A CONFIDENCE MEASURE FOR KEY LABELLING

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

Athanasios Lykartsis
Audio Communication Group,
TU Berlin
athanasios.lykartsis@tu-berlin.de

Michael Stein
Native Instruments GmbH
michael.stein@native-instruments.de

ABSTRACT

We present a new measure for automatically estimating the confidence of musical key classification. Our approach leverages the degree of harmonic information held within a musical audio signal (its "keyness") as well as the steadiness of local key detections across the its duration (its "stability"). Using this confidence measure, musical tracks which are likely to be misclassified, i.e. those with low confidence, can then be handled differently from those analysed by standard, fully automatic key detection methods. By means of a listening test, we demonstrate that our developed features significantly correlate with listeners' ratings of harmonic complexity, steadiness and the uniqueness of key. Furthermore, we demonstrate that tracks which are incorrectly labelled using an existing key detection system obtain low confidence values. Finally, we introduce a new method called "root note heuristics" for the special treatment of tracks with low confidence. We show that by applying these root note heuristics, key detection results can be improved for minimalistic music.

1. INTRODUCTION

A major commercial use case of musical key detection is its application in DJ software programs including Native Instruments' Traktor [1] and Pioneer's rekordbox [2]. It represents the basis for harmonic music mixing [9], a DJing technique which is mostly bounded to electronic dance music (EDM). However, the concept of musical key is not universally applicable to all styles of music, especially those of a minimalistic nature, which is often the case in (EDM) [7, 10, 21]. A particular challenge of key detection in EDM is that the music often does not follow classic Western music standards in terms of its harmonic composition and progression. This applies to a broad range of contemporary EDM music which can be composed in a chromatic space or, if following classic characteristics, uses more "exotic" modes such as e.g. Phrygian [19], which is actually predominant for certain genres such as Acid House, Electronic Body Music (EBM) and New Beat, which, since the 1980s represent a prominent source of inspiration for contemporary EDM. A further difficulty is the tendency of certain electronic music to be strongly percussive and very minimalistic in terms of its harmonic content [5]. In fact, following pioneering groups like Kraftwerk, melodic minimalism is a main characteristic of techno music [13]. Today, a wide range of EDM productions are exclusively percussion-based. The lack of harmonic information clearly leads to problems in assigning an unambiguous key label, which is still the most widely used way to describe a track in its harmonic composition [21].

In the recent years, confidence measures have gained interested in the field of MIR, namely related to tempo estimation [8,17]. The described scenario motivates to establish such measure for key detection tasks. Crucial factors to consider are the degree to which a musical audio signal conforms to the concept of musical key, and furthermore to explore where a single key persists throughout a recording. Being able to capture this information automatically could therefore serve as an indicator to predict potential misclassifications. It may also be used to define a threshold to decide whether to label a track with a key or alternatively simply with a root note [10], within a genre-specific framework [21] or in spatial coordinates [2,3,12]. Alternatively, multiple key labels could be assigned for tracks containing key changes [16]. We collate this information to derive a key detection confidence measure and present an alternative means for handling music where a traditional key assignment is not be possible. The remainder of this paper is structured as follows: in Section 2, we present the development of the confidence features as well as a special key detection method for tracks of a minimalistic nature. Section 3 outlines our evaluation of the developed features and the special treatment of low confidence scoring tracks. Finally, we conclude our work and provide an outlook for future work in Section 4.

2. METHOD

To establish the confidence measure, we follow two hypotheses and for each we develop a feature: First, there must be sufficient harmonic information within the signal

[1] https://www.native-instruments.com/de/products/traktor/
[2] https://rekordbox.com/de/

to reliably determine a key, i.e., it would be inappropriate to label a track consisting exclusively of percussive content with a meaningful key. Consequently, we denote our first confidence feature as *keyness* to indicate the amount of harmonic content within a musical piece. Second, we state that any local key changes throughout the duration of a track will inevitably lead to a discrepancy between a given global label and at least some regions. Our second confidence feature, which measures the steadiness of key information, will be referred to as *stability*. The development of both features is discussed in the following subsections.

2.1 Keyness

Various approaches have been taken to the problem of assigning a musical key designation based on the information retrieved from an audio signal. A straightforward method would be to follow the well-known key template approach introduced by Krumhansl et al. [15], where the correlation of an input signal's chroma distribution with the chosen key's template could be used as a keyness measure. Often, these templates are not needed, for instance when the key detection is handled within a tonal space model like Chew's Spiral Array [3] or Harte et al.'s Tonal Centroid Space [12]. To avoid the necessity of computing the correlations and to keep our approach most simple, we bypass this option and retrieve keyness information directly from the chromagram. For this, we use a chromagram representation which empahsizes tonal content, based on a perceptually inspired filtering process in [10]. This procedure removes energy in the chromagram evoked by noisy and/or percussive sounds, which are especially present in EDM. We then apply Chuan et al.'s fuzzy analysis technique [4] to further "clean" the chromagram. Figure 1 shows the resulting chromagram of an EDM track [3] with a temporal resolution of $250\,\mathrm{ms}$ and below it, the curve resulting from the sum of the frame-wise individual chroma energies $E(c,t)$ ranging from 0 to 1 for each chroma c at time-frame t:

$$E_c(t) = \sum_{c=1}^{12} E(c,t). \tag{1}$$

We denote $E_c(t)$, the *chroma energy*. By inspection of the resulting curve, a raw subdivision of the track into three partly recurring harmonic structures can be observed: The first with a chroma energy equal (or close to) zero is present in the purely percussive regions which accord to our represent regions of *low* keyness. The second structure describes the G# power chord (where G# is the root and D# the fifth), which reaches chroma energy values of 1 to approximately 1.75 for $E_c(t)$. The power chord is widely used in EDM productions and is ambiguous in terms of the mode of its tonic's key due to the third missing. Finally, the third structure in the middle of the track holds a chroma energy level of approximately 4 which far exceeds the other regions. In fact, it is the only region that contains a sufficient number of notes present to use as the basis for detecting the key. As this representative example demonstrates, the straightforward calculation of chroma energy can be informative about how much harmonic information is contained in a musical audio signal.

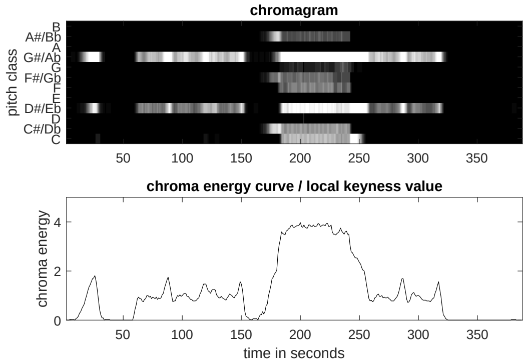

Figure 1. Chromagram (upper plot) and local keyness curve (lower plot) of an EDM track derived from the framewise energies of the chromagram.

To obtain a global keyness measure, we average the chroma energy vector $E_c(t)$ over the full duration T of the track and obtain the keyness value, K:

$$K = \frac{1}{T} \cdot \sum_{t=0}^{T} E_c(t). \tag{2}$$

2.2 Stability

The second confidence feature, stability, is derived from the steadiness of key classifications throughout the full duration of the track. For this purpose, we take into account the vector of local key detections using a template-based approach on temporal frames with $250\,\mathrm{ms}$ length and $125\,\mathrm{ms}$ hop-size. In DJ software which was the framework of our research, the 24 key classes are usually displayed in the 12-dimensional subspace of so-called "Camelot numbers" [6] each of which corresponds to a certain "hour" on the circle of fifths. This implies that a major key and its relative minor are considered equivalent. The middle plot of Figure 2 shows the progression of Camelot classifications over time. It is important to note that both the vertical axis of the middle plot and the horizontal axis of the lower histogram plot are circular i.e. the chroma has been "wrapped". In our example, the most frequently detected Camelot number is 1 (B/G# m) which is followed by its direct neighbour one fifth above, 2 (F# / Ebm). The right tail of the distribution fades out with small counts for numbers 3 (Db/Bbm) and 4 (Ab/Fm), whereas the left tail's only present value is 11 (A/F# m). For a high degree of stability, we would expect a low angular spread of camelot detections throughout, which we compute in terms of the circular variance $V(cam)$ of the distribution according to [1] . In terms of a numeric measure for the stability

[3] Praise You 2009 (Fatboy Slim vs. Fedde Le Grand Dub): https://www.discogs.com/de/Fatboy-Slim-vs-Fedde-Le-Grand-Praise-You-2009/release/1967533

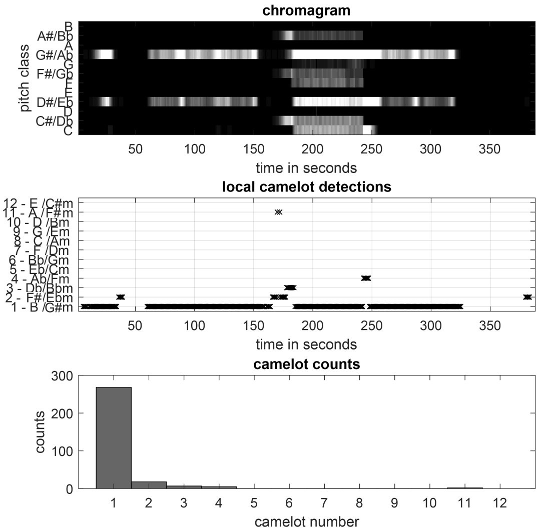

Figure 2. Local camelot decisions (middle plot) and histogram of absolute camelot counts (lower plot). The number describes the "hour" on the circle of fifths.

of the whole track, we define the confidence feature of stability, S, as:

$$S = 1 - V(cam), \tag{3}$$

with $V(cam)$ depicting the circular variance of the camelot vector. Thus, the stability of a track will be 0 for a uniform histogram and 1 for maximum stability (where only one camelot number is detected throughout). In more complex compositions in classical music, we can expect key changes throughout musical pieces. However, these key changes are usually small moves on the circle of fifths and consequently small steps on the Camelot wheel (e.g. just one "hour" for a fifth). When using the circular histogram, these key changes would not have a strong impact on the variance of the distribution and would therefore exert only a small influence on the stability feature. In the special case of pop or EDM, key modulation is mostly absent [7].

2.3 An Overall Confidence Feature

In the two previous subsections, we discussed the development of two features to measure the *keyness* and *stability* of musical audio, both representing independent approaches to find a quantitative measure for the overall confidence of a key detection. As discussed, the two features focus on different characteristics of the music signal. While the keyness measure describes the amount of harmonic information held by a track, the stability feature focusses on the steadiness of key detections throughout a whole track. Collectively these features will penalise the presence of key changes within a track as well as "random" labels from a key classification system caused by harmonic structures which don't conform to the classic major / minor distribution. We state that, for a "trustworthy"' key detection which is informative for harmonic mixing, a given track should score high for both of these features. Thus, we define an overall confidence feature as the linear combination, C, of the subfeatures K and S with variable weighting parameters κ and σ. We quantise K and S and discretise them individually to evenly distributed percentiles, resulting in C_k for K and C_s for S. As a result, the lowest percentile of 1 comprises tracks scoring lower in K (or S respectively) than 99% of the database which is discussed in Section 3.2. This is done to ensure an even distribution of the subfeature values over all tracks as well as to map both to a range from 1 to 100:

$$C = \frac{\kappa \cdot C_k + \sigma \cdot C_s}{\kappa + \sigma} \tag{4}$$

We consider the choice of κ and σ to be genre-dependent. For minimalistic music such as EDM, where we do not expect highly complex harmonic structure or key changes that would eventually lead to a low score for C_s, we believe greater emphasis should be given to C_k to filter e.g. purely percussive tracks. However, for the analysis of classical music, more importance should be attributed to the stability feature C_s. Here, we should not expect a lack of harmonic information, but frequent and "far" key changes would lead to less clarity about the key the piece is composed in. In this paper, we set the values of $\kappa = 5$ and $\sigma = 2$ for the evaluation of a database mainly containing EDM tracks, however we intend to explore the effect of modifying these values and genre-specific parameterisations in future work.

2.4 Root Note Heuristics

With the proposed confidence feature, C, it is possible to determine a threshold below which a key detection should not be considered reliable. This raises an important question of how to treat problematic (i.e. low confidence) tracks in terms of assigning a key label. One option could be the use of multiple key labels for tracks with low stability [16] or to use root note labelling for tracks with low keyness [10]. Alternatively, for EDM, minimalistic tracks could be labelled as the root note's minor key due to the strong bias towards minor mode in this genre [7, 14]. We call this procedure *"root note heuristics"* and apply it to tracks whose keyness falls below a certain threshold. For the case of root note detection, we first accumulate the chroma energies $E(c,t)$ over time to obtain a global chroma energy vector $\underline{E}(c)$:

$$\underline{E}(c) = \sum_{t=0}^{T} E(c,t). \tag{5}$$

To detect the most predominant chroma, and hence *root note*, we apply a simple binary template $T(c)$ in which the referenced chroma and its dominant are given an equal weight of 1, with all pitch classes set to 0. Consideration of the fifth interval is made to explicitly take power chords into account and allow them to point towards their root. We shift this template circularly by one step for each chroma value accordingly and calculate the inner product per shift.

This results in the likelihood $R(c)$ of the chroma c to be the root of the track:

$$R(c) = < T(c), \underline{E}(c) > \tag{6}$$

Finally, the minor mode of the chroma with the highest value of $R(c)$ is assigned to the track as a whole.

3. EVALUATION

For an extensive analysis of our developed confidence features, we undertook two separate evaluation procedures. First, to examine the validity of our subfeatures *keyness* and *stability*, we conducted a listening test where we asked participants to rate a set of musical audio examples according to three questions concerning their harmonic content. Second, we evaluated the degree to which the calculated confidence score for each single track would be associated with a given genre label and whether it was detected correctly by a key detection system and - if not - whether the error was close to the ground truth key label or not. Hence, we would then be able to use the confidence score as a prediction measure for the potential rejection of a key decision and eventually the special treatment of the corresponding tracks. Both approaches are discussed in the following subsections.

3.1 Listening Test for Subfeature Evaluation

The listening experiment was performed as an online survey, in which we presented 12 different representative excerpts[4] of length 120 s which we considered sufficient to allow the perception of any potential key changes. These 12 excerpts could be characterised by the following four properties A - D:

A: Clear and unique key throughout (Track IDs 1, 8, 12)
B: Change in key structure (Track IDs 2, 7, 10)
C: Non-Western melodic content (Track IDs 3, 4, 6)
D: No or little melodic content (Track IDs 5, 9, 11)

After listening to the audio samples, participants were asked to rate them on a 10-point Likert scale in terms of their harmonic complexity, i.e. whether the tracks followed the major/minor scheme and how clearly they adhered to one unique key throughout. In order to prevent any bias in the participant ratings, no information about the developed features was provided. However, a short training phase was set up before the test to ensure participants understood the questions they were going to be asked. In total, we recruited 29 participants (22 male, 7 female) who self-reported as musically trained. The participants' ages ranged from 23 to 66 with an average of 10 years of musical training. In the following sections, the relatedness of the ratings with the computed subfeatures C_k, C_s as well as the overall confidence C will be discussed.

3.1.1 Keyness

To assess the subfeature of keyness, we asked participants to rate the audio excerpts according to two questions. With the first, we aimed to test if the concept of the keyness feature as a general measure for tonal density or complexity (not necessarily relating to a key) would prove appropriate:

Q1: *"To which degree do you find the presented audio harmonically complex?"*

We hypothesised a positive correlation between the ratings and the computed values of C_k, however we made no assumption about the coherence of the ratings with C_s as harmonically complex excerpts could still be unstable in harmony or key. The mean ratings as well as the corresponding feature values C, C_k and C_s are displayed in the leftmost column of Figure 3. For a measure of relatedness, we calculated Spearman's rho correlation measure for the ratings' means across participants and the feature values. With a choice of $\alpha = 0.05$ as the level of significance, the observed strong positive correlation ($\boldsymbol{r_s = 0.63, p < 0.05}$) between the ratings and computed values for the keyness feature C_k supports our initial hypothesis. However, some outliers can be identified, for which the formulation of the question might have been misleading: Excerpt 7 (second rated from category B) exhibits strong break beat percussion and a rather chaotic melodic progression with a short minor mode piano passage, which would contribute to a low score for C_k. Feedback from some participants revealed the excerpt was considered as rather challenging, which caused it to be rated high in terms of complexity. Excerpt 9 (the highest rated excerpt from category D) is also mostly percussive with pitched voice samples and sounds. Its relatively unusual composition might also have caused some participants to rate it "complex". The excerpts from category A consist of quite common, repetetive chord structures which therefore may not have been perceived as particularly complex in a musical sense. However, they all feature a high amount of harmonic content, and therefore represent "complex" musical excerpts in line with our keyness definition. As discussed in 2.1, the keyness feature is derived from the average amount of tonal information throughout the analysed signal. We argued that in the case of Western music, a high amount of tonal information usually indicates the presence of a major or minor scheme as harmonic layerings of notes deviating from Western scales rarely appear [20] and thus a higher density of tonal information should point towards the clear presence of a musical key. To examine the validity of this assumption, the second question of the listening test focussed on whether the keyness feature could in fact be used as an indicator for the presence of a major/minor scheme within the audio:

Q2: *"To which degree does the presented audio fit the major/minor scheme?"*

[4] A link to the examples will be provided in the camera ready copy.

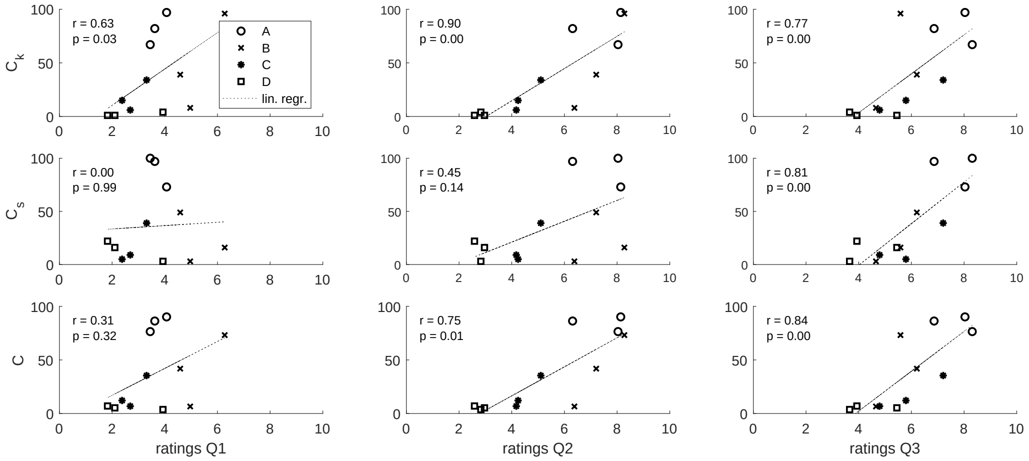

Figure 3. Mean ratings on the questions **Q1**, **Q2** and **Q3** for the 12 stimuli and their corresponding feature values C, C_k and C_s with the respective Spearman correlation coefficients r.

Again, we hypothesised a positive correlation between the ratings and C_k, but again, not C_s. The results are presented in the subplots in the middle column of Figure 3. Our hypothesis regarding C_k was supported with a very strong positive correlation of $\boldsymbol{r_s = 0.90, p < 0.01}$. Remarkably, it even exceeds the correlation of the stronger hypothesis we explored in **Q1** regarding its relatedness to the complexity ratings, as discussed in 3.1.1: Of the four outliers discussed above, namely one excerpt from category D and all the excerpts from category A, all agree much more strongly with the C_k value. As with **Q1**, no significant correlation between the ratings and the values of C_s was observed.

3.1.2 Stability

To evaluate the stability subfeature C_s, participants were asked to rate the stimuli according to the question:

Q3: *"To which certainty does the audio correspond to one unique and distinct key?"*

We expected the ratings for **Q3** to be correlated with the computed values of C_s, as key changes should results in lower the ratings and stability. In addition, we also hypothesised a positive correlation to C_k as a lack of harmonic information could complicate a clear assignment to one unique key. The subplots in the rightmost column of Figure 3 show the outcomes of the third question. As can be seen, both subfeatures exhibit a significant correlation with the mean ratings. While C_k shows a strong positive correlation with a Spearman coefficient of $\boldsymbol{r_s = 0.77, p < 0.01}$, the correlation of C_s is even stronger ($\boldsymbol{r_s = 0.81, p < 0.01}$). The combination of both in the overall confidence feature C results in an even higher correlation $\boldsymbol{r_s = 0.84, p < 0.01}$, which fortifies our choice to combine both features in order to explain the certainty of a unique key decision and therefore the confidence of a key assignment.

3.2 Evaluation on an annotated dataset

In the second part of our evaluation progress, we tested how the computed confidence scores relate to genre labels and whether a track's key classification was correct or not. We based our analysis on a private commercial database comprised of 834 tracks consisting mainly of EDM (697 total) as well as 137 tracks from Harte's [11] Beatles dataset with key labels forming the ground truth. A subset of 101 of the EDM tracks were labelled "Inharmonic" and represented tracks that were considered ambiguous or unclassifiable by musical experts.

3.2.1 Genre Specific Differences

For a first observation, we compare the means of C_k and C_s for the three different subsets, namely the Beatles, the "Inharmonic" labelled EDM subset, and the remainder of the EDM tracks. According to our model, C_k should be high for the Beatles dataset, since it contains mostly melodic music. However, we should expect lower values for the EDM set, following the hypothesis that EDM is often of a more minimalistic melodic nature. For the subset of EDM tracks labelled "Inharmonic" we shouldn't expect much harmonic information, and hence lows score for C_k. Alternatively, a lack of clarity about the label might occur due to the use of a non-Western scale, and would therefore result in a low value for C_s. We hypothesised C_s to reach higher scores for the remaining EDM tracks as we expected a more stable melodic structure for these than the Beatles tracks which inherit a number of key changes and sometimes unconventional harmonic content. The results in table 1 show that our expectations are confirmed. The "Inharmonic" subset scores substantially lower in all (sub)-features, while the Beatles dataset scores high in keyness whereas the remainder of the EDM dataset achieves high values in stability.

Subset	C_k	C_s	C
EDM	49.4	55.8	51.2
EDM Inharmonic	14.3	30.6	19.0
Beatles	82.0	42.1	70.6

Table 1. Confidence score means for the different subsets, in the range 1 - 100.

3.2.2 *Prediction of Misclassification*

We aimed to assess whether the the confidence feature would be an appropriate indicator of the degree to which an automatic key detection could be considered trustworthy, primilary for the application of harmonic mixing. To provide automatic estimates of musical key, we used a key-template based system built into a state-of-the-art DJ software, which was modified by incorporating the preprocessing stage as proposed in [10]. Given our equalisation of relative keys to equal Camelot numbers as discussed in 3.1.2, we defined three different labelling categories: *Match* for key detections matching the ground truth label, *Fifth* for fifth related errors and thus, one Camelot number away from the ground truth and *Other* for detections greater than one Camelot number apart. Across the 834 tracks, we counted 627 *Matches*, 117 *Fifths* and 90 *Others*. Three hypotheses were put forward: We expected tracks for which our key detection result matched the ground truth to score higher in confidence than those from both other categories. We were less sure about the tracks from the *Fifth* category, but intuitively expected them to score higher than those from *Other*. Figure 4 shows the distributions of the confidence scores C within the three groups. We performed a Welch-ANOVA which supported this hypothesis with high significance, $\boldsymbol{F(2, 170.41) = 64.16, p < .001}$. To test the mean differences between the three groups, we conducted a Games-Howell post-hoc analysis which showed significant differences between all three pairs for $\alpha = 0.01$.

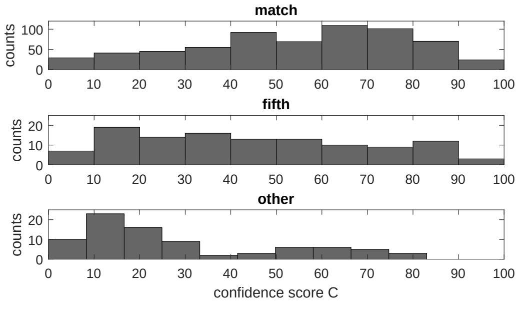

Figure 4. Distributions of the confidence scores C within the three different labelling categories.

3.2.3 *Root Note Heuristics*

Finally, we evaluated the special treatment of the “root note heuristics” introduced in 2.4. For this, we took into consideration the counts of the three labelling categories between the different subsets. As a preliminary investigation, we applied the heuristics to the lowest sixth quantile scoring tracks. The resulting absolute counts for the labelling categories are shown in Table 2. While the Beatles and normal EDM subsets are barely affected, a clear improvement is achieved within the “Inharmonic” subset. Using the root note heurestics, the number of correctly detected tracks could be increased by 18%. Furthermore, we were able to reduce the number of *Other* classified errors by 26%.

Subset	Match	Fifth	Other
EDM	469 / **468**	86 / **85**	41 / **43**
EDM Inharmonic	50 / **59**	17 / **17**	34 / **25**
Beatles	108 / **108**	14 / **14**	15 / **15**

Table 2. Counts of labelling categories for the three subsets without / **with** the application of the root note heuristics method.

4. CONCLUSIONS

In this paper, we described the development of a confidence feature for key labelling, as a means to measure the likelihood of an automatic key classification being correct. For this, we developed two subfeatures, keyness and stability, to estimate the amount of tonal content of musical audio as well as the steadiness of key detections throughout the full duration of the track respectively. Both subfeatures were evaluated by means of a listening test. Our analysis demonstrated high correlations for harmonic complexity, accordance to the major/minor scheme and the uniqueness of one key between the participants’ ratings and the developed features. Furthermore, we showed that our confidence feature can be helpful indicator of cases where an automatic estimated key label can be trusted. Our confidence measure may also be used as a threshold to switch between different key detection approaches. To this end, we introduced a root note heuristics method that can be used as a special key detection approach for tracks of harmonically minimalistic nature, and we showed that the application of this procedure could positively affect key detection performance. However, the presented root note heuristics approach is still at an early stage of development, therefore these promising results motivate continued research towards adjusting the threshold and further development of alternative key detection methods. This work has mostly been focussed on EDM. A major area of future work would therefore be to generalise the key confidence concept for other genres, where it would be neccessary to also take into account relative errors instead of considering only in the Camelot subspace. Also, other possible ways to use the developed features can be considered: Since the keyness feature is sequentially analysed over time, this allows inference about individual segments of a track. In the context of harmonic mixing, this information could be extremely useful by allowing a DJ to locate appropriate regions for executing the transition between two tracks, thus avoiding harmonic clashes [9, 18].

5. REFERENCES

[1] P. Berens. Circstat: A MATLAB toolbox for circular statistics. *Journal of Statistical Software*, 31(10):1–21, 2009.

[2] G. Bernandes, D. Cocharro, M. Caetano, and M.E.P. Davies. A multi-level tonal interval space for modelling pitch relatedness and musical consonance. *Journal of New Music Research*, 45(4):281–294, 2016.

[3] E. Chew. *Towards a mathematical model of Tonality*. Ph.D. thesis, MIT, Cambridge, MA, 2000.

[4] C.-H. Chuan and E. Chew. Fuzzy analysis in pitch class determination for polyphonic audio key finding. In *Proc. of the 6th International Society for Music Information Retrieval (ISMIR 2005) Conference*, pages 296–303, 2005.

[5] G. Dayal and E. Ferrigno. *Electronic Dance Music*. Grove Music Online, Oxford University Press, 2012.

[6] Á Faraldo. *Tonality Estimation in Electronic Dance Music*. Ph.D. thesis, UPF, Barcelona, 2017.

[7] Á Faraldo, E. Gómez, S. Jordà, and P. Herrera. Key estimation in electronic dance music. In *Proc. of the 38th European Conference on Information Retrieval*, pages 335–347, 2016.

[8] F. Font and X. Serra. Tempo estimation for music loops and a simple confidence measure. In *Proc. of the 17th International Society for Music Information Retrieval Conference (ISMIR 2016)*, pages 269–275, 2016.

[9] R.B. Gebhardt, M.E.P. Davies, and B.U. Seeber. Psychoacoustic approaches for harmonic music mixing. *Applied Sciences*, 6(5):123, 2016.

[10] R.B. Gebhardt and J. Margraf. Applying psychoacoustics to key detection and root note extraction in EDM. In *Proc. of the 13th International Symp. on CMMR*, pages 482–492, 2017.

[11] C. Harte. *Towards automatic extraction of harmony from music signals*. Ph.D. thesis, University of London, London, 2010.

[12] C. Harte, M. Sandler, and M. Gasser. Detecting harmonic change in musical audio. In *Proc. of the 1st ACM workshop on Audio and music computing multimedia*, pages 21–26, 2006.

[13] J. Hemming. *Methoden der Erforschung populärer Musik*. Springer VS, Wiesbaden, 2016. In German.

[14] 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 corrections. In *Proc. of the 16th International Society for Music Information Retrieval (ISMIR 2015) Conference*, pages 364–370, 2015.

[15] C. Krumhansl, E. Kessler, and J. Edward. Tracing the dynamic changes in perceived tonal organization in a spatial representation of musical keys. *Psychological Review*, 89(4):334–368, 1982.

[16] K. Noland and M.B. Sandler. Key estimation using a hidden markov model. In *Proc. of the 7th International Society for Music Information Retrieval (ISMIR 2006) Conference*, pages 121–126, 2006.

[17] J. Pauwels, K. O'Hanlon, G. Fazekas, and M.B. Sandler. Confidence measures and their applications in music labelling systems based on hidden markov models. In *Proc. of the 18th Conference of the International Society for Music Information Retrieval (ISMIR 2017)*, pages 279–279, 2017.

[18] M. Spicer. (ac)cumulative form in pop-rock music. *Twentieth Century Music*, 1(1):29 – 64, 2004.

[19] P. Tagg. From refrain to rave: The decline of figure and the rise of ground. *Popular Music*, 13(2):209 – 222, 1994.

[20] D. Temperley. *Music and Probability*. MIT Press, Cambridge, MA, 2007.

[21] R. Wooller and A. Brown. A framework for discussing tonality in electronic dance music. In *Proc. Sound: Space - The Australasian Computer Music Conference*, pages 91 – 95, 2008.