AUTOMATIC DETECTION OF AUDIO PROBLEMS FOR QUALITY CONTROL IN DIGITAL MUSIC DISTRIBUTION

A PREPRINT

Pablo Alonso-Jiménez[a], Luis Joglar-Ongay[a,b], Xavier Serra[a], and Dmitry Bogdanov[a]

[a]Music Technology Group - Universitat Pompeu Fabra
[b]SonoSuite
Correspondence should be addressed to pablo.alonso@upf.edu *or* luis@sonosuite.com

April 3, 2019

Abstract

Providing contents within the industry quality standards is crucial for digital music distribution companies. For this reason, an excellent quality control (QC) support is paramount to ensure that the music does not contain audio defects. Manual QC is a very effective and widely used method, but it is very time and resources consuming. Therefore, automation is needed in order to develop an efficient and scalable QC service. In this paper we outline the main needs to solve together with the implementation of digital signal processing algorithms and perceptual heuristics to improve the QC workflow. The algorithms are validated on a large music collection of more than 300,000 tracks.

1 Introduction

As digital music services become the largest source of music consumption, artists are more and more interested in getting their creations on streaming and other digital services. In many cases, artists cannot upload their music without a middleman handling contracts and royalties for them. Music distribution companies provide such a service, but small and independent distributors cannot afford to develop their own technology to handle their catalogs, deliver releases to digital service providers and collect their royalties. To provide these services, there exist white label software-as-a-service platforms, such as SonoSuite.[1]

One of the challenges faced by such distribution services is to ensure a great standard in the quality of audio tracks uploaded to the system before sending them to the DSPs to avoid any flaws in the distribution process. To this end, the services may rely on manual quality control (QC), which is very time consuming as a QC agent needs to carefully listen to the tracks and ensure that there are no issues in the audio.

In this paper, we present how digital signal processing algorithms can be used for automation of some steps of the audio QC process, helping in identifying audio problems. We describe a taxonomy for the principal audio problems reported by the SonoSuite QC team. For each problem, a detection algorithm is developed. All presented algorithms are developed by the Music Technology Group in collaboration with SonoSuite, as a part of Essentia [1],[2] an open-source library for audio and music analysis, description and synthesis, and are integrated into the SonoSuite platform. The evaluations are done using the SonoSuite's catalog comprised of more than 300,000 tracks.

2 Audio Problems

The audio problems we address can be divided into five categories according to their nature. Inappropriate silence margins are commonly introduced by human error during the music rendering process:

- *Excessive silence* at the beginning or end of the track.
- *Insufficient silence* at the beginning or end of the track.

Stereo problems are related to incorrect management of the audio channels:

- *False stereo* can be introduced by several causes such as setting twice the same channel on the master output or by digitizing an old mono recording in a stereo setup.

[1] `https://sonosuite.com/`

[2] `https://essentia.upf.edu/`

- *Phase problems* appear when channels are strongly out of phase. Some components of the music will be lost on certain speaker setups.

Digital audio artifacts are characterized by an unnatural disposition of one or more audio samples that do not correspond to a realistic waveform. Sometimes these problems are originated during the process of copy to storage devices as CDs or hard drives. We distinguish the following type of problems:

- *Gaps* are empty audio fragments where the signal goes down to zero or get stuck on some value.
- *Clicks* and *pops* are impulsive noise that may be originated from a variety of causes as plosive sounds on vocal recordings or digitization processes artifacts.
- *Discontinuities* in the waveform are abrupt, unnatural jumps in the waveform. They can be originated by losing some samples of the waveform or by inappropriate tracks cross-fading.
- *Noise bursts* are successions of artifactual samples, typically they are originated by errors in the codification process.
- *Clipping* occurs when the amplitude of the waveform exceeds the available dynamic range.

Loudness and *saturation* problems are related to incorrect use of compression or limiting. In the last decades, the progressive loudness increase in commercial music led to a phenomenon known as "loudness wars" [2], where mastering engineers tried to maximize the loudness of music as much as possible in order to increase the perceived level of excitement in comparison to quieter tracks. However, nowadays most of the music is consumed through streaming services or radios where the playback level is normalized, making such an extreme maximization pointless or even detrimental. We are considering:

- *Loudness* problems appear when the music level is very far from the recommended ranges.
- *Saturation* may appear under extreme maximization parameters.

Noise problems are stationary processes that are present during the whole duration of the track. For this project, we are dividing them into narrowband and broadband noises:

- *Humming tones* are low-frequency narrowband noises typically due to the electricity mains (in the range of 50-80 Hz).
- *Broadband noises* include a variety of noises including vinyl crackle or hiss noise inherent to electronic devices.

3 Related Work

Most of the existing audio quality research is related to the quantification of audio degradation in terms of human perception [3, 4, 5, 6]. However, in our work, the overall quality perception is not as important as finding critical audio problems that can help to speed up QC decisions. In line with this idea, Artega made a taxonomy of audio problems with a detailed review of available solutions [7]. This work is however limited to detection problems related to digitizing vinyl records, being out of scope of our work.

Mühlbauer proposed a collection of detection algorithms for a number of audio defects including *gaps* and *discontinuities*, problems that are also present in our taxonomy [8]. His *gaps* detection algorithm works with the envelope of the rectified audio waveform, using an energy threshold to find gap candidates. In order to distinguish musical pauses from gap artifacts, it is proposed to measure the cumulative power of a small slice before each gap candidate. The idea behind this is that musical pauses tend to be preceded by a fade-out time while this phenomenon is not present on artifactual *gaps*.

For *discontinuity* detection, a Linear Prediction Coefficients (LPC) analysis [9] is performed in a frame-wise manner. LPC coefficients are designed to model autoregressive processes. The prediction error is computed by subtracting the signal reconstructed from the LPC to the original one. Impulsive artifacts, such as *discontinuities* in the waveform, cannot be properly modeled generating huge narrow peaks in these prediction error measurements. Detections are performed by applying an adaptive threshold that dependands on the standard deviation of the prediction error.

Vaseghi studied the problem of detection and correction of *clicks* and *pops* in music with the focus on restoration of music recorded in old analog formats such as phonographic cylinders and gramophone records [10]. Similarly, an LPC analysis is used to describe the stationary part of the audio. *Clicks* are modeled as deltas on the error signal and a matched filter [11] is used to enhance them on the prediction error signals. More recently, Laney proposed fingerprinting techniques in the wavelet domain to identify common flaws on wax cylinders and vinyl records [12].

Brandt and Bitzer investigated the presence of low frequency *humming tones* in music [13]. Their algorithm works by measuring the steadiness of the periodogram frequency bins on 10-30 seconds audio segments. The frequency-wise power stability vector for the section is given by the ratio between the 10th and 55th percentiles of each bin. A stability matrix can be obtained by repeating this algorithm with a moving window throughout the track duration. *Humming tones* are found by thresholding the peaks on the stability matrix.

Solutions for *loudness* and *clipping* have already been standardized in recommendations supported by the broadcasting, audio and film industry. Analog Volume Units meters (VU) [14] were introduced in the 1940s and have been a standard for loudness metering in radio and film mixing. Bob Katz proposed the K-System [15] in the early 2000s as a tool for standardizing loudness in the digital

music industry. Nowadays the recommendation R 128 by the European Broadcasting Union (EBU) [16] has gained a huge popularity as it is proven to have a good correlation to the human perception of loudness [17].

Similarly, digital *clipping* detectors have been available since the first digital audio workstations (DAWs). Various heuristics were proposed for *clipping* detection taking into account auditory perception of *clipping*. For instance, experts say that at least three or four samples in a row have to clip in order to be audible. The recent recommendation ITU-R BS.1770-4 proposes the use of upsampling to make a better estimation of the true value of the peaks in the waveform [18].

Finally, phase correlation meters are used for visual inspection and quantitative measurement of the stereo image [19]. The frame-wise correlation can be computed with the Pearson coefficient between left and right audio channels.

4 Proposed Solutions

In this section, we describe the algorithms developed or adapted for the detection of the audio problems described in the previous section. All the algorithms are publicly available as part of the open-source library Essentia. The source code can be consulted for more details.[3]

Inappropriate silence durations at the track beginning and/or end are detected with a frame-wise power estimator and time thresholds. All frames with a power below a user-specified threshold are considered as silent. The silence problems are detected when the first non-silent frame occurs too soon (*insufficient silence* margin) or too late (*excessive silence* duration). The same checks are applied to the endings of the tracks.

False stereo is detected by computing the Pearson correlation coefficient between both audio channels. When the coefficient is sufficiently close to 1 it is considered a *false stereo* track. The algorithm allows small differences between the channels that may be due to the dithering noise. In contrast, *phase problems* are detected when the coefficient has a negative value meaning that a considerable amount of the left and right signals may be out of phase.

Gaps are detected using a modified version of the algorithm presented by Mühlbauer [8]. As a contribution to the original algorithm, a median filter is used to provide extra filtering and discard too short silent regions, that cannot be perceived as a gap.

Clicks are detected using the already referenced algorithm proposed by Vaseghi [10]. As an addition to the algorithm, an adaptive threshold depending on the standard deviation of the excitation signal was chosen. The threshold is scaled by an arbitrary constant that has to be chosen considering the tradeoff between false positives and misdetections. In order to prevent the threshold to be raised too much by the actual *clicks*, the samples of the excitation signal are clipped to the value of 5 times the median of the excitation vector.

Discontinuities are detected as peaks on the prediction error signal of the LPC analysis similarly to *clicks*. The main difference is that we cannot model them as delta functions in order to apply the matched filter refinement. Instead of that, a median filter is used to smoothen the prediction error while preserving the narrow peaks caused by the *discontinuities*. It was empirically found that a threshold depending on both, the standard deviation and the median of the prediction error showed the best performance. Additionally, it was found that most of the false positives were happening in frames that are partially silent. This is because the average low energy leads to a very low threshold. In order to solve this, an energy threshold is used to mask out silent parts of the frames for the threshold computation.

Noise bursts are detected in the waveform by thresholding the peaks of the second derivative of the signal. We use this algorithm because LPCs are inefficient when many artifacts are present in the same frame. The threshold is computed using an exponential moving average (EMA) filter over the quadratic mean value of the second derivative of the input frame. EMA is used to slightly smoothen the threshold fluctuation across frames.

Clipping is detected following the BS.1770 recommendation [18]. The proposed algorithm works by oversampling the input signal by a factor of 4. After this, there is an optional high shelf filter and a DC blocker. The peaks above 0 dB on the absolute value of the resampled signal are considered clipped.

Loudness is measured following the EBU R 128 standard. The input stereo signal is preprocessed with a K-weighting filter [16]. Integrated loudness is an averaged value over the whole track duration analyzed on blocks of 400 ms. The algorithm uses an absolute and a relative threshold for gating quiet frames.

Saturation is detected by an algorithm that employs two masks. The first one is activated when the energy of the signal is higher than a power threshold. The second one is activated when the derivative of the signal is smaller than a threshold. The intersections between these two masks are considered saturation candidates. Finally, regions shorter than a certain time threshold in milliseconds are discarded because if the saturated region lasts just a couple of samples the degradation will probably not be perceived.

The presence of noise is estimated in terms of signal-to-noise ratio (SNR). The power spectral density (PSD) of the noise is estimated on the audio at the beginning of the track before the music starts. We use a power estimator to detect this start position, and proceed with SNR computation if a sufficient amount of noisy frames have been collected before. The algorithm relies on the minimum mean-square error spectral estimator proposed by Ephraim and Malah [20] to estimate the frame-wise SNR for each frequency bin. An overall SNR estimation for the whole

[3] `https://github.com/MTG/essentia`

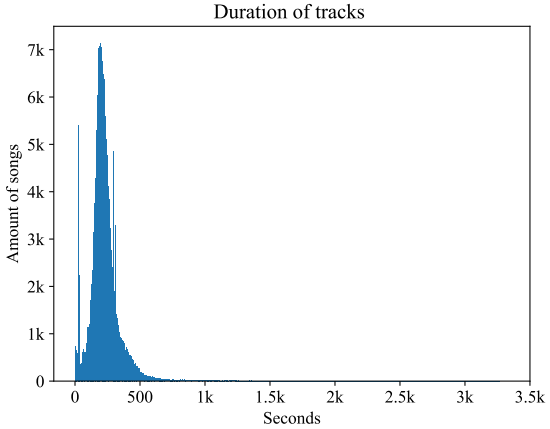


Figure 1: Track durations (s) in the analyzed collection.

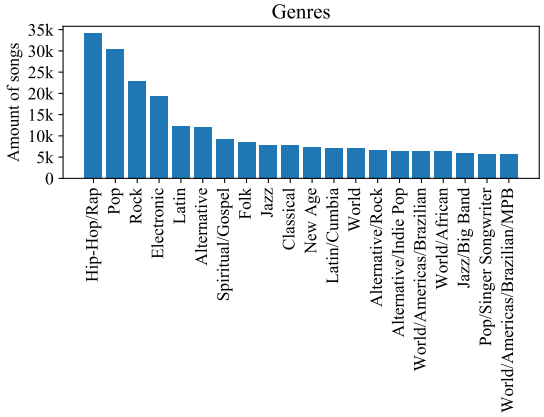


Figure 2: Most common genres in the collection.

spectrum is then computed using an exponential moving average filter.

Low-frequency *humming tones* are found by measuring the steadiness of the PSD of the signal by obtaining the quantile ratios as described in [13]. After this, the salience of a given frequency is computed as the sum of the weighted energies found at integer harmonics of that frequency as done in [21]. The most predominant low-frequency tonal contours are retrieved by tracking the peaks of the salience function.

5 Evaluation

We evaluated the proposed algorithms on a large collection of music tracks from the SonoSuite catalog uploaded by its users. About 80% of these tracks have been distributed and passed through a QC process. Over the years, the manual quality control process at SonoSuite has constantly improved, but only in the last couple of years the procedure has achieved the best quality standards from industry. Therefore, even though distributed tracks have passed a human QC process, some of them still have problems. The other 20% of the catalog has never been distributed so it is more likely to have quality problems. As will be shown by the results, it is clear that the quality of the catalog can improve by using the detection algorithms.

Figure 1 shows the distribution of track durations in the analyzed collection. The vast majority of the tracks are within the 100-300 seconds range with a maximum duration around 54 minutes. There is an isolated peak at the 30 seconds mark, which can be attributed to the ringtones present in the collection. In addition, Figure 2 presents the distribution of the 20 most frequent genres with hip-hop, pop and rock being the most common.

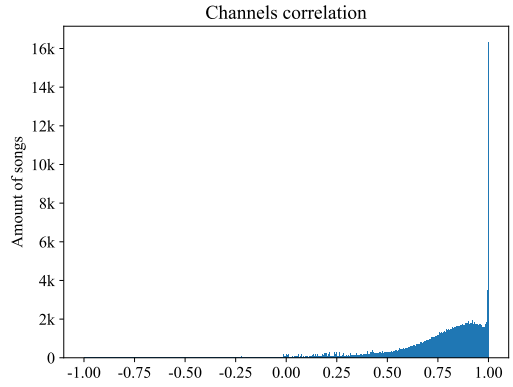
Table 1 shows the percentage of tracks in the collection for which the problems were detected. The histograms in Figure 3 inform about the distribution of the encountered problems.

Figure 3a shows a lobe that peaks around 0.9. This is because most of the stereo music still have a lot of information in the center of the image (typically low frequency instruments such as kick or bass). However, 3.10% of the collection was found to have a correlation coefficient of 1, meaning that there is no stereo information at all. *Phase problems* (negative correlation coefficient) are present on 0.93% of the tracks.

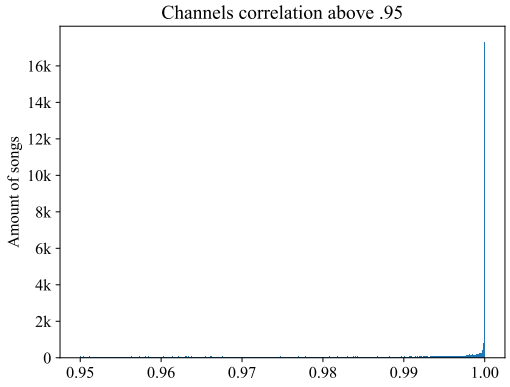
Figures 3c to 3f show the amount of *gaps*, *clicks*, *discontinuities* and *clipping* per time unit. As it was expected, all the distributions peak at a low value where the perceptibility of the problems should be assessed by the QC team. However, all histograms reveal a long right tail with a very high likelihood of audio degradation in the files. For example, 0.46% of the tracks have more than 0.07% of their duration detected as *gaps*. 2.1% of the tracks have more than 5 samples detected as *clicks* or *pops* per minute; 1.09% of the tracks have more than 0.54 *discontinuities* per minute and 5.39% of the tracks have more than 684 clipped samples per minute.

Excessive silence at the start	0.07%
Excessive silence at the end	5.32%
Insufficient silence start	16.32%
Insufficient silence end	9.10%
False stereo	3.10%
Phase problems	0.93%
Gaps	0.75%
Clicks and pops	12.64%
Discontinuities	2.63%
Clipping	49.09%
Loudness problems	19.39%
Noise bursts	24.77%
Saturation	1.63%
Humming	62.63%

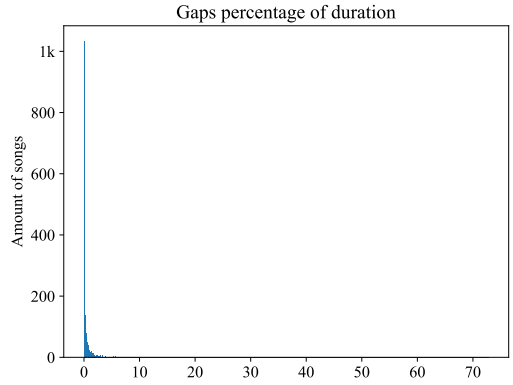
Table 1: Percentage of problems detected.



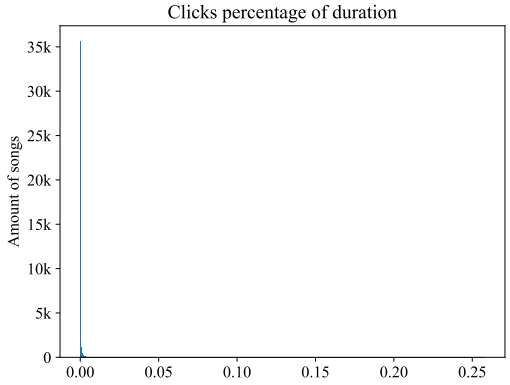
(a) Pearson correlation between left and right channels. Last bin represents tracks with correlation equal to 1.



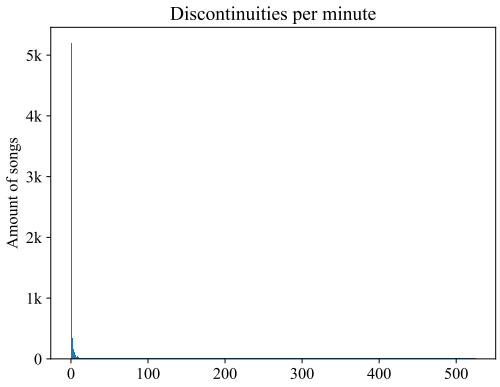
(b) Pearson correlation between left and right channels for tracks with correlation above 0.95.



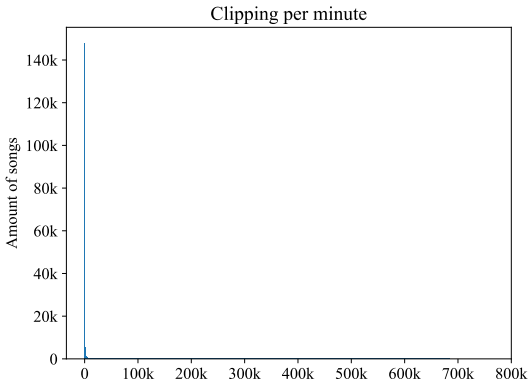
(c) Percentage of a track duration occupied by *gaps* in tracks with detected *gaps* problems.



(d) Percentage of the samples of a track that are part of a click in tracks with detected *clicks* problems.

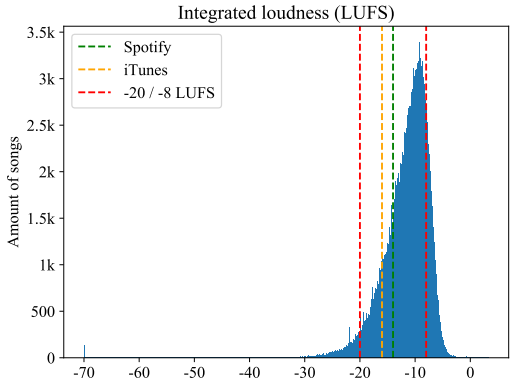


(e) Number of *discontinuities* per minute of audio in tracks with detected *discontinuities* problems.

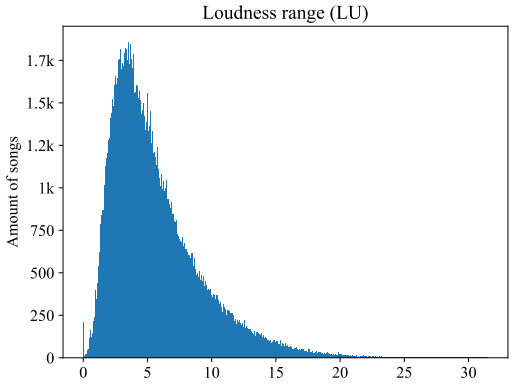


(f) Amount of clipped samples per minute in tracks with detected *clipping* problems.

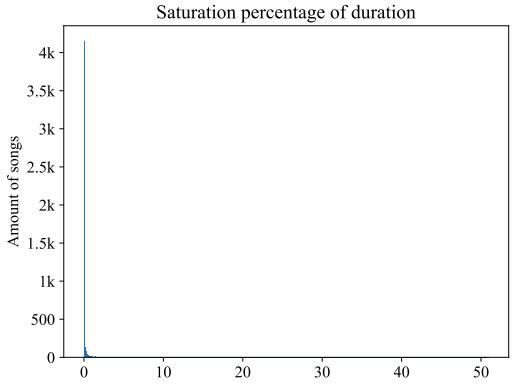
As we can observe in Figure 3g, the integrated loudness value of the analyzed collection peaks at -10 LUFS, which is consistent with a typical target level when mastering for digital downloads or CDs. The vast majority of the collection has this value greater than the loudness playback level used in the streaming platforms to whom SonoSuite



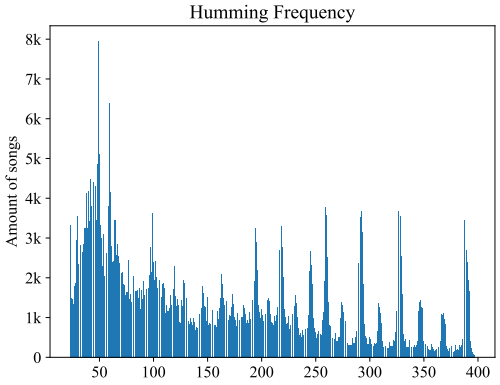
(g) Integrated loudness (LUFS). Streaming loudness levels in Spotify and iTunes are included as a reference.



(h) Loudness range (LU).



(i) Percentage of a track duration occupied by *saturation* regions in tracks with detected *saturation* problems.



(j) *Humming* frequencies (Hz) in tracks with detected *humming* issues.

Figure 3: Distribution of audio problems, loudness and stereo image in tracks in the analyzed music collection.

distributes, such as Spotify or iTunes. The fact that these tracks were not specifically mastered targeting the industry standards for streaming means that their gain will be reduced to a reference level on playback. This may be undesirable, as these tracks will contain less dynamic range and peak headroom due to the extra mastering compression. We cannot see any peak or lobe in the LUFS distribution around the reference levels for streaming suggesting that, for now, there is no tendency from the users to follow them.

To estimate the amount of problematic files, we have set a lower bound of -20 LUFS as a minimum acceptable loudness following the recommendation TD1004.1.15 by AES [22]. For the upper bound, this recommendation proposes a limit of -16 LUFS which is not realistic outside the streaming scenario. Given that -10 LUFS is now typical in CD mastering, we have given a margin of 2 LU, setting the upper boundary at -8 LUFS. It can be seen that 19.39% of the catalog falls out of this range, with 3.42% being too quiet and 15.97% being excessively loud. It is also interesting to see that there are 139 entirely silent tracks (-70 LUFS), accounting for 0.04% of the catalog.

Figure 3i shows that *saturation* also follows a long right tail pattern. In 0.40% of the tracks more than 0.05% of the waveform is saturated. In the most extreme case, one track where more than 50% of the waveform is squashed was found.

Figure 3j shows the distribution of *humming* frequencies among the tracks in the collection. The largest peak occurs around 50 Hz, which is a well known phenomenon associated with alternating current at the frequency of the mains electricity. However, other peaks match the frequency of musical notes, suggesting that the algorithm is sometimes tracking the bassline when it is sufficiently stable. Further refinement of the algorithm parameters should be done to prevent this behaviour.

6 Discussion

The conducted preliminary evaluation demonstrates the utility of the proposed algorithms as a useful tool for the QC process. Analyzing their output we found problematic outliers for all considered audio problems. Therefore, it will be trivial to establish quality thresholds serving as an indicator of a potentially unacceptable degradation for the QC team.

On the other hand, we have detected that some algorithms such as the *hum* and *noise burst* detectors are performing with an excessive sensibility due to inappropriate parametrization. Nevertheless, in our evaluation we have obtained examples of problematic tracks that can now be curated by the QC team and used to fine-tune the algorithms.

Even though the proposed algorithms are capable of detecting common audio problems, there is an open question of whether they are well aligned with the human perception of audio quality. For this reason, the proposed tools are valuable to help the manual QC process, but they are not powerful enough to build an entirely autonomous system able to discern between valid and invalid audio files. We envision that in order to create such a system one can employ machine learning algorithms operating on the raw audio input and the extracted audio quality descriptors. In that case, some of the potential problems to tackle would be gathering a sufficient amount of annotated tracks or finding appropriate parameters for the algorithms showing excessive sensibility such as the *noise bursts* detector.

7 Conclusion

We presented a collection of algorithms that allow to detect common audio quality problems related to digital music distribution and conducted their initial evaluation on a large corpus of music from the SonoSuite platform. All the algorithms are implemented as a part of Essentia, an open-source library for audio and music analysis, description and synthesis, and are configurable and adaptable for reuse in similar problems.

The proposed algorithms can be easily suited for different applications such as curation of audio collections and quality filtering. Such functionality can be highly beneficial for different industrial scenarios including music distribution, music radios, podcasts, audio for video production, DJ applications, music production or user-generated content platforms such as Youtube or Freesound.

In our future work, we will expand the evaluation, including manual validation with the QC team and generation of annotated datasets of audio problems that will allow a thorough evaluation in terms of precision and recall. This will subsequently help to fine-tune parameters of the algorithms for producing outputs that are better correlated with human perception of audio quality.

8 Acknowledgements

This research has received funding from the Centre for the Development of Industrial Technology (CDTI), a public entity under the Spanish Ministry of Economy, Industry and Competitiveness, under the grant agreement No. IDI-20170768.

References

[1] Dmitry Bogdanov, Nicolas Wack, Emilia Gómez, Sankalp Gulati, Perfecto Herrera, O. Mayor, Gerard Roma, Justin Salamon, J. R. Zapata, and Xavier Serra. Essentia: An audio analysis library for music information retrieval. In *International Society for Music Information Retrieval Conference (ISMIR'13)*, pages 493–498, Curitiba, Brazil, 04/11/2013 2013.

[2] Earl Vickers. The loudness war: background, speculation, and recommendations. In *Audio Engineering Society Convention 129*. Audio Engineering Society, Nov 2010.

[3] BS.1387-1, ITU-R. *Method for objective measurements of perceived audio quality*, 2001. first edition.

[4] Rainer Huber and Birger Kollmeier. Pemo-q—a new method for objective audio quality assessment using a model of auditory perception. *IEEE Transactions on audio, speech, and language processing*, 14(6):1902–1911, 2006.

[5] James M Kates and Kathryn H Arehart. The Hearing-Aid Speech Quality Index (HASQI). *Journal of the Audio Engineering Society*, 58(5):363–381, 2010.

[6] Paul Kendrick, Iain R Jackson, Francis F Li, BM Fazenda, TJ Cox, et al. Perceived audio quality of sounds degraded by non-linear distortions and single-ended assessment using hasqi. *Journal of the Audio Engineering Society*, 63(9):698–712, 2015.

[7] Ignasi Adell Arteaga. Automatic detection of audio defects in personal music collections. Master's thesis, Universitat Pompeu Fabra, 2016.

[8] Rudolf Mühlbauer. Automatic audio defect detection, 2010.

[9] Lawrence R Rabiner and Ronald W Schafer. *Digital processing of speech signals*, volume 100. Prentice-hall Englewood Cliffs, NJ, 1978.

[10] Saeed V. Vaseghi. *Advanced digital signal processing and noise reduction*, chapter 13, pages 351–353. John Wiley & Sons, fourth edition, 2008.

[11] George Turin. An introduction to matched filters. *IRE transactions on Information theory*, 6(3):311–329, 1960.

[12] Ryan Laney. Automatic detection of flaws in recorded music using wavelet fingerprinting, 2011.

[13] Matthias Brandt and Joerg Bitzer. Automatic detection of hum in audio signals. *Journal of the Audio Engineering Society*, 62(9):584–595, 2014.

[14] Howard A. Chinn, D.K. Gannett, and R.M. Morris. A new standard volume indicator and reference level. *Proceedings of the IRE*, 28(1):1–17, 1940.

[15] Bob Katz. Integrated approach to metering, monitoring, and leveling practices, part 1: Two-channel metering. *Journal of the Audio Engineering Society*, 48(9):800–809, 2000.

[16] TECH 3343, EBU. *Practical guidelines for production of programmes in accordance with EBU-R 128*, 2011. Version 2.0.

[17] Fabian Begnert, Håkan Ekman, and Jan Berg. Difference between the ebu r-128 meter recommendation and human subjective loudness perception. In *Audio Engineering Society Convention 131*. Audio Engineering Society, 2011.

[18] BS.1770-4, ITU-R. *Algorithms to measure audio programme loudness and true-peak audio level*, 2015. fourth edition.

[19] Hugh Robjohns. What are my phase-correlation meters telling me? `https://www.soundonsound.com/sound-advice/q-what-are-my-phase-correlation-meters-telling-me`, 2016. accessed on 2019-01-09.

[20] Yariv Ephraim and David Malah. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Transactions on acoustics, speech, and signal processing*, 32(6):1109–1121, 1984.

[21] Justin Salamon and Emilia Gómez. Melody extraction from polyphonic music signals using pitch contour characteristics. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(6):1759–1770, 2012.

[22] TD1004.1.15-10, AES. *Recommendation for loudness of audio streaming and network file playback*, 2015. first edition.