WHAT IF THE 'WHEN' IMPLIES THE 'WHAT'?: HUMAN HARMONIC ANALYSIS DATASETS CLARIFY THE RELATIVE ROLE OF THE SEPARATE STEPS IN AUTOMATIC TONAL ANALYSIS

Mark Gotham[1,2], Rainer Kleinertz[2], Christof Weiß[3], Meinard Müller[3], Stephanie Klauk[2]

[1] Institut für Musik und Musikwissenschaft, Technische Universität Dortmund, Germany
[2] Institut für Musikwissenschaft, Saarland University, Germany
[3] International Audio Laboratories Erlangen, Germany

ABSTRACT

This paper uses the emerging provision of human harmonic analyses to assess how reliably we can map from knowing only *when* chords and keys change to a full identification of *what* those chords and keys are. We do this with a simple implementation of pitch class profile matching methods, partly to provide a benchmark score against which to judge the performance of less readily interpretable machine learning systems, many of which explicitly separate these *when* and *what* tasks and provide performance evaluation for these separate stages. Additionally, as this 'oracle'-style, 'perfect' segmentation information will not usually be available in practice, we test the sensitivity of these methods to slight modifications in the position of segment boundaries by introducing deliberate errors. This study examines several corpora. The focus on is symbolic data, though we include one audio dataset for comparison. The code and corpora (of symbolic scores and analyses) are available within: https://github.com/MarkGotham/When-in-Rome

1. INTRODUCTION

Since the pioneering work of Carol Krumhansl and colleagues in the 1980s, [1,2] there have been several proposals for prototypical key profiles in tonal music based on the relative importance of the constituent pitches in a key (examples follow in context below). The motivations and data for this approach derive usually from either psychological tests (asking 'how well does this pitch fit this key context?', for instance), empirical usage data ('how often is it used?'), or a combination of the two. The working hypothesis is that there exists a link between this pair: that substantial past exposure to the statistical regularities of a musical style forms a mental representation which affects our expectations when listening.

The main task for these 'prototypical' profiles in the empirical domain is automatic key finding, either for an entire work ('what is *the* key of this piece'), or for passages (*keys* plural) with the latter often approached in terms of key matching for the usage within ranges delimited by a moving window [3,4].

The idea is that if these *prototypical* pitch profiles give us a strong sense of the relative usage of each pitch in a key, and we also also have data for the *actual* pitch usage in a score or audio source of interest, then we can simply compare the source with the prototypes for each key and find the one that fits best.

Several more sophisticated algorithms have been proposed to replace the whole practice of matching profiles [5,6], or enhance that practice in place [7], but there is an enduring attraction to the clarity and simplicity of this approach. That clarity and simplicity could have a particular significance now for evaluating the more opaque machine learning approaches that increasingly dominate this field. While these architectures can be hard to interpret, they often separate the constituent tasks (notably here segmentation from identification) such that their performance can be compared with simpler techniques. [1]

1.1 Prototype Profiles: Comparing Like with Like

At their simplest, prototype profiles consist of a binary separation with (typically) a value of 1 for membership of the collection, and 0 otherwise across the pitch classes (0–11), discounting the differences of octave or enharmonic spelling. For instance, such a 'binary' profile for the *chord* of C-major (CEG) would be given as: [2]

```
[1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0]
```

while the *key* of C-major (CDEFGAB) is represented by:

```
[1, 0, 1, 0, 1, 1, 0, 1, 0, 1, 0, 1]
```


[1] On segmentation as part of chord analysis with machine learning in symbolic data, see especially [8,9].

[2] Note that we speak only of 'representation' here: there is clearly more to both chords and keys that these simple pitch class profiles.

'Rotations' of these profiles cover all the transposition-equivalent sets, in this cases encompassing all major triads and all major keys respectively.

Alternatively, profiles can be constructed from the empirical evidence of either psychological 'goodness of fit' studies or from musical practice, for instance, by taking the pitch class profile (hereafter, PCP) of all events in a dataset ('symbolic' or audio) considered relevant.

These more subtly weighted profiles help to address vexed issues like scale degrees 6 and 7 in the minor mode. They can also help to introduce other, potentially important contextual sensitivity in response to variables that limit the effectiveness of a one-size-fits-all approach to the profile. Literature on symbolic datasets of Western classical music specifically include accounts of: [3]

- **Repertoire.** [11] demonstrates changing PCPs for repertoire created during the historical period in which there is a move from modality to tonality. [4]
- **Specific keys.** [14] demonstrate a small but significant effect of key on the resulting usage profiles, reinforcing the received wisdom that tonal composers do not regard transposition as a neutral change.
- **Partial pieces.** [11, 14, 15] and others demonstrate that drawing prototype profiles from a short passage at the beginning and/or end improves performance.

1.2 The Part versus the Whole

The last of these points is significant for our purposes. Given that common practice tonal music almost always starts and ends unequivocally in the same, main key, it makes perfect sense that these regions would be more tonally-stable, and thus provide a better indicator of what the PCP for within-key music looks like.

And while the benefit may be only slight for identifying the key of an overall piece, we can realistically expect a greater improvement for shorter, partial piece comparisons. This is because the whole practice of profile matching depends on comparing like with like.

In the extreme case, if we used a chord template for key matching or vice versa, we should expect the system to perform worse on average. More realistically, this applies to the vast grey area between chord and key. For instance, what could be a clearer statement of key than a simple V7-I progression? Assuming the two chords have equal length and exactly one instance of each constituent pitch, this equates to an overall PCP in C major of:

```
[1, 0, 1, 0, 1, 1, 0, 2, 0, 0, 0, 1].
```

Several of the main PCPs in the literature would confuse this short passage in C major for one in G major despite the presence of F and lack of F♯: the double weighting of the pitch G is enough to tip the balance. This is not to criticise any specific scholarship or PCP, but simply to illustrate the problem with applying a key-matching profile to a passage that is too short.

In general, to build PCPs that perform well for the task of identifying keys within a piece effectively, we may have most success by creating those profiles from passages that are firmly attested to be in a given key, rather than assumed to be so. The barrier to generating these PCP models in this way is that it requires us to know what the keys are and where they change in the first place, yet that is the problem that we are trying to solve [14, p.532]. One solution is to build up corpora of human annotations in other, relevant ('similar') music, which can then serve as a model for an automatic approach to new cases that do not come with a manual analysis.

1.3 Human analyses for keys and chords

Fortunately, recent years have seen the creation of several human harmonic analysis corpora. [5] Equipped with this data, we can do better than taking entire works or even shorter spans of an arbitrary length as the basis for our PCPs, focussing instead on passages corresponding exactly to these human-defined segmentations, and building up profiles from passages more robustly 'known' to be within-key. [6] Profiles from individual passages can then be combined with other, comparable passages according to the user's priorities (e.g., repertoire, length, key, and position in work) to yield new models for profile matching.

Moreover, these full harmonic analysis datasets enable us to apply the same logic to the equivalent task for chords. While the field of automatic chordal analysis is at least as established as the equivalent for key, repertoire-based PCP models for chord recognition are rarely available, at least for 'classical' music and symbolic data. This is entirely understandable given that chord-level analysis is much more fine-grained than key-only: they take longer to create, and are harder to manage when alignment issues are concerned (e.g., for multiple sources).

In both cases, the human analyses provide both full details of *what* the chords and keys are, but also *when* they change. This allows us to take the *when* information alone, segment the corpora into short segments for each chord or key, and compare the PCP of that passage to a reference profile to assess how reliably the *when* implies the *what*.

This paper undertakes that comparison for the case of both chords and keys, and across several corpora. Further, we consider how dependent this process is on the exact segmentation by introducing systematic errors to see how deleteriously this affects the results. In all cases, we at least start with simple, highly interpretable conditions: ℓ^1-normalisation and the Manhattan comparison metric; simple (e.g., binary) reference profiles; and a clear separation of 'known' information (the *'when'* of segmentation) from the 'tested' part (the *'what'* of identification).

[3] See [10] for a recent overview of the audio literature on these topics that has no overlap with the symbolic papers cited here.

[4] There has also been a notable, recent return to probe-tone psychological study of repertoire difference. See [12] (after [13]) on distinguishing 'classical' and 'rock' styles.

[5] Datasets with full (chord and key) analyses of Western classical music include [16–20].

[6] This does not, of course, account for inter-analyst disagreement. We should avoid the term 'ground truth' for human annotations.

2. CORPORA

For comparison, this study draws together a range of sources across **repertories** (within the symbolic domain) and **data types** (both audio and symbolic representations of one repertoire). Specifically, for the audio-symbolic comparison, we use the **Beethoven sonata first movements** with score and audio data from [21] and analyses originally from [17]. The audio-analysis alignment includes key- but not chord-level section data, so this is currently limited to the key-level study.

The corresponding score-analysis alignment is as detailed at the 'When in Rome' repository, [7] and includes both keys and chord. 'When in Rome' provides a single, consistent, human- and computer-readable format for all publicly-shared, encoded corpora of Roman numeral harmonic analyses of notated works. [8] First reported in [19], the repository continues to grow and currently includes ca.450 analyses of works by 100 composers (unevenly distributed) for a total of ca.100,000 Roman numerals.

The largest of the new datasets within this framework comprises over 150 analyses of 19th-century songs from the **OpenScore Lieder Corpus**. [9] This provides a useful and interesting counterpoint for across-repertoire comparisons, balancing similarity and difference. The songs are from a similar time period to the Beethoven sonata movements (overlapping, though mostly later), for similar forces (the solo piano now joined by a voice part), and generally slightly shorter (but not always).

The Beethoven sonatas and lieder provide the main two symbolic corpora studies here, supplemented in one case by the **Bach Well Tempered Clavier** collection, also from the 'When in Rome' meta-corpus, and as reported in [19].

2.1 Data preparations for across-domain comparison

Despite the self-evident differences between audio and symbolic data, we seek to make the representations as similar and comparable as possible. To that effect, as well as taking a frame-by-frame approach to audio (sampling rate 10Hz), we approach the symbolic data in a comparable way, converting encodings via musicXML to a similar 'slice' representation that encodes a new data point for each change of pitch in the score. [10]

In both cases, given segment timing information, frames and slices within the corresponding segment can be combined into single PCPs and compared (via the same normalisation) to reference profiles. For the symbolic data, we provide the full set of these 'slice' files as well as another set of files recording the pitch-class profiles for each section asserted to be in one key at the 'When in Rome' repository. This makes processing faster and less dependent on external libraries, which in turn makes replication studies more practical.

[7] https://github.com/MarkGotham/When-in-Rome

[8] In addition to [17], this includes [16, 18] and more.

[9] Originally reported in [22]; now released as an MIR dataset in [23].

[10] See [24] (after [25]) for more details and code.

Specifically, for every symbolic source, these files provide the metadata (including title and composer) and record for each key-section at least the:

- **key:** tonic pitch (such as E♭) and mode (major or minor, indicated by case, e.g. 'F♯' versus 'f♯');
- **profile:** the raw (not normalised) PCP of usage;
- **start and end 'offset ':** as measured from the start of the piece in 'quarter note' symbolic values as well as the **'quarter length'** recording the difference;
- **start, end, and length in measures:** the equivalent measurements using symbolic 'measures'.

For the audio sources, we use absolute duration in seconds as the primary measurement of time.

From this point, it is easy and computationally inexpensive to build new model PCPs from the entries relevant to a specific use case. For instance, to assess the best-fit minor key for passages of 20 measures' duration, we may want to build and use a model profile from all minor-key entries of between, say, 15 and 25 measures, ignoring shorter or longer passages, and perhaps also restricting the sample to the composer and / or genre in question. Alternatively, these 'typical' ranges could also be used to inform parameter setting for variable window size.

2.2 Two qualifications: subjectivity and similarity

First, it bears repeating that human analysis datasets – valuable as they are – are naturally and necessarily subjective. While we often see strong agreement for simple cases, analysts differ greatly in their view of more complex passages. Then again, that is exactly the object of this research area. If there were strict, comprehensive rules mapping from a score or audio source to a single, 'correct' analysis, there would be no need for either the corpora or the studies presented by this paper and the wider field.

Second, while it is expedient initially to work with simple, True/False data for the presence/absence of a match between human and computational key choice (as we do here), chords and keys really exist in a relative proximity relation. For more on the *kinds* of 'errors' that are typical, see [26]'s early examination of the tendencis of certain reference profiles and [20, 27] on discrepancies between a computer reading and several manual annotations of local key in Schubert's *Winterreise* song cycle.

3. CHORD-LEVEL SEGMENTATION

We begin with the case of chord identification, a problem operationally defined here as the selection from among 9 distinct chord types in any of the 12 transpositions, making for 108 options in total. In these studies we test chord and key identification from the corresponding segmentation separately, so chords are defined in 'absolute' terms

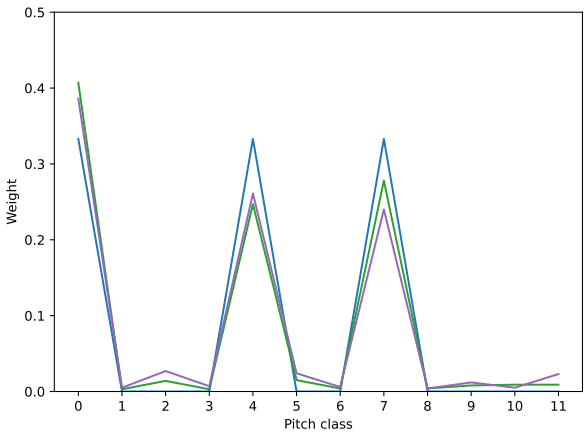

Figure 1. A comparison of ℓ^1-normalised profiles for the major triad with binary values (blue), alongside values from the Lieder (green) and Beethoven (purple) corpora.

(e.g. the *triad* C major) rather than in relation to a key ('tonic' or 'I' in the *key* of C-major). [11]

3.1 Included v.s. excluded chords by type and %

The 9 chords types in question are the four triads (major, minor, diminished, augmented) and five of the most common sevenths (dominant, major, minor, diminished, half-diminished). Around 95% of all chords in the corpora are accounted for by one of these. The remaining ca.5% of cases deemed outside the scope for the current study include: additional seventh types such as sevenths built on augmented triads (relatively rare); all further tertian chords (i.e., 9ths: there are no 11th or 13th chords in the corpora); some chromatic chords like augmented sixths; and detailed entries using RomanText's syntax for supporting missing, added, and altered tones.

To support other approaches to this problem, we provide code for simplifying these chords to what might be considered the 'nearest' corresponding member of the canonical 9 types. For instance, this means removing the 9th of a 9th chord to yield a 7th chord, and 'completing' incomplete triads. All the same, we operationally exclude these cases in the present comparisons as they do not directly reflect the analyst's stated view. The 'N/A' values on Table 1 provide exact numbers of these cases, corpus by corpus.

3.2 Repertoire-specific profiles

Complementing the binary profiles, we create and use a new set of PCPs extracted from the corpora at hand. Given robustly aligned data, the method for this is straightforward: for each chord (in the analysis), identify the triad or seventh type, extract the PCP (from the corresponding range of the score), rotate it to place the chord's root on C (pitch class 0), add each PCP usage value to running totals for the relevant triad or seventh type.

[11] This testing of key and chord separately accounts for most of the relevant considerations, though it is worth noting that the Roman numeral encoding includes other, intermediary information such as 'secondary' key tonicizations.

Repertoire	Matches analysis			Total	% True from
	True	False	N/A		True+False
Binary reference profiles:					
Bach	1149	865	95	2109	57.051
Beethoven	3782	2058	61	5901	64.760
Lieder	8395	2897	505	11797	74.345
Winterreise	1899	660	96	2655	74.209
Profiles from Beethoven:					
Bach	1182	832	95	2109	58.689
Beethoven	3880	1960	61	5901	66.438
Lieder	8187	3105	505	11797	72.503
Winterreise	1874	685	96	2655	73.232
Profiles from Lieder:					
Bach	1237	777	95	2109	61.420
Beethoven	4101	1739	61	5901	70.223
Lieder	8695	2597	505	11797	77.001
Winterreise	2013	546	96	2655	78.664

Table 1. Chord-level segmentation to chord identification, separating values for correct ('True'), incorrect ('False') and out of scope ('N/A').

An alternative strategy would keep separate PCPs without transposition. We decide against that approach here. First, the datasets are not that large, and some chord types (such as augmented triads) are rather rare: this suggests erring on the side of caution, creating fewer distinct chord type PCPs from a greater number of repertoire instances. Second, the lieder dataset is central here, and it is common practice to transpose those songs to a variety of keys depending on the vocal range of the performer. Composers are aware of this, and it stands to reason that key-specific writing is rarer here than in symphonies, say. Finally, and related, it is not obviously better to keep chords separate by absolute triad (e.g., recording a PCP for C major) than by within-key, functional status (e.g., for all tonic major triads). We anticipate further investigation of these areas as the provision of analysis corpora grows and matures.

For illustration, Figure 1 plots the ℓ^1-normalised profiles for the major triad as extracted from the Beethoven and lieder corpora along with the (also normalised) binary profile. Note how the binary profile (blue line) preserves equal weighting of C, E, and G (pitch classes 0, 4, and 7), while the repertoire cases place greater emphasis on the tonic, C, less on E and G, and include some use of the non-chord tone pitch classes.

3.3 Results

Table 1 provides comparative data for this task across the corpora and prototypes. Specifically, we begin with the binary profiles discussed above, applying these to the Bach, Beethoven and Lieder corpora, as well as single collection from within the corpus (Schubert's *Winterreise*) for comparison. We then apply the same method with new chord PCPs extracted from the Beethoven and Lieder corpora.

In all cases, we ℓ^1-normalise both the source PCP and the 108 reference profiles, take the Manhattan distance between the two, and return the top-choice from among those 108 options. Each individual case is a match if and only if the top-choice is the same as that given by the analyst. The

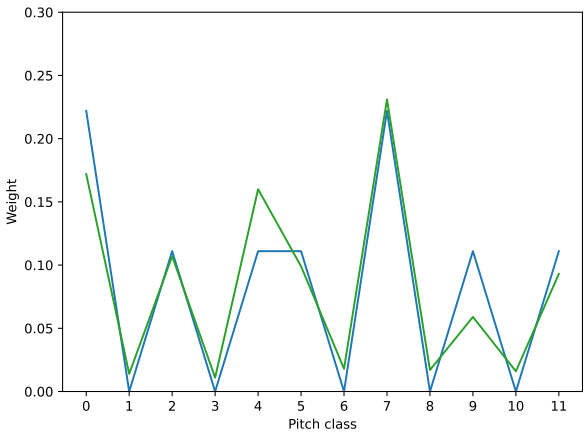

Figure 2. A comparison of ℓ^1-normalised profiles for the major mode with values from C.S. (blue) and Q.W. (green).

final score is given by the percentage of 'correct' responses (computational process matches human judgement) from the total cases (excluding those out of scope).

These results suggest two main observations. First, the success rate of ca.70% across the board is relatively high, considering the simplicity of the algorithm, and the fact that any divergence counts as a failure, including the often slight difference between V and V7. This appears to indicate that the *where* of segmentation is a significant part of the problem: once 'solved', even simple algorithms perform very well in determining the *what* of identification.

Secondly, we should consider the effect of using repertoire-specific PCPs. For these tasks, the lieder PCPs substantially improve results on all corpora, while the Beethoven PCPs are more mixed: they perform better than the binary PCPs on the Bach and Beethoven corpora, but worse for the lieder. The success of the lieder PCPs may speak to the benefit of using PCPs that reflect averages across a more diverse corpora. This has significant ramifications for the field, given that most corpora still pursue a single, focussed repertoire of works by one composer.

4. LOCAL-KEY LEVEL SEGMENTATION

Turning to the equivalent task for local key, there are several, existing PCPs available. Figure 2 illustrates the difference between two contrasting profiles from the literature used here. First, **Craig Sapp (C.S.)** [26] provides deliberately simple (nearly binary) profiles. Specifically, Sapp starts with values of 1 for within-key and 0 for chromatic pitches, but adds 'an additional value of 1' to the tonic and dominant pitch classes (0 and 7) to mitigate 'modal confusion between relative major and (natural) minor keys'.

This profile provides a clear, discrete point of comparison for more fine-tuned, continuous alternatives, notably **Quinn and White (Q.W.)** [14] which distinguishes itself as the only set of profiles to go beyond transposition-equivalence and offer key-specific usage profiles. [12] As

[12] Note that although Figure 2 provides a major-key composite for direct comparison, the present study uses their key-specific profiles.

Reference Profile	Manhattan to human	Euclidean to human	Manhattan to Euclidean
Lieder			
C.S.	74.640	73.583	92.123
A.S.	76.177	74.928	92.315
Q.W.	**77.618**	76.081	93.756
Beethoven (symbolic)			
C.S.	70.529	73.300	87.657
A.S.	79.345	80.353	92.191
Q.W.	**85.39**	82.620	93.199
Beethoven (audio)			
C.S.	51.263	51.136	82.828
A.S.	63.763	61.490	84.722
Q.W.	**68.813**	67.298	87.879

Table 2. The percentage of segments for which the comparison metrics match the human judgement.

part of the 'When in Rome' repository, we provide these along with all published profiles, enabling others to experiment with the full range. As Figure 2 shows, the two distributions differ primarily in their handling of scale degrees 3 and 6 (pitch classes 4 and 9).

For this test, we expand the corpora to include the Beethoven audio, and add a third reference profile from **Albrecht and Shanahan (A.S.)** [15]. We also include an additional comparison between using the ℓ^1-normalisation with Manhattan distance metric, and the ℓ^2-normalisation with the Euclidean distance, comparing each to the human-asserted key and additionally to each other.

Table 2 sets out the results in the form of percentages of segments determined by analysts to be in a single key for which the comparison metrics yield a match, choosing from among the 24 keys (12 major, 12 minor). In all cases, the Q.W. profile, ℓ^1-normalisation and Manhattan distance metric perform best (as highlighted in bold on the Table). The reference profiles are particularly compelling (the normalisation/distance metric paints a more mixed picture). It is perhaps also reassuring that the symbolic and audio corpora follow the same trend.

5. SENSITIVITY TO SEGMENTATION ERROR

While it is useful to know how profile matching performs given 'perfect' segmentation data, for most prospective use cases, we need to know the effect of 'near-misses' in the segmentation. This is important given the subjectivity of human analysis annotation in general, and the fact that segmentation appears to be a particularly variable element.

To that effect, let us return to the Beethoven audio corpus as a test case for considering the effect of segmentation 'error'. For this final test, we introduce systematic segmentation errors across the range of likely 'near misses'. This means varying both the *length* of the segment (making it longer or shorter) as well as the *position* of that modification: adjusting the start of the segment, the end, or both (sharing the length change equally between start and end, centring the new span form on the original range).

The dataset comprises 792 key-segments with a mean duration of 17.8 seconds and a standard deviation of 24.2

Extra length in seconds (total)	Applying adjustment to: the end	the start	both (shared)
-8 sec.	35.859	39.394	35.101
-4 sec.	49.369	51.136	49.116
-2 sec.	58.838	57.323	57.828
-1 sec.	65.404	61.869	61.995
0 sec.	**68.813**	–	–
1 sec.	66.162	68.182	66.793
2 sec.	61.742	62.626	63.889
4 sec.	53.914	49.495	52.146
8 sec.	45.202	35.606	38.510

Table 3. The percentage of segments yielding a match when deliberately diverging from the analyst-defined section length (using Q.W.'s PCPs). The central duration of 0 seconds refers to the 'correct' (analyst defined) length; negative values are shorter; positive are longer.

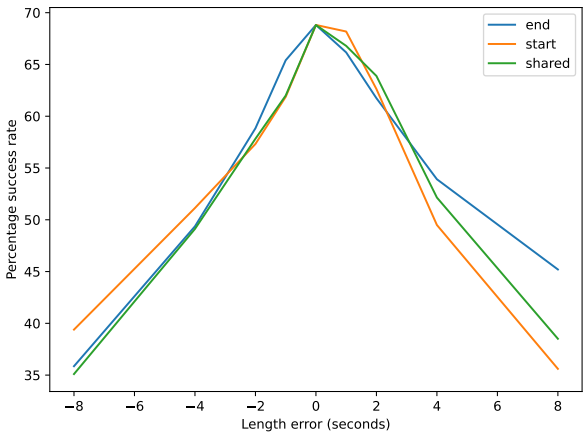

Figure 3. The effect of segmentation errors on the Beethoven audio corpus by error size (x-axis, seconds) and the position of the error (start of segment, end, or shared).

due to a long tail (many segments last more than a minute), and we adjust the length by $\pm$ 1, 2, 4, and 8 seconds per segment. In the original, segment boundaries create contiguous blocks: the end of one segment is the start of the next. As such, increases (positive length errors) mean overlapping segments such that frames originally near the boundary will be considered as part of both the foregoing and following segments. In this scenario, the first and last segments of each movement also extend into the preceding and following silence.

Decreases (negative length error) mean introducing gaps such that there are boundary passages between segments that are not considered at all. In a few cases (of short segments subject to a large change), the segment may be shortened by more than its total length, thus creating a segment with a (musically meaningless) negative duration.

To operationalise an approach to these situations, both the silent *frames* and the negative-duration (non-existent) *segments* return a flat profile with equal weighting for each pitch class. This, in turn, always makes the same choice of best key (an arbitrary one that is usually wrong).

Table 3 and Figure 3 set out the results. Perhaps the most notable outcome here is the asymmetry: shortening a segment is typically more damaging than extending it, particularly in the 'start' condition and the most relevant range of small timing errors. While some of this effect may be due to the handling of negative length described above, that would affect start and end conditions equally and almost never have a bearing on small changes of $\pm$ 1 second. Instead, a start error of $+1$ second leads to a drop in performance of only 0.631%, while the equivalent error of -1 yields more than 10 times the drop: 6.944%.

On the one hand this is surprising. Reducing the segment length means the new passage is still within the range defined by the analyst to be in-key. In this case, we have lost some of the relevant, within-key material, but not added anything from the neighbouring sections in different keys. We might expect this to be barely any more difficult than no change, and certainly less problematic than extending into another segment in a different key.

On the other hand, the moments at which we enter a new key area often announce themselves with material characteristic of the new key such as a dominant seventh chord. According to that view, starting late means missing important material. Viewed this way, it is unsurprising at least that *late starts* adversely affect key recognition.

6. CONCLUSION

This paper has sought to demonstrate both the utility of human analyses for evaluating automatic key- and chord-detection in general and specifically how the very simple information contained therein for *when* chords and keys change can be significant for determining the *what* of full harmonic analyses. We demonstrate this with very simple algorithms that are fully transparent, interpretable, and computationally lightweight.

At a minimum, the results provide important benchmark values for the equivalent task within machine learning architectures that have become popular tools for this field. It may also suggest more efficient work-flows for producing human analyses by separating the tasks which computational processes can perform well from those for which we really need expert annotators.

Having demonstrated the relatively high performance of such simple methods for exact matches, a final section considers the effect of small errors in segmentation. There appears to be an asymmetrical effect of error type by length and position, with late starts being notably damaging for even the shortest adjustments. As these artificial errors emulate a more realistic scenario for many data-driven processes that do not have segmentation information available, this result may have significant implications, for instance in setting window size and tolerance thresholds.

In short, the *what* is highly interrelated with the *when*, at least in the 'idealized' case of full, manual, human harmonic analysis. Segmentation may not be all we need, but it certainly does contribute a great deal, especially relative to the simplicity of the information it encodes.

Acknowledgements: This work was supported by the German Research Foundation (DFG MU 2686/7-2, KL 864/4-2). The International Audio Laboratories Erlangen are a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and Fraunhofer Institut für Integrierte Schaltungen IIS.

7. REFERENCES

[1] C. L. Krumhansl and E. J. Kessler, "Tracing the dynamic changes in perceived tonal organization in a spatial representation of musical keys," *Psychological Review*, vol. 89, no. 4, pp. 334–368, 1982.

[2] C. L. Krumhansl, *Cognitive Foundations of Musical Pitch*. Oxford University Press, 1990, 2001.

[3] C. S. Sapp, "Visual hierarchical key analysis," *Computers in Entertainment (CIE)*, vol. 3, no. 4, pp. 1–19, Oct. 2005. [Online]. Available: http://doi.acm.org/10.1145/1095534.1095544

[4] C. Weiß and J. Habryka, "Chroma-based scale matching for audio tonality analysis," in *Proceedings of the Conference on Interdisciplinary Musicology (CIM)*, Berlin, Germany, 2014, pp. 168–173.

[5] I. Quinn, "Are pitch-class profiles really "key for key"?" *Zeitschrift der Gesellschaft für Musiktheorie [Journal of the German-Speaking Society of Music Theory]*, vol. 7, 01 2010.

[6] L. Feisthauer, L. Bigo, M. Giraud, and F. Levé, "Estimating keys and modulations in musical pieces," in *Proceedings of the 17th Sound and Music Computing Conference*, 2020, pp. 323–330.

[7] N. Nápoles López, C. Arthur, and I. Fujinaga, "Key-finding based on a hidden markov model and key profiles," in *Proceedings of the 6th International Conference on Digital Libraries for Musicology*, ser. DLfM '19. New York, NY, USA: ACM, 2019. [Online]. Available: http://doi.acm.org/10.1145/3358664.3358675

[8] T.-P. Chen and L. Su, "Harmony Transformer: Incorporating Chord Segmentation into Harmony Recognition," in *Proceedings of the 20th International Society for Music Information Retrieval Conference*. Delft, The Netherlands: ISMIR, Nov. 2019, pp. 259–267. [Online]. Available: https://doi.org/10.5281/zenodo.3527794

[9] A. McLeod and M. Rohrmeier, "A modular system for the harmonic analysis of musical scores using a large vocabulary," in *International Society for Music Information Retrieval Conference (ISMIR)*, 2021, forthcoming.

[10] J. Pauwels, K. O'Hanlon, E. Gomez, and M. B. Sandler, "20 years of automatic chord recognition from audio," in *Proceedings of the 20th International Society for Music Information Retrieval Conference*. Delft, The Netherlands: ISMIR, Nov. 2019, pp. 54–63. [Online]. Available: https://doi.org/10.5281/zenodo.3527739

[11] J. D. Albrecht and D. Huron, "A Statistical Approach to Tracing the Historical Development of Major and Minor Pitch Distributions, 1400-1750," *Music Perception*, vol. 31, no. 3, pp. 223–243, 02 2014. [Online]. Available: https://doi.org/10.1525/mp.2014.31.3.223

[12] D. T. Vuvan and B. Hughes, "Probe tone paradigm reveals less differentiated tonal hierarchy in rock music," *Music Perception: An Interdisciplinary Journal*, vol. 38, no. 5, pp. 425–434, 2021.

[13] T. de Clercq and D. Temperley, "A corpus analysis of rock harmony," *Popular Music*, vol. 30, no. 1, pp. 47–70, 001 2011.

[14] I. Quinn and C. W. White, "Corpus-Derived Key Profiles Are Not Transpositionally Equivalent," *Music Perception*, vol. 34, no. 5, pp. 531–540, 06 2017. [Online]. Available: https://doi.org/10.1525/mp.2017.34.5.531

[15] J. Albrecht and D. Shanahan, "The use of large corpora to train a new type of key-finding algorithm," *Music Perception: An Interdisciplinary Journal*, vol. 31, no. 1, pp. 59–67, 2013.

[16] J. Devaney, C. Arthur, N. Condit-Schultz, and K. Nisula, "Theme and variation encodings with roman numerals (TAVERN): A new data set for symbolic music analysis," in *Proceedings of the 16th International Society for Music Information Retrieval Conference, ISMIR 2015, Málaga, Spain, October 26-30, 2015*, 2015, pp. 728–734. [Online]. Available: http://ismir2015.uma.es/articles/261_Paper.pdf

[17] T. Chen and L. Su, "Functional harmony recognition of symbolic music data with multi-task recurrent neural networks," in *Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR 2018, Paris, France, September 23-27, 2018*, E. Gómez, X. Hu, E. Humphrey, and E. Benetos, Eds., 2018, pp. 90–97. [Online]. Available: http://ismir2018.ircam.fr/doc/pdfs/178_Paper.pdf

[18] M. Neuwirth, D. Harasim, F. C. Moss, and M. Rohrmeier, "The Annotated Beethoven Corpus (ABC): A Dataset of Harmonic Analyses of All Beethoven String Quartets," *Frontiers in Digital Humanities*, vol. 5, no. 16, 2018. [Online]. Available: https://www.frontiersin.org/article/10.3389/fdigh.2018.00016

[19] D. Tymoczko, M. Gotham, M. S. Cuthbert, and C. Ariza, "The Romantext format: A flexible and standard method for representing Roman numeral analyses," in *Proceedings of the 20th International*

Society for Music Information Retrieval Conference, ISMIR 2019, Delft, The Netherlands, November 4-8, 2019, A. Flexer, G. Peeters, J. Urbano, and A. Volk, Eds., 2019, pp. 123–129. [Online]. Available: http://archives.ismir.net/ismir2019/paper/000012.pdf

[20] C. Weiß, F. Zalkow, V. Arifi-Müller, M. Müller, H. V. Koops, A. Volk, and H. G. Grohganz, "Schubert Winterreise dataset: A multimodal scenario for music analysis," *Journal on Computing and Cultural Heritage*, vol. 14, no. 2, May 2021. [Online]. Available: https://doi.org/10.1145/3429743

[21] C. Weiß, S. Klauk, M. Gotham, M. Müller, and R. Kleinertz, "Discourse not dualism: An interdisciplinary dialogue on sonata form in Beethoven's early piano sonatas," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Montréal, Canada, 2020, pp. 199–206.

[22] M. Gotham, P. Jonas, B. Bower, W. Bosworth, D. Rootham, and L. VanHandel, "Scores of scores: An OpenScore project to encode and share sheet music," in *Proceedings of the 5th International Conference on Digital Libraries for Musicology*, ser. DLfM '18. New York, NY, USA: ACM, 2018, pp. 87–95. [Online]. Available: http://doi.acm.org/10.1145/3273024.3273026

[23] M. Gotham and P. Jonas, "The OpenScore Lieder Corpus," in *Music Encoding Conference*, ser. MEC '21. MEC, 2021, forthcoming.

[24] M. Gotham, "Moments musicaux," in *6th International Conference on Digital Libraries for Musicology*, ser. DLfM '19. New York, NY, USA: Association for Computing Machinery, 2019, pp. 70–78. [Online]. Available: https://doi.org/10.1145/3358664.3358676

[25] C. W. White and I. Quinn, "The Yale-Classical Archives Corpus," *Empirical Musicology Review*, vol. 11, no. 1, 2016.

[26] C. S. Sapp, "Computational methods for the analysis of musical structure," Ph.D. dissertation, Stanford University, 2011.

[27] C. Weiß, H. Schreiber, and M. Müller, "Local key estimation in music recordings: A case study across songs, versions, and annotators," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 1–1, 2020.