APPLYING PSYCHOACOUSTICS TO KEY DETECTION AND ROOT NOTE EXTRACTION

Roman B. Gebhardt, Jonas Margraf

Audio Communication Group
Technische Universität Berlin

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 Richard Parncutt we develop a new way of key detection. We also introduce a new method of harmonic labeling 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 such based on key detection in terms of ratings in a listening test.

Index Terms— Harmony, Music Mixing, Key Detection, Root Note, Psychoacoustics

1. INTRODUCTION

The basic challenge in music mixing is to align two or more tracks of music both in its temporal and spectral dimension. Speaking in more musical terminology, the pieces have to be fitting respectively 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 of each other using time stretching and pitch shifting techniques. On top of that, automatic beat matching (so-called 'syncing') allows to align two tracks in their tempo. While this method has been explored quite extensively, harmonic alignment turns out to be a more challenging task. Whereas the 'Sync' button allows automatic temporal fitting, no such function exists for automatic harmonic alignment. Information about harmony is indeed 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 are fitting according to the theory of the circle of fifths, say tracks that share even or parallel keys. Apart from this key detection method turning out to be error-prone in several cases as key changes cannot be taken into account by giving a global key label or the key might be ambiguous in some cases, we state that for either harmonically simplistic music or music not following the classical western major / minor scheme, as for example atonal or chromatic music, the key detection method is deemed to be a conceptually wrong approach. 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 forced to a piece of music by a key detection algorithm.
The remainder of this paper is structured as follows: in Section 2 we present a draft outline of the architecture of Parncutt's model and give a short review of research on the topic of harmonic mixing. In Section 3, we give an overview of the structure of our algorithm. In Section 4 we first show the results of a key detection method based on our model in comparison with 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 5.

2. RELATED WORK

The theory of the root of a chord of musical sonority has already been covered by the research of Rameau on the *basse fondamentale* in the early 18th century [2]. Since then, the theory has been extended by Terhardt [3], Hofmann-Engl [4] and Parncutt [5]. The latter of these proposed a computable psychoacoustic model for - among other aspects - the extraction of the most salient note upon which we build our approach.

Limitations of key detection for music mixing and alternative methods for harmonic alignment have already been discussed by Gebhardt et al. in previous works [6, 7]. While approaches based on the psychoacoustic measure of roughness already showed promising outcomes, similarity measures of information derived from harmonic feature extraction by aid of Parncutt's model did not improve the results. In this work, instead of similarity measures for automatic alignment we approach to use the model with means towards harmonic extraction for user recommendation.

3. ALGORITHM OVERVIEW

3.1. Theory of Harmonic Extraction

Parncutt's model represents a black-box system that takes a set of N partials, defined by their frequencies f in Hz and magnitudes M in dBSPL as input and allows to output a variety of gradually computed values modeling the cognitive perception of a simultaneous sound. To enable harmonic pattern recognition, as explained below, the partials are converted from frequency to pitch categories in semitone-distance.
The pitch categories, P, are defined by their center frequencies in Hz:

$$P(f_n) = 12 \cdot \log_2(\frac{f_n}{440Hz}) + 57 \quad (1)$$

Hence, the standard pitch of 440 Hz (musical note A_4) is represented by pitch category 57.

In terms of audibilities of single partials, various psychoacoustic aspects are taken into account: A successive filtering is applied to the magnitude levels taking into account loudness differences of partials in different frequency regions, masking effects and the theory of virtual pitch, meaning the perception of the fundamental of a harmonic pattern. Such a tone is defined by its harmonics of which the first harmonic reaches up 12 pitch categories above the fundamental (corresponding to one octave or double frequency), the second harmonic 19 (one octave and a fifth or triple frequency), and so on. The more partials of the analysed sonority fall into this harmonic template, the higher is the given value of its fundamental's audibility. On top of this, the multiplicity, meaning the number of partials potentially being perceived at one time is determined. Based on this, a set of pitch categories are calculated, which are defined by their number expressing pitch and their respective **salience** $S(P)$, which indicate how distinctive they are perceived as part of the sonority. By wrapping these pitch categories over the registers r to a 12-dimensional salience-vector, the **chroma salience** S_c is computed for each note or chroma c of the octave:

$$S_c(c) = \sum_r S(c + 12r) \quad (2)$$

According to equation 1, the note C is linked to the chroma number $c = 0$, whereas for Bb, c would take a value of 11. 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.

3.2. Application on Audio Data

As stated in the above subsection 3.1, Parncutt's harmonic model takes a set of partials as input. Due to the fact that a piece of music consists of a progressive harmonic structure, we interpret it as a series of simultaneous sounds. Therefore, to obtain the partials of this progression of sounds we apply a Short-Time Fourier Transform (STFT) to the audio signal. As parameters, we chose the levels of 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 bin-size of 5.38 Hz for a sampling frequency of 44.1 kHz to be a necessary minimum as especially low frequencies, which hold critical information on harmonic content [8] would show poor spectral resolution for smaller window sizes, whereas bigger ones would not coincide with the assumption of the time frames as simultaneities. 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[1]. To reduce computational time demand as well as to equalise frequency and magnitude deviations deriving from the spectral analysis, we average over 8 time frames by means of the median for both frequency and magnitude values. Subsequently, partials contained in these averaged 'time blocks' are transferred to the harmonic model, which computes the chroma saliences for each block. The resulting chroma saliences of an example audio file are represented visually in figure 1.

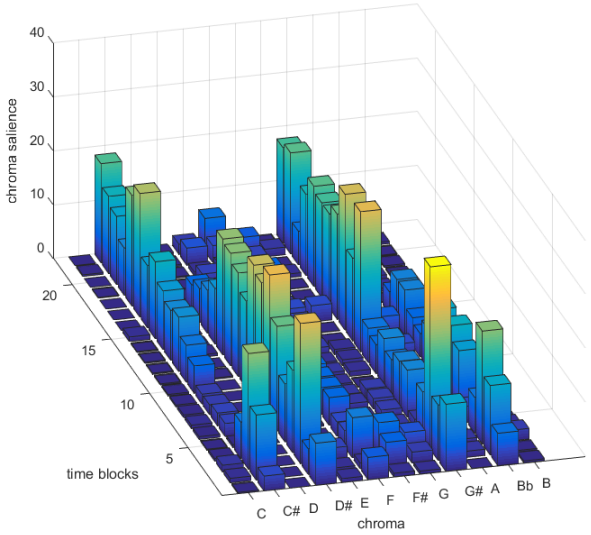

Fig. 1. *Blockwise chroma saliences computed from a musical excerpt. High bars indicate a high likeliness for the note to be perceived as the root of the time block's sonority.*

To attain a global assertion, following Parncutt's definition of **chroma** tally we sum the T blockwise chroma salience vectors over time and obtain 12 chroma tally values. We de-

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

fine the chroma obtaining the highest overall tally to be the **root** R of the track:

$$R = argmax(\sum_{t=1}^{T} S_c(c,t)) \quad (3)$$

4. EVALUATION

4.1. Ground Truth Dataset

In order to test our algorithm, we employ the Giant Steps data set [9], which consists of 2-minute excerpts of electronic dance music. These excerpts include annotations for each track, such as key and genre information. This key and genre data serves as our ground truth. In order to balance computation time and conclusiveness of the results 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. Our asumption is that *Techno* is either more harmonically minimalistic or ambiguous in terms of its key, and that *Trance* pieces will have a more clearly identifiable key signature. We also expect *Techno* to be less clearly categorizable as having a major or minor key signature in comparison to Trance.

4.2. Key Detection

As a first step, we analyzed the selected music excerpts using current industry standard DJ software Traktor. We then 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 [10] and Hofmann-Engl [4]. Using this procedure, we performed a psychacoustics-based key detection and compared these results to Traktor's key detection and the ground truth. Due to the very high computation time of our algorithm, we limited the length of the audio excerpts to be analyzed 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 analyzing 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.

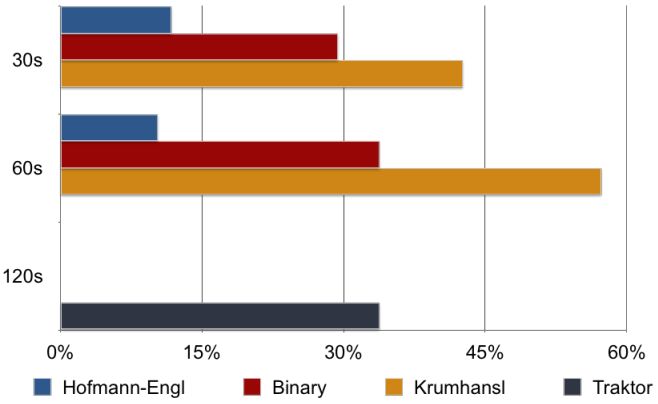

Fig. 2. *Amounts of correctly detected keys for different key detection approaches. While there were two separate runs using 30 and 60 seconds long music excerpts for our tally based method, the analysis using Traktor took the full 120 seconds of the excerpts.*

4.3. Root Note Extraction

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, and that our approach would perform better for the more ambiguous *Techno* pieces. The results of the listening experiment are shown in Figure 3. It is particularly noteworty 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. This conforms highly to our expectations.

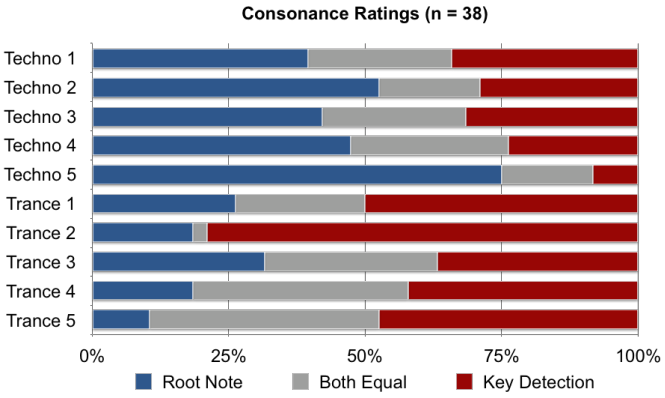

Fig. 3. *Consonance ratings derived from the listening test. For each horizontal bar, the votes for the root note based mix (blue) and the key match based mix (red) matched to a respective anchor track. The X-decision is respresented by the grey area of the bar.*

5. 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. We have shown that considering psychoacoustics can provide relevant insights for harmonic content analysis of music.

Our results show that key detection methods which take into account psychoacoustic models can outperform state-of-the-art commercial software.

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 limited in its effectiveness because it is rather processing-intense. We have identified the third-party sinusoidal tracking function as being the most time-consuming. While our algorithm is certainly useable 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 future work, we are interested in evaluating our model when applied to other genres of music, as we 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.

6. REFERENCES

[1] "Native Instruments Traktor Pro 2 web presence," https://www.native-instruments.com/en/products/traktor/dj-software/traktor-pro-2/, Accessed: 2016-08-19.

[2] T. Christensen, "Rameau's "l'art de la basse fondamentale"," in *Music Theory Spectrum Vol. 9*, Society for Music Theory, Ed., pp. 18–410. University of California Press, Berkeley, CA, USA, 1987.

[3] E. Terhardt, *Akustische Kommunikation*, Springer, Berlin, 1998.

[4] L. Hofman-Engl, "Virtual pitch and pitch salience in contemporary composing," in *Proceedings of the VI Brazilian Symposium on Computer Music*, 1999.

[5] R. Parncutt, *Harmony: A psychoacoustical approach*, Springer, Berlin, 1989.

[6] R. B. Gebhardt, M. E. P. Davies, and B. Seeber, "Harmonic mixing based on roughness and pitch commonality," in *Proceedings of the 18th Int. Conference on Digital Audio Effects (DAFx-15)*, Trondheim, Norway, December 2015, pp. 185–192.

[7] R. B. Gebhardt, M. E. P. Davies, and B. Seeber, "Psychoacoustic approaches for harmonic music mixing," *Applied Sciences*, vol. 6, no. 5, pp. 123, 2016.

[8] S. Hainsworth and M. Macleod, "Onset detection in musical audio signals," in *In Proceedings of International Computer Music Conference (ICMC)*, Singapore, December 2003, pp. 136–166.

[9] P. Knees, Á. Faraldo, P. Herrera, R. Vogl, S. Böck, F. Hörschläger, and M. Le Goff, "Two data sets for tempo estimation and key detection in electronic dance music annotated from user correction," in *Proceedings of the 16th ISMIR Conference*, Malaga, Spain, October 2015, pp. 364–370.

[10] C. Krumhansl, *Cognitive Foundations of Musical Pitch*, Oxford University Press, New York, 1990.