low levels. The data is smoothed for readability purposes. The 1000 Hz band is always set to 0 dB. The white line in the middle shows the weighted power spectrum for the optimum monitoring level, with the 6 dB constraint being illustrated by the rectangle.

which results in 91 weighted spectra per standard. Each power spectrum is then translated so that its value is 0 dB at 1000 Hz.

Fig. 9, top, shows the power spectrum for the bass track from one of the songs in the corpus. Fig. 9, bottom, shows the 91 resulting weighted spectra. For each frequency spectrum, we identify three frequency zones: a low frequency roll-off, a high frequency roll-off, and a middle "flat" zone, whose slope is moderate. Algorithmically, we evaluate the middle zone as the largest contiguous set of frequency bands for which the weighted spectrum can be contained inside a 6 dB span.

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

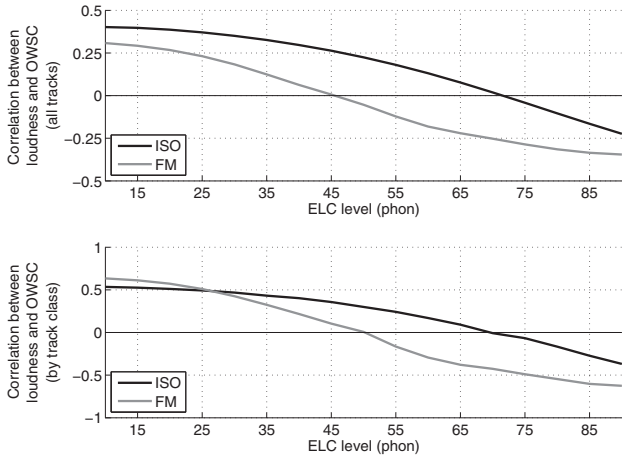

Fig. 10. Correlations between loudness and OWSC. The top diagram shows correlations evaluated on all 392 couples of values. The bottom diagram shows correlations evaluated on the 7 couples of values corresponding to the track classes.

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). Siginificance 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,

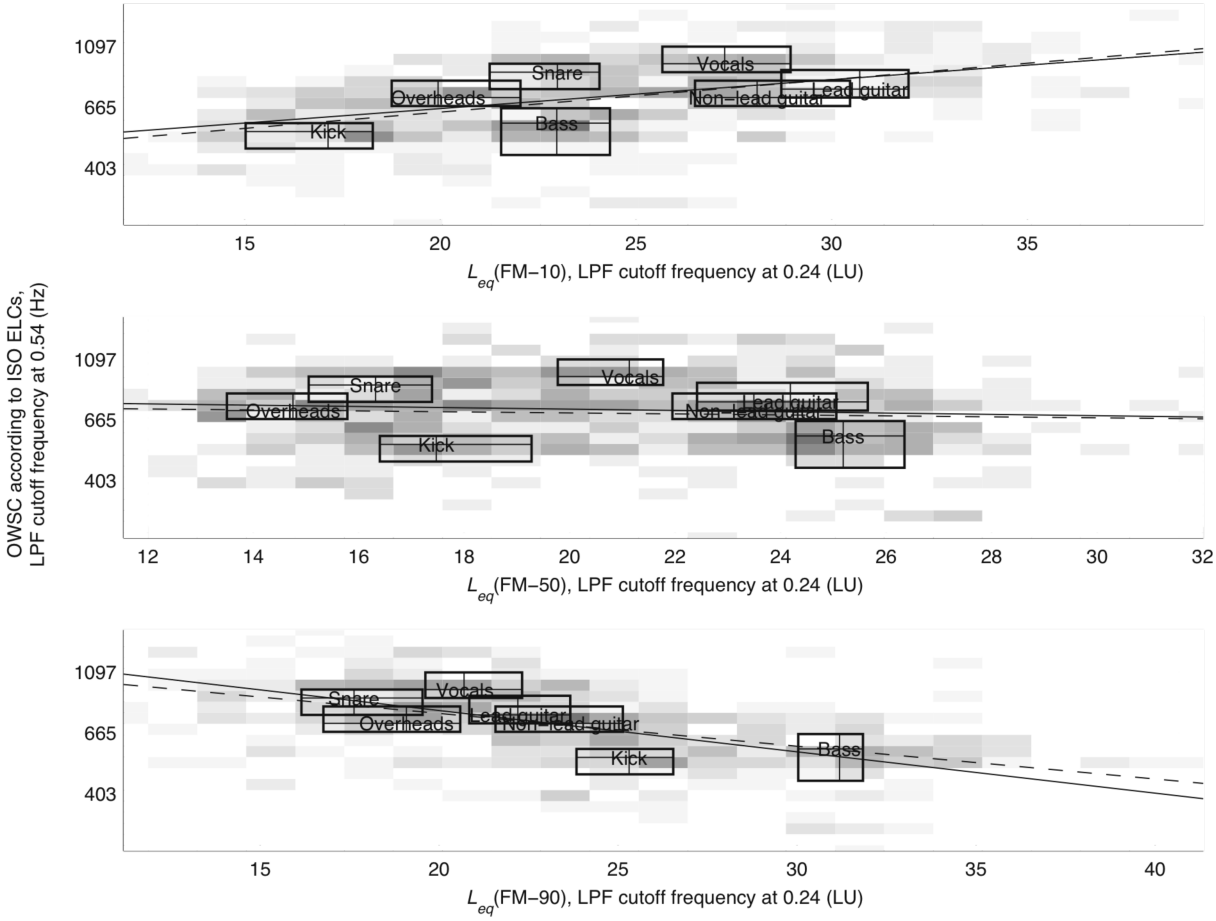

Fig. 11. From top to bottom, OWSC as a function of the loudness values corresponding to 10, 50, and 90 phon FWCs respectively. The loudness model is based on Fletcher–Munson FWCs smoothed using a low-pass filter with a normalized cutoff frequency at 0.24, and the OWSC model on ISO226:2003 FWCs smoothed using a low-pass filter with a normalized cutoff frequency at 0.54. This particular combination of models provides the strongest relations between loudness and OWSC. The name of the track class lies at the median values. The surrounding rectangles indicate the 25th and 75th percentiles. The solid line segment shows the linear regression for the median values. The dashed line segment shows the linear regression for values over all track classes. The grayed areas indicate the distribution of OWSC and loudness values across all track classes.

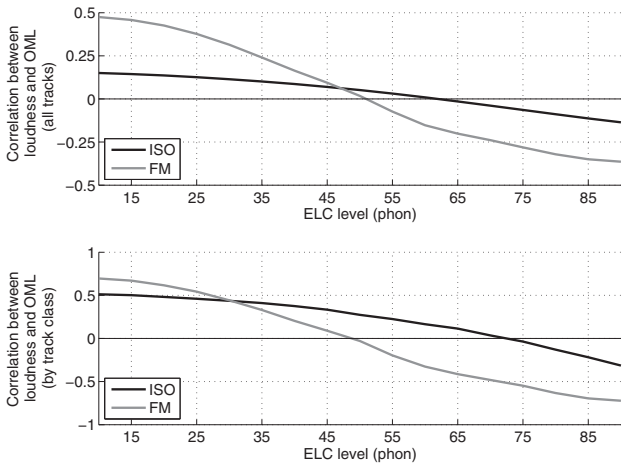

Fig. 12. Correlations between loudness and OML. The top diagram shows correlations evaluated on all 392 couples of values. The bottom diagram shows correlations evaluated on the 7 couples of values corresponding to the track classes.

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

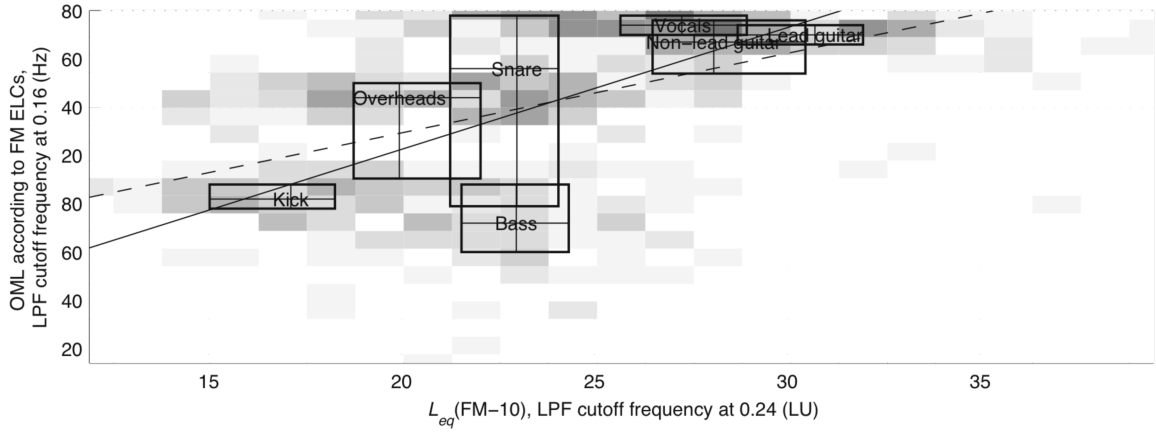

Fig. 13. OML as a function of the loudness values corresponding to 10-phon FWCs. The loudness model is based on Fletcher–Munson FWCs smoothed using a low-pass filter with a normalized cutoff frequency at 0.24, and the OWSC model on Fletcher–Munson FWCs smoothed using a low-pass filter with a normalized cutoff frequency at 0.16. This particular combination of models provides the strongest relations between loudness and OML. The name of the track class lies at the median values. The surrounding rectangles indicate the 25th and 75th percentiles. The solid line segment shows the linear regression for the median values. The dashed line segment shows the linear regression for values over all track classes. The grayed areas indicate the distribution of OWSC and loudness values across all track classes.

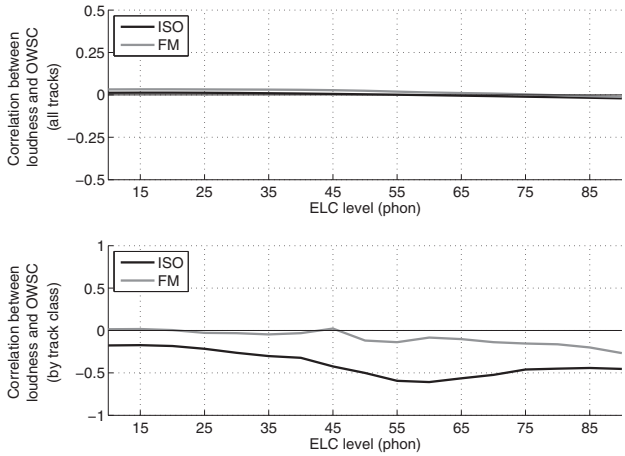

Fig. 14. Correlations between loudness and OWSC based on the degraded corpus. The top diagram shows correlations evaluated on all single couples of values. The bottom diagram shows correlations evaluated on the 7 couples of values corresponding to the track classes.

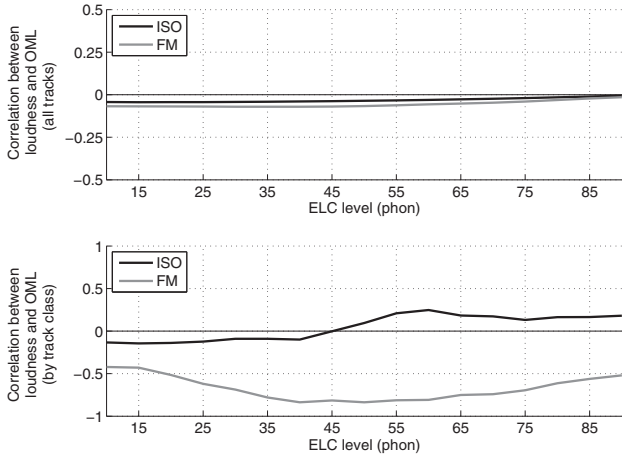

Fig. 15. Correlations between loudness and OML based on the degraded corpus. The top diagram shows correlations evaluated on all single couples of values. The bottom diagram shows correlations evaluated on the 7 couples of values corresponding to the track classes.

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

[1] A. Case, *Mix Smart: Professional Techniques for the Home Studio*, Focal Press, Taylor & Francis, 2011.

[2] M. Senior, *Mixing Secrets*, Taylor & Francis, 2012.

[3] B. De Man and J.D. Reiss, "A knowledge-engineered autonomous mixing system," in *AES 135th Convention*, New York, USA, 2013.

[4] E. Perez-Gonzales, *Advanced Automatic Mixing Tools for Music*, submitted forthePh.D. degree of Queen Mary University Of London, September 2010.

[5] E. Perez-Gonzales and J.D. Reiss, "Improved control for selective minimization of masking using interchannel dependency effects," *Proceedings of the 11th Int. Conference on Digital Audio Effects DAFX-08*, pp. 75–81, 2008.

[6] E. Perez-Gonzales and J.D. Reiss," Automatic equalization of multi-channel audio using cross-adaptive methods," in *AES 127th Convention*, New York, USA, October 9–12, 2009.

[7] S. Mansbridge and al., "Implementation and Evaluation of Autonomous Multi-Track Fader Control," in *AES 132nd Convention*, Budapest, Hungary, April, 2012.

[8] P. Aichinger, A. Sontacchi and B. Schneider-Stickler, "Describing the transparency of mix- downs: The masked-to-unmasked-ratio," in *AES 130rd Convention*, London, UK, May 2011.

[9] D. Ward, J.D. Reiss, and C. Athwal, "Multitrack Mixing Using a Model of Loudness and Partial Loudness," in *AES 133rd Convention*, San Francisco, CA, 2012.

[10] D. Barchiesi and J. D. Reiss, "Reverse Engineering of a Mix," *Journal of the Audio Engineering Society*, vol. 58, pp. 563–576 (2010 July/Aug.).

[11] R.B. Dannenberg, "An Intelligent Multi Track Audio Editor," *Proceedings of the 2007 International Computer Music Conference*, San Francisco, 2007.

[12] P.D. Pestana, Z. Ma, J.D. Reiss, A. Barbosa and D.A.A. Black, "Spectral Characteristics of Popular Commercial Recordings 1950–2010, Convention Paper," in *AES 135th Convention*, New York, USA, 2013.

[13] Z. Ma, J.D. Reiss, and D.A.A. Black, "Implementation of an intelligent equalization tool using Yule-Walker for music mixing and mastering," in *AES 134th convention*, Rome, Italy, 2013.

[14] B. Katz, *Mastering Audio - the Art and the Science*, Focal Press, 2007.

[15] P. Pestana and J.D. Reiss, "Intelligent Audio Production Strategies Informed by Best Practices," in *AES 53rd International Conference*, London, UK, January 2014.

[16] J. Scott and Y.E. Kim," Analysis of acoustic features for automated multi-track mixing," *Proceedings of the 12th International Society for Music Information Retrieval Conference*, pp. 621–626, 2011.

[17] P. White, "Mixing Multitracked Drums," Sound on Sound, February 2001.

[18] Janice Brown, "Thom Cadley: Mixing Rock Bands for Rock Band," retrieved (2014 September) from http://www.sonicscoop.com/2010/03/31/thom-cadley-mixing-rock-bands-for-rock-band.

[19] M. Senior, "Mix Mistakes," Sound on Sound, September 2001.

[20] M. Florentine, A. Popper and R.R. Fay, "Loudness," *Springer Handbook of Auditory Research*, vol. 37, no. 14, 2011.

[21] B.C.J. Moore, B.R. Glasberg and T.A. Baer, "A Model for the Prediction of Thresholds, Loudness, and Partial Loudness," *Journal of the Audio Engineering Society*, vol. 45, no. 4, pp. 224–240, 1997.

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

[23] E. Skovenborg and S.H. Nielsen, "Evaluation of Different Loudness Models with Music and Speech Material," in *AES 117th Convention*, San Francisco, CA, 2004.

[24] G. Soulodre, "Objective Measures of Loudness," *Canadian Acoustics/Acoustique Canadienne*, vol. 32 no. 3, pp. 152–153, 2004.

[25] H. Fletcher, W. A. Munson, "Loudness, Its Definition, Measurement and Calculation," *Bell System Technical Journal*, vol. 12, no. 4, pages 377–430, October 1933.

[26] ISO, "Normal equal-loudness-level contours," Technical Report 226, 2003.

[27] Yoiti Suzuki and Hisashi Takeshima, "Equal-loudness-level contours for pure tones," *Journal of the Acoustic Society of America*, 116 (2), August 2004.

[28] M. Walker, "Mixing on Headphones," Sound on Sound, January 2007.

[29] "EBU - TECH 3341. Loudness metering: 'EBU mode' metering to supplement loudness normalisation in accordance with EBU R 128," EBU/UER, August 2011.

[30] E. Zwicker and H. Fastl, *Psychoacoustics Facts and Models*, 2nd updated edition Springer-Verlag, Berlin/Heidelberg, 1999.

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

[32] P. White, "EQ," *Sound on Sound*, February 1997.

[33] D. Mellor, "EQ: how and when to use it," Sound on Sound, March 1995.

[34] P.R. Newell, K.R. Holland and J.P. Newell, "The Yamaha NS10M: twenty years a reference monitor. Why?," *Proceedings of the Institute of Acoustics*, 23, (8), pp. 29–40, 2001.

[35] Adam Professional Audio, "S3X-V," retrieved (2014 June) from http://www.adam-audio.com/en/pro-audio/products/s3x-v/description.

[36] Phil Ward, "Monitors versus Hi-fi Speakers for Project Studio Monitoring," Sound on Sound, June 2002.

[37] Genelec, "1032B Bi-Amplified Monitor System," retrieved (2014 June) from http://www.genelec.com/products/1032b.

[38] Neumann, "O 410 - Active Studio Monitor," retrieved (2014 June) from http://www.neumann-kh-line.com/neumann-kh/home_en.nsf/root/prof-monitoring_studio-monitors_midfield-monitors_O410.

[39] Recording. de, "Auratone clone/Bosetones," retrieved (2014 June) from http://recording.de/Community/Forum/Recording_und_Studiotechnik/Do-It-Yourself_(DIY)/176701/Post_1945360.html.

[40] Avantone Pro, "Avantone Active MixCube Powered Full-Range Mini Reference Monitors," retrieved (2014 June) from http://www.avantonepro.com/Avantone-Active-MixCube-Powered-Full-Range-Mini-Reference-Monitors.html.

[41] Equator Audio, "D5 Studio Monitors with DSP (Pair)," retrieved (2014 June) from http://www.equatoraudio.com/D5-Coaxial-Studio-Monitors-p/d5.htm.

[42] Genelec, "6010A Bi-Amplified Monitor System," retrieved (2014 June) from http://www.genelec.com/products/previous-models/6010a.

[43] Phil Ward, "The Yamaha NS10 Story, How A Hi-fi Speaker Conquered The Studio World," Sound on Sound, September 2008.

[44] P. Tingen, "Jeff Bhasker on mixing 'We Are Young', Inside Track, Secrets Of The Mix Engineers," Sound on Sound, October 2012.

[45] C. Korff, "Product Review - Spotlight: Secondary Monitors," Sound on Sound, May 2014.

[46] E. Schubert, J. Wolfe and A. Tarnopolsky, "Spectral centroid and timbre in complex, multiple instrumental textures," *Proceedings of the 8th International Conference on Music Perception & Cognition*, Evanston, Illinois, 2004.

[47] G. Milner, *Perfecting sound forever: An aural history of recorded music*, Faber & Faber, 2009.

[48] P. Gonin, *The Cure Pornography*, Discogonie, 2014.

[49] P. White, "Side-chaining In The Software Studio," Sound on Sound, November 2006.

[50] P. Tingen, "Phil Ramone: Producer," Sound on Sound, April 2005.

[51] P. Tingen, "Phil Thornalley: Torn, From Rock Producer To Pop Songwriter," Sound on Sound, June 2010.

THE AUTHORS

Emmanuel Deruty

François Pachet

Pierre Roy

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 lconducts 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.