Applying Psychoacoustics to Key Detection and Root Note Extraction in EDM

Roman B. Gebhardt and Jonas Margraf

Audio Communication Group, TU Berlin
r.gebhardt@campus.tu-berlin.de

Abstract. In this paper, we present alternate methods for harmonic content analyis of music that factor in psychoacoustic aspects. Based on a model for harmonic analysis by Parncutt we develop a new key detection. We also introduce a new method of harmonic labelling by extraction of the note that is most likely to be perceived as the harmonic "root" of a piece of music, which we call root note extraction. We show that psychoacoustics-based key detection creates results that match a ground truth dataset better than a built-in key detection from commercial software. On top of that we show that for music not fitting the classical major / minor scheme, music mixes based on root note extraction outperform those based on key detection in terms of ratings in a listening test.

Keywords: Harmony, Music Mixing, Key Detection, Root Note, Psychoacoustics, EDM

1 Introduction

The basic challenge in music mixing is to align two or more tracks of music both in their temporal and spectral dimension. Speaking in more musical terminology, the pieces have to fit both in tempo and harmony. State-of-the-art DJ Software like Native Instruments' Traktor Pro 2[1] (hereafter referred to as Traktor) offers methods to alter both dimensions independently using time stretching and pitch shifting techniques. On top of that, automatic beat matching allows the alignment of two tracks in tempo. While beat tracking has been explored quite extensively (see e.g. [1]), harmonic alignment turns out to be a more challenging task. Whereas in Traktor, the built in synchronisation option offers automatic temporal fitting, no such function exists for automatic harmonic alignment. Information about harmony is extracted and provided in form of a major or minor key value. By means of the displayed key value, the user can choose tracks to mix that fit according to the theory of the circle of fifths, e.g. tracks that share even or parallel keys. However, this key detection method can be error-prone: providing incorrect key estimates; failing to capture local key changes; or further, that the

[1] https://www.native-instruments.com/en/products/traktor/dj-software/traktor-pro-2/

key may be ambiguous. Indeed, we believe that for either harmonically simplistic music or music not following the classical western major / minor scheme, e.g. atonal or chromatic music, the key detection method is conceptually the wrong approach for harmonic mixing. In cases like these, the idea of a "root note", which we define to be the sole chromatic note that represents the harmony of a piece best, should be a better descriptor than a key that is applied as a label to a piece of music by a key classification.

The theory of the root of a chord of musical sonority has a long tradition in musicology and has already been covered by the research of Rameau on the *basse fondamentale* in the early 18th century [2]. It has been extended to more mathematical models by Terhardt [3], Hofmann-Engl [4] and Parncutt [5]. The latter of these proposed a computable psychoacoustic model for, among other aspects, the analysis for the salience of different chromas within a simultaneous sound, upon which we build our approach.

Limitations of key detection for harmonic mixing and alternative methods for harmonic alignment have already been discussed by the authors in previous works [6], [7]. While approaches based on the psychoacoustic measure of roughness already showed promising outcomes, similarity measure in terms of pitch commonality derived from harmonic feature extraction by the aid of Parncutt's model did not improve the results. In this work, instead of similarity measures for automatic harmonic alignment we use the model with means towards harmonic extraction in the form of a harmonic label as it can for example be used for user recommendation.

In summary, our work provides two novelties: First, we use and well-founded psychoacoustic model as front-end to our classification stage. Other authors like Pauws [8] have indeed used perceptually inspired methods for their key detection systems. However, to the best of the authors' knowledge, Parncutt's [5] established model has not been used in this context up to the current state. The second innovation and major motivation is the establishment of an alternative system for harmonic analysis in form of the root note. The allocation of a single note from a harmonic entity may be realised with Chew's Spiral Array approach [9], that represents an alternative to key template based classification. However, Chew does not make use of this fact and also builds up her system on 24 pitch classes. We believe our root note approach to be the first of its kind within the field of harmonic extraction from music.

The remainder of this paper is structured as follows: in Section 2, we give an overview of the structure of our algorithm. In Section 3, we first show the results of a key detection method based on our model in comparison with the built-in key detection in state-of-the art DJ software. We then present the results of a listening test comparing mixes resulting from our root-note approach with key-detection based mixes. Finally, we conclude our findings and give an outlook on future work in Section 4.

2 Algorithm Overview

2.1 Pre-Processing

Parncutt's harmonic model deals with simultaneous sounds and takes a set of sinusoids as input. To obtain a series of these from the mono audio files we apply a Short-Time Fourier Transform (STFT) to the signal which has been downsampled to 10.25 kHz beforehand. As parameters, we chose 4096 samples for size of the Fast Fourier Transform (FFT) window (Hanning-type) and 2048 samples for the distance between the analysis windows' centers (hop size). We found the resulting resolution of 2.7 Hz for the sampling frequency to be a necessary minimum especially for low frequencies, which hold critical information on harmonic content [10]. We consequently perform a sinusoidal tracking on the resulting spectrogram to extract the partials. For this purpose, we use Dan Ellis' open access *"Sinewave and Sinusoid + Noise Analysis/Synthesis"* toolbox for MATLAB[2]. To reduce computational time demand and to equalise frequency and magnitude deviations deriving from the spectral analysis, we average over ten time frames by means of the median for both frequency and magnitude values. Subsequently, partials contained in these averaged temporal frames are input to the harmonic model.

2.2 Parncutt's Approach for Harmony Extraction

Parncutt's [5] model represents a black-box system that takes a set of N partials, defined by their frequencies f in Hz and magnitudes M in dB SPL as input and outputs several values modelling the cognitive perception of a simultaneous sound which, following Parncutt, we will further refer to as *simultaneity*). For means of analysis within the musical pitch presentation, the partials are converted from frequency to **pitch categories**, P, in semitone resolution.
The pitch category of the $n-th$ partial is defined by its center frequency f_{n} in Hz:

$$P(f_{\mathrm{n}}) = 12 \cdot \log_2 \left(\frac{f_{\mathrm{n}}}{440\ \mathrm{Hz}} \right) + 57 \tag{1}$$

Hence, the standard pitch of 440 Hz (musical note A_4) would be represented by pitch category 57. Taking into consideration the relative perceptual loudness differences of partials in different regions of pitch, the magnitudes of the **auditory levels** $\Upsilon L(P)$ is calculated

$$\Upsilon L(P) = \max(SPL(P) - TL(P),\ 0), \tag{2}$$

[2] `http://www.ee.columbia.edu/ln/rosa/matlab/sinemodel/`

where $TL(P)$ is a curve modelling equal loudness for sinusoidal sounds:

$$TL(P) = \Re\left\{91 - 49 \cdot \log_{10}(P - 7)\right\}. \tag{3}$$

For the considered pitch range of the modern piano, this function represents an approximation to the threshold in quiet proposed by Terhardt et al. [11]. Next, simultaneous masking effects are considered. Every pitch category's masking effect is modelled according to its respective masking curve, which has its maximum value of $\Upsilon L(P*)$ at the masking pitch category P_*, which declines with increasing pitch distance from P_*. Due to the dissimilar shape of the masking curves in the chroma domain, the pitch classes are transformed to pure tone height $H_p(P)$, which can be understood as equivalent to the *mel* scale of pitch [12]:

$$H_p(P) = \sqrt{\left(\frac{P}{5} - 10\right)^2 + 44} + \left(\frac{P}{5} - 10\right) - 2 \tag{4}$$

As in the *mel*-domain the masking curves hold a pitch-independent triangular shape, the **partial masking levels** $ml(P)$ for each masking pitch category (P_*) can be expressed as a negated absolute value function with a maximum at the masker's auditory level $\Upsilon L(P)$. By applying this procedure for every pitch category masked $(P_\text{-})$, we obtain a two-dimensional matrix containing the masking curves resulting for all pitch categories P_* of a simultaneity or in other words the temporal slice of auditory levels of N pitch categories.

$$ml(P_\text{-}, P_*) = \Upsilon L(P_*) - 25 \cdot |(H_p(P_*) - H_p(P_\text{-}))| \tag{5}$$

For each $P_\text{-}$, by means of simple SPL addition of $ml(P_\text{-}, P_*)$, the global masking level $ML(P)$ is computed, which describes to which degree a pitch category is masked by all others present in the sound. In consequence, the **audible level** $AL(P)$ is derived, which describes the loudness of a pitch category after consideration of equal loudness and masking effects:

$$AL(P) = \max(\Upsilon L(P) - ML(P),\, 0) \tag{6}$$

The third and last filter simulates virtual pitch effects, that is, the perception of the fundamental of a harmonic complex tone by its overtone pattern's structure. In chroma representation, the first overtone lies 12 pitch categories above the fundamental for (corresponding to one octave or double frequency), the second 19 (one octave and a fifth or triple frequency), and so on. Hence, a

harmonic complex tone can be represented by a template with the above described chroma distances. To consider stronger impact of lower harmonics, the template's partials are weighted according to

$$w_H = \frac{1}{H} \tag{7}$$

with H expressing the partial's harmonic numer (1 for the fundamental, 2 for the first octave, and so on). The template is now shifted over the full pitch category range. Matches of the template with present pitch categories are weighted according to eqn. 7, added up and assigned to the its fundamental's pitch category. Hence, the more pitch categories of the analysed sonority fall into this harmonic template, the stronger the virtual perception of its fundamental pitch category. Accordingly, a set of **saliences** $S(P)$ is calculated, which indicates how distinctive the pitch categories are perceived as part of the sonority. By wrapping the pitch categories over the registers r to a 12-dimensional vector, the **chroma salience** S_c is computed for each note or chroma c of the octave:

$$S_\mathrm{c}(c) = \sum_r S(c + 12r) \tag{8}$$

Following eqn. 1, the note C is linked to the chroma number $c = 0$, whereas for Bb, it would take a value of 11. A series of the chroma saliences from an audio file can be displayed in a chromagram representation. Figure 1 shows a comparison of the chromagrams of the same musical excerpt; above derived from the Cannam and Landone's Feature Extraction Plugin of the well-known Sonic Visualizer software[3] (with equal FFT parameters as described above) and below from the harmonic model. It becomes obvious, that our chroma salience representation offers a clearer distinction of chroma as especially vertical structures in the chromagram are reduced. This "de-noised" representation of the chromagram may be explained by two main aspects of our psychoacoustic model: First, the Equal Loudness Filter mainly tackles frequencies in the lower regions that also tend to be smeared to a higher degree because of the frequency independent bin-size of the STFT processing. Second, the Masking Filter proves to be a suitable way to deal with pitches that evoke energy in neighbouring pitch categories: As closer pitch categories have a stronger masking effect towards each other, two neighbouring categories with non-zero levels produced by faulty overlap of bins in the STFT, the masking filter will strengthen the relative difference between them tidy the pitches that produce smeared structures. Apart from that, the sinusoidal tracking that we use prior to the filtering process reduces the signal to tonal information which may prevent distortion by unpitched percussive sounds. We expect the clear nature of the chromagram to have positive impact on the classic key detection procedure.

[3] `http://http://www.sonicvisualiser.org/`

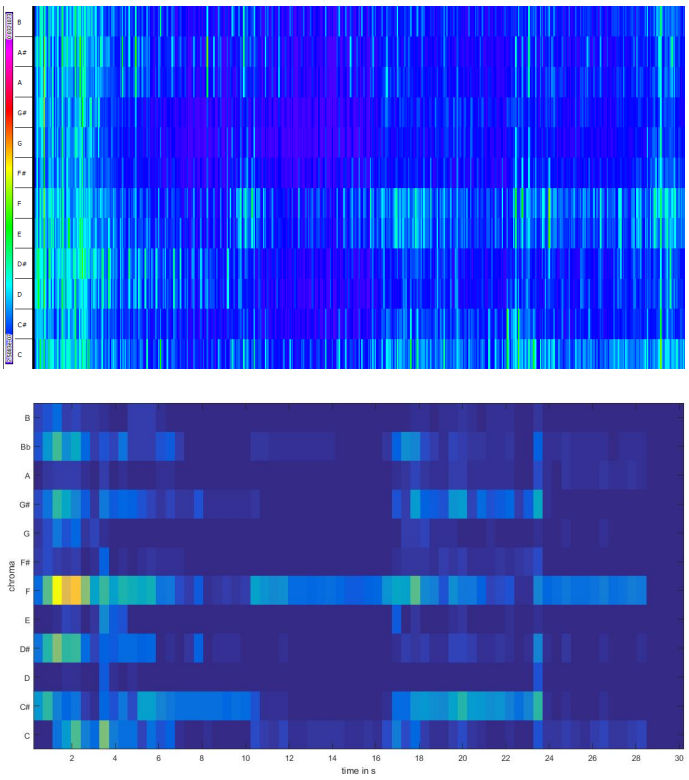

Fig. 1: Chromagram representations of the same musical excerpt derived from the Feature Extraction Plugin for Sonic Visualizer (above) and our system (below).

2.3 Root Note Extraction

Finally, according to Parncutt [5], the measure of chroma salience can be understood as the probability of a note being noticed as the root of the sonority which in our case is corresponding to the *root note* we aim to extract. To attain a global assertion of harmony throughout the whole track, following Parncutt's definition of **chroma tally** we sum the T blockwise chroma salience vectors over time and obtain 12 chroma tally values. We define the chroma obtaining the highest overall tally to be the root note R of the track:

$$R = \text{argmax}(\sum_{t=1}^{T} S_{\text{c}}(c, t)). \tag{9}$$

3 Evaluation

3.1 Ground Truth Dataset

In order to test our algorithm, we employed the Giant Steps data set [13], which consists of 2-minute excerpts of electronic dance music. The data set includes diverse annotations for each track, such as key and genre information. This key and genre data serves as our ground truth. We selected 68 tracks from the full data set, 34 of which had the genre label *Techno* and 34 were labelled *Trance*. These two genres were chosen because we consider them to be very different in the types of harmonies they employ. We assumed *Techno* to be more harmonically minimalistic or ambiguous in terms of its key, and *Trance* pieces to have a more clearly identifiable key. Accordingly, we expected *Techno* to be less clearly categorisable in terms of major or minor key signature in comparison to Trance.

We ran the algorithm on a consumer-grade laptop (2.2 GHz Intel Core i7 with 8 GB of RAM) and identified the third-party sinusoidal tracking function as being the most time-consuming. Our algorithm needed up to several minutes to analyse a single piece. We observed greatly varying execution time for different pieces. We assume this to be dependent on the degree of tonal content within a track and hence, the differing number of sinusoids to be tracked and further processed.

3.2 Key Detection

We applied three different key detection templates to the chroma tallies resulting from our model, namely a binary one (weighting all intervals contained in a scale with one and all others with 0) and those proposed by Krumhansl [14] and Hofmann-Engl [4]. For direct comparison with a state-of-the-art industry standard, we analysed the selected music excerpts using the current standard DJ software Traktor. Consequently, we performed our psychacoustics-based key detection and compared these results to Traktor's key detection and the ground truth. Due to the high computation time of our algorithm, we limited the length of the audio excerpts to be analysed to only 30 seconds in a first run through and 60 seconds in a second pass. The results of these different key detection approaches are presented in Figure 2.

While Traktor's key detection matched the ground truth in 23 out of 68 cases, key detection based on the template introduced by Krumhansl found the correct key for up to 39 music excerpts. Results for the Hofmann-Engl template show poor performance in detecting the correct key for both 30 and 60 second excerpts (8, respectively 7 correct matches). However, for the 30 seconds run the Krumhansl template already outperforms Traktor's detection (29 correct matches), even though only a fourth of the duration used for the Traktor analysis is taken into account. Since there is less harmonic information provided to the algorithm when using a shorter music excerpt, we assume this should actually decrease the precision of the algorithm. Presumably, a more accurate result should be obtained by analysing a longer excerpt. This theory is strengthened

by the better results for the binary (30s: 20 correct matches, 60s: 23 correct matches) and Krumhansl templates in the 60 seconds run where the binary template performs equally well as Traktor and the Krumhansl template correctly detects 39 out of 68 pieces. The relatively bad performance of all methods described may be explained by the fact that the Giant Steps data set draws its ground truth towards key from discussions in online forums. The motivation for such discussions implies that for these tracks there is already a certain degree of uncertainity or ambiguousity of key within the music.

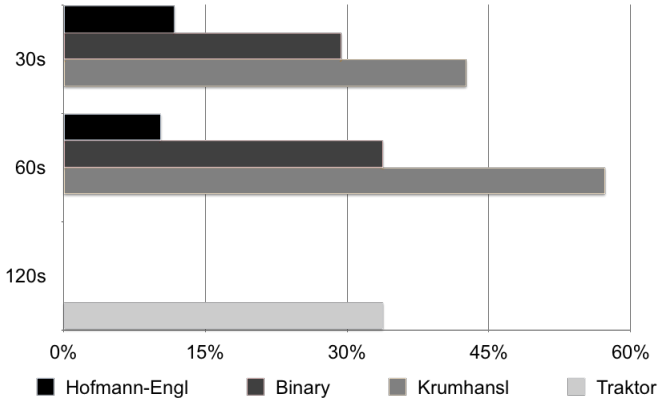

Fig. 2: *Amounts of correctly detected keys for different key detection approaches. While there were two separate runs using 30 and 60 seconds long music excerpts for our tally based method, the analysis using Traktor took the full 120 seconds of the excerpts.*

3.3 Root Note Based Mixing

To evaluate our root note approach we performed an online listening experiment with 38 participants. The experiment was designed as an ABX-comparison. Condition A was a mix of 2 music excerpts that shared the same key according to Traktor's key analysis, while condition B was a mix of 2 key-matched music excerpts according to our model. Conditions A and B shared one "anchor track", while their respective other tracks had to mismatch. We limited our choice of mixes sharing the same anchor track to not match with condition B when condition A matched and vice versa.

Each participant heard 5 ABX-comparisons of *Techno* tracks and 5 ABX-comparisons of *Trance* tracks. For each comparison, participants were asked to choose if they considered one of the two conditions to be more consonant (A / B) or if they perceived both as equally consonant (X). Our expectation was that Traktor's key detection approach would give a good performance for the

Trance pieces, which employed relatively simple and explicit harmonies that can easily mapped to one distinct and unique key label. Moreover, we hypothesised our approach to perform better for the more ambiguous *Techno* pieces as their simplistic harmonic content would not suffice an explicit classification; e.g. tracks with composed with a single-note bassline would both fit the corresponding major or minor key or may also easily be misinterpreted as the dominant or subdominant keys as in many key templates, the fifth is equally weighted as the tonic. The results of the listening experiment are shown in Figure 3. It is particularly noteworthy that listeners' preferences are strongly dependent on the genre of the music excerpts. In 5 out of 5 cases for *Techno*, most listeners preferred the mix based on root note, while for 4 out of the 5 *Trance* cases, most listeners preferred the key detection-based mix. In one *Trance* mix, both methods were rated as equally consonant.
The distribution of consonance ratings conforms to our expectations and strengthens our hypothesis about the root note being a more suitable descriptor for harmonically minimalist music.

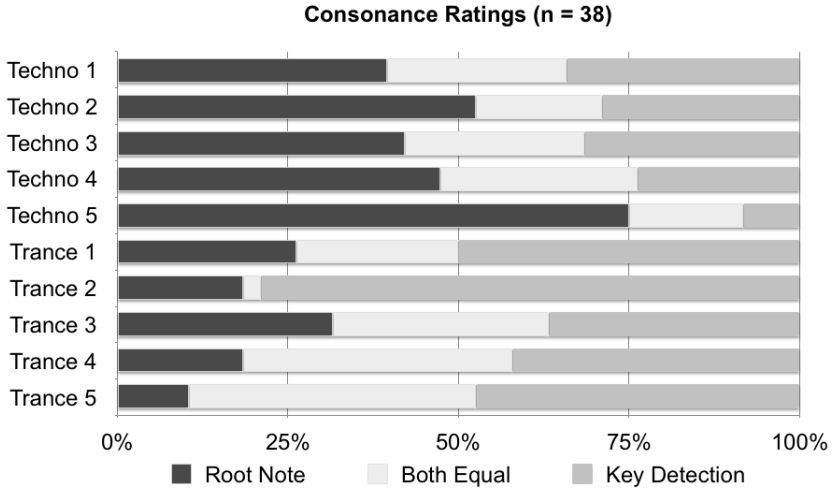

Fig. 3: *Consonance ratings derived from the listening test. For each horizontal bar, the votes for the root note based mix (blue) and the key match based mix (red) matched to a respective anchor track. The X-decision is respresented by the grey area of the bar.*

4 Conclusion

In this paper, we have presented a new approach for key detection and a novel method for harmonic extraction which we call root note extraction for EDM. We have shown that considering psychoacoustics can provide a profitable way for pre-processing music signals for harmonic content analysis.

Our results show that key detection methods which take into account psychoacoustic models can outperform state-of-the-art commercial software which we assumed not to be psychoacoustic approaches. This might be explained by the very clean representation of the chromagram that results from the filtering process which restricts the signal to musically meaningful information. The proposed chromagram computation might as well be interesting to be used as input to a machine learning based back end method for key detection to further improve results.

Our work also shows that for music with an unambiguous key signature, current standard key detection approaches work well, while our root note approach succeeds for music outside the major / minor scale framework.

In its current state, our algorithm is rather processing-intensive. While our algorithm is certainly usable for research purposes, it would be beneficial to optimize our code for speed in order to make further research more efficient or to employ the algorithm in any sort of real-world application.

For this work, we have narrowed down our analysis to the two genres of techno and trance, which we selected due to their contrasting harmonic nature. For future work, we are interested in evaluating our model when applied to other genres of music, which would not necessarily have to be EDM-related. Also music that does not fit the major / minor scheme because of following completely alternative structures would be an interesting test object especially for the root note approach.

We also believe this algorithm can have interesting commercial applications beyond DJ software, such as in music production, where it could be used for harmonic fitting of pitched instrument sounds (such as tonal percussion like toms), but also in automated playlist generation and recommendation systems.

References

1. Zapata, J.R., Davies, M.E.P., Gomez, E.: Multi-Feature Beat Tracking. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 22(4), 816–825 (2014)
2. Christensen, T.: Rameau's "L'Art de la Basse Fondamentale". In: Society for Music Theory (ed.) Music Theory Spectrum, Vol. 9. University of California Press, Berkeley, CA, USA (1987)
3. Terhardt, E.: Akustische Kommunikation. Springer, Berlin (1998)
4. Hofman-Engl, L.: Virtual Pitch and Pitch Salience in Contemporary Composing. In: Proceedings of the VI Brazilian Symposium on Computer Music. Rio de Janeiro, Brazil (1999)
5. Parncutt, R.: Harmony: A psychoacoustical approach. Springer, Berlin (1989)
6. Gebhardt, R.B., Davies, M.E.P., Seeber, B.: Harmonic Mixing Based on Roughness and Pitch Commonality. In: Proceedings of the 18th Int. Conference on Digital Audio Effects (DAFx-15), pp. 185–192. Trondheim, Norway (2015)
7. Gebhardt, R.B., Davies, M.E.P., Seeber, B.: Psychoacoustic Approaches for Harmonic Music Mixing. In: Applied Sciences, Vol. 6, No. 5 (2016)
8. Pauws, S.: Musical key extraction from audio. In: Proceedings of the 5th ISMIR Conference, pp. 96-99. Barcelona, Spain (2004)
9. Chew, E.: Towards a mathematical model of tonality (Ph.D.). Massachusetts Institute of Technology (2000)
10. Faraldo, Á., Gómez, E., Jordà S., Herrera, P.: Key Estimation in Electronic Dance Music. In: Proceedings of the 38th European Conference on Information Retrieval, pp. 335–347. Padua, Italy (2016)
11. Terhardt, E., Paulus, E., Zwicker, E.: Automatic speech recognition using psychoacoustic models. In: Journal of the Acoustical Society of America, Vol. 65, pp. 487 – 498 (1979)
12. Terhardt, E., Feldtkeller, R.: Das Ohr als Nachrichtenempfänger. Hirzel-Verlag, Stuttgart (1967)
13. Knees, P., Faraldo, Á, Herrera, P., Vogl R., Böck S., Hörschläger, F., Le Goff, M.: Two Data Sets for Tempo Estimation and Key Detection in Electronic Dance Music Annotated from User Correction. In: Proceedings of the 16th ISMIR Conference, pp. 364–370. Malaga, Spain (2015)
14. Krumhansl, C.: Cognitive Foundations of Musical Pitch. Oxford University Press, New York (1990)