On Local Keys, Modulations, and Tonicizations

A Dataset and Methodology for Evaluating Changes of Key

Néstor Nápoles López*
nestor.napoleslopez@mail.mcgill.ca
McGill University, CIRMMT
Montréal, Canada

Laurent Feisthauer*
laurent.feisthauer@univ-lille.fr
CRIStAL, UMR 9189, CNRS, Université de Lille, Algomus
Lille, France

Florence Levé
florence.leve@u-picardie.fr
Université de Picardie Jules Verne, MIS, Algomus
Amiens, France

Ichiro Fujinaga
ichiro.fujinaga@mcgill.ca
McGill University, CIRMMT
Montréal, Canada

ABSTRACT

Throughout the common-practice period (1650–1900), it is customary to find changes of musical key within a piece of music. In current music theory terminology, the concepts of *modulation* and *tonicization* are helpful to explain many of these changes of key. Conversely, in computational musicology and music information retrieval, the preferred way to denote changes of key are *local key* features, which are oftentimes predicted by computational models. Therefore, the three concepts, local keys, modulations, and tonicizations describe changes of key. What is, however, the relationship between the local keys, modulations, and tonicizations of the same musical fragment?

In this paper, we contribute to this research question by 1) reviewing the current methods of local-key estimation, 2) providing a new dataset with annotated modulations and tonicizations, and 3) applying all the annotations (i.e., local keys, modulations, and tonicizations) in an experiment that connects the three concepts together. In our experiment, instead of assuming the music-theoretical meaning of the local keys predicted by an algorithm, we evaluate whether these coincide better with the modulation or tonicization annotations of the same musical fragment. Three existing models of symbolic local-key estimation, together with the annotated modulations and tonicizations of five music theory textbooks are considered during our evaluation.

We provide the methodology of our experiment and our dataset (available at https://github.com/DDMAL/key_modulation_dataset) to motivate future research in the relationship between local keys, modulations, and tonicizations.

*Both authors contributed equally to this work.

DLfM '20, October 16, 2020, Montréal, QC, Canada

ACM ISBN 978-1-4503-8760-6/20/10…$15.00
https://doi.org/10.1145/3424911.3425515

CCS CONCEPTS

• **Applied computing** → **Sound and music computing**; *Fine arts*; • **Information systems** → **Retrieval tasks and goals**.

KEYWORDS

computational music theory, local key estimation, music information retrieval, roman numeral analysis, tonality, harmony

ACM Reference Format:
Néstor Nápoles López, Laurent Feisthauer, Florence Levé, and Ichiro Fujinaga. 2020. On Local Keys, Modulations, and Tonicizations: A Dataset and Methodology for Evaluating Changes of Key. In *7th International Conference on Digital Libraries for Musicology (DLfM '20), October 16, 2020, Montréal, QC, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3424911.3425515

1 INTRODUCTION

Key identification is a fundamental task in the analysis of tonal music. It is often a preliminary or concurrent step to other common musicological tasks like harmonic analysis and cadence detection. In particular, the knowledge of the musical key can help a music analyst to find boundaries in a musical piece, interpret the role of notes and chords, or suggest a musical form to which the analyzed piece conforms. Due to its importance, key estimation is a well-studied research topic in Music Information Retrieval (MIR), and multiple key-analysis algorithms have emerged during the last decades. Broadly, there are two types of key-estimation algorithms: those that find the *main key* of the piece (often called *global-key-estimation* models in the context of computational musicology and MIR), and those that find the changes of key *within* the piece (often called *local-key-estimation* models). The annotations provided by these models have found applications in music technologies.

In some contexts such as guitar tablatures [18] and electronic dance music [3], global-key annotations are useful as one of the search parameters of a digital music library. Local-key annotations, however, have not yet been used for this purpose. It would be useful to complement key-related searches with local-key annotations, using them to search for musical pieces, based on their underlying changes of key. However, the "interpretability" of local-key annotations requires some attention first.

Changes of key may belong to different categories. In music theory, terms like *modulation* and *tonicization* are helpful for interpreting the context of a change of key. Yet, most local-key-estimation

research omits an investigation of the relationship between local-key annotations and these categories of changes of key. Therefore, as they stand, local-key annotations lack the characteristics that would make them useful in real applications, such as searching for the musical pieces in a large database that showcase similar modulation or tonicization patterns. One could think that these queries would be interesting, and quite different to, for example, searching for pieces of music that share the same global key.

The ideas presented in this paper may be useful to improve the interpretability of the annotations provided by these models.

1.1 Global-Key-Estimation Models

Researchers have designed a number of global-key-estimation algorithms throughout the years. The first one, to our knowledge, is the one by Longuet-Higgins [29] from 1971. Starting from the beginning of a score, the algorithm considers each pitch in order of occurrence and discards the keys that do not include that pitch within their diatonic scale degrees. This process is repeated until only one key remains, and some heuristics are applied afterward to help with the most difficult cases. This algorithm was able to retrieve the key of the fugues in J. S. Bach's *Well-tempered Clavier*. It also served as a reference for later models, such as the one by Vos and van Geenen [47].

With the introduction of the *probe-tone* technique and *key profiles* in 1979 [27], and their later application to the design of a global-key-estimation algorithm in 1990 [25], research regarding global-key-estimation models saw a shift toward more *distributional* approaches [45]. Key profiles, originally introduced as listener ratings by Krumhansl et al. [26, 27], have transitioned into probability distributions that can be used to predict the key of a musical piece. Alternative key-profile distributions—and techniques for applying them—have been proposed over the years.

Key profiles are the basis of many global-key-estimation models for symbolic music files and, starting from the 2000s, audio files as well. More exhaustive surveys of modern global-key-estimation techniques, with a focus on audio, are available [5, 23]. Key profiles have also been useful in the design of local-key estimation models.

1.2 Local Keys, Modulation, and Tonicization

In MIR, it is common to describe algorithms that model *changes of key* as local-key-estimation algorithms. The *local keys* being the predictions that these models generate. Conversely, in music theory, the concepts of *modulation* and *tonicization* are often the manner in which changes of key are explained, and the term *local key* is virtually non-existent.

Therefore, the three concepts, local keys, modulations, and tonicizations describe changes of key. Yet, what is the meaning of these terms? And what is the relationship between the local keys, modulations, and tonicizations of the same musical fragment?

According to the Grove Music Online dictionary, a modulation "refers to a firmly established change of key, as opposed to a passing reference to another key, known as a 'tonicization' " [41]. Moreover, a tonicization is "the act of establishing a new key centre, or of giving a degree other than the first the role of tonic" [16].

A formal definition of local keys is difficult to find. According to Papadopoulos et al. [35], a local key is the "key of each segment" of a "[segmented] musical piece [...] according to the points of modulation".

However, after these definitions, it is still difficult to understand the distinction between modulations and tonicizations. Kostka and Payne have suggested that such distinction is not possible: "The line between modulation and tonicization is not clearly defined in tonal music, nor is it meant to be" [24].

Regarding local keys, most researchers, as Papadopoulos et al. [35], associate them with modulations, however, this relationship has not been explored sufficiently. It would certainly benefit the computational musicology and MIR communities to engage in this exploration, in order to understand what is it that local-key-estimation algorithms predict.

For the scope of this work, we define these terms as follows:

1.2.1 Modulation. Is the change from one key to another. We refer to the initial key as the *departure* key, and the second key as the *destination* key.

1.2.2 Tonicization. Is a brief deviation to a different key, usually with the intention of emphasizing a certain scale degree or harmony. The tonicization often returns to the original key briefly after the deviation.

1.2.3 Local keys. Are the predictions of the musical key provided by a local-key-estimation algorithm. These predictions are given at a finer level of granularity than the entire piece (e.g., notes, onsets, fixed-duration timesteps, audio frames, etc.). In principle, no music-theoretical meaning is inferred from them. They may coincide with modulations or tonicizations.

1.3 Local-Key-Estimation Models

Contrary to the global-key estimation approaches, local-key estimation models have a relatively recent history.

Purwins et al. introduced a method for tracking changes of key in audio signals [37]. Their goal is to track the tone center and its variation during the piece. Their references annotate both modulations and tonicizations but consider that the ground truth is the one indicated by the tonicizations.

Chew [11] measured the distance from a sequence of pitches to a key using the *spiral array* [10]. The succession of keys is then modeled as a sequence of *boundaries* dividing the score in different key areas.

Chai and Vercoe designed a model based on a Hidden Markov Model (HMM) to detect changes of key [7]. They describe the term *modulation* as "the change of key at some point". Their model detects, at first, the tonal center, and then, the mode of the key.

Catteau et al. [6] introduced a model for scale and chord recognition, assuming that there is a correspondence between a major scale and a major key, and between a harmonic minor scale and a minor key. Their model is based on the key profiles by Temperley [44] and Lerdahl's *tonal pitch spaces* [28].

Izmirli introduced a model to find local keys from audio sources, based on non-negative matrix factorization for segmentation [21]. Izmirli also attempted to disambiguate modulations and tonicizations in the following manner: "Secondary functions and tonicizations are heard as short deviations from the well-grounded key

in which they appear—although the boundary between modulation and tonicization is not clear cut. A modulation unambiguously instigates a shift in the key center".

Papadopoulos and Peeters adopted a similar approach to Izmirli for audio local-key estimation [35]. Their model attempts to segment the score based on the points of modulation. They introduced key dependencies on the harmonic and metric structures of global-key-finding methods, in order to convert them into local-key-finding ones.

Rocher et al. introduced a model that provides (chord, key) duples for each audio frame of an input excerpt. The model is based on a graph and the *best-path* estimation method [40]. For evaluating key distances, they used the key profiles by Temperley [44]. The authors alluded to the term modulation when discussing their key predictions.

Mearns et al. used an HMM to estimate modulations over audio transcriptions of Bach chorales [30]. The HMM is trained with chord progressions. The emission probability distributions are obtained from two tables with the probabilities of chords existing in a given key. These tables are based on the work by Schoenberg and Krumhansl. Applied chords (i.e., tonicizations) are not described in these charts, therefore, the authors do not deal with tonicizations.

In 2014, Pauwels and Martens present a probabilistic framework for the simultaneous estimation of chords and keys from audio [36]. They mention the importance of "integrating prior musical knowledge" into a local-key-estimation system, however, they do not allude to the terms modulation and tonicization. The same year, Weiss et al. proposed an audio scale estimator [48]. They argue that this estimator can help to determine the local tonality based on Gárdonyi's scale analysis method. They did not use the term tonicization, however, they discussed "short-time local modulations", which resemble tonicizations.

Machine learning approaches, especially using neural networks, have recently gained popularity in MIR research, including key estimation. Independently, Chen et al. [8, 9] and Micchi et al. [31] designed models that estimate local keys as well as roman numeral analysis annotations. Tonicization information is implied by the roman numeral analysis annotations.

Nápoles López et al. introduced a model to find changes of key (local-key estimation) as well as the main key of a piece (global-key estimation), using an HMM [32]. The model is also capable of working with symbolic and audio data. They do not allude to the terms modulation or tonicization, always referring to their predictions as *local keys*.

One of the most recent models for finding changes of key is by Feisthauer et al. [17], which has been designed to detect modulations in Classical music. It uses three proximity measures established from pitch compatibility, tonality anchoring, and tonality proximity. The model computes the cost of being in a key on a given beat, and estimates the succession of keys using dynamic programming techniques.

1.4 Existing Datasets to Evaluate Local-Key-Estimation Models

Most of the models discussed have been evaluated using different datasets, which are presented in Table 1.

Table 1: Datasets used to evaluate local-key-estimation models.

Model	Files	Dataset
Catteau [6]	10	Manually-built chord sequences
Chai [7]	10	Various (Classical)
Chen [8, 9], Micchi [31]	23	Beethoven
Chew [11]	2	Bach
Feisthauer [17]	38	Mozart (Classical)
Izmirli [21]	17	Pop songs
	152	Naxos set (Classical)
	17	Kostka-Payne (Classical)
Mearns [30]	12	Bach Chorales
Micchi [31]	27	TAVERN (Classical)
	70	ABC (Beethoven)
	72	Roman Text (Classical)
Papadopoulos [35]	5	Mozart
Pauwels [36]	142	SEMA Set (Pop)
	210	MIREX 2009
Purwins [37] and Napolés Lopéz [32]	1	Chopin
Rocher [40]	174	Beatles
Weiss [48]	10	Various (Classical)

The datasets used for evaluating local-key-estimation algorithms are typically small. Additionally, each dataset has often been used to evaluate a single model, which makes the comparison between models somewhat dubious. In this paper, we contribute further discussion around this topic by focusing on the following question: *What is the relationship between the local keys, modulations, and tonicizations of the same musical fragment?* For this purpose, we describe: (1) our methodology for comparing annotations of local keys, modulations, and tonicizations, (2) a dataset that we collected from five music theory textbooks, and (3) an experiment where we evaluated three existing local-key-estimation models.

2 LOCAL KEYS AND MUSIC THEORY

It is challenging to compare, in principle, local key annotations—intrinsically subject to computational considerations (e.g., input constraints or training data)—to modulations and tonicizations, which are rooted in music theory.

In order to achieve a comparison, an initial step involves finding a common representation between local keys, modulations, and tonicizations. For this purpose, we convert an annotated score with the three classes of annotations into a sequence of key labels that share the same level of granularity. There are multiple ways to determine the "slicing" of the musical excerpt or level of granularity. We opted for *onsets*.[1] That is, for every note *onset* in the score, we have a corresponding annotation of the key. This level of granularity is especially convenient for encoding roman numeral analysis annotations, the way in which we encode the modulations and tonicizations in our dataset.

[1]The precise moment when a note (or simultaneously-sounding group of notes) is played [4].

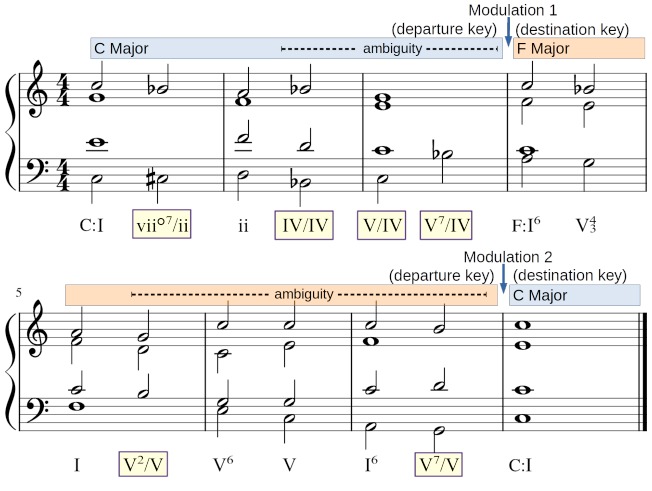


Figure 1: Example 3-17b in Rimsky-Korsakov's *Practical Manual of Harmony* [39]. Roman numeral annotations describe the harmonies on each onset of the score and have been written by the theorist. Some of the roman numeral annotations also indicate tonicizations. These have been framed.

In Figure 1, we introduce a musical score with annotated modulations and tonicizations. Two modulations are observed. The first one, departing from C major and arriving to F major (measure 4). The second one, departing from F major and returning to C major. Between the two modulation events, the *destination* key of one modulation becomes the *departure* key of the next modulation.

Throughout the excerpt, there are also six tonicizations, identified by the presence of at least one forward slash symbol ("/") in the roman numeral annotation.[2] The tonicizations that occur during the first modulation (*departure* key), deviate briefly from C major to D minor (measure 1) and F major (measures 2 and 3); the ones in the second modulation deviate from F major to C major (measures 5 and 7).

Table 2 shows the generated modulation-and-tonicization labels of every onset. There are several conventions for digitally encoding roman numeral annotations [20, 33, 46]. We opted for *harmalysis* [33], an extension of ***harm* [20]. Once the roman numeral annotations have been digitally encoded within the scores, it is fairly simple to retrieve the sequence of key labels.

2.1 Encoding Modulations

We derive the ground-truth keys of the modulation column based on the *departure* keys. For every onset slice before a *destination* key is reached, the *departure* key is written as the ground-truth label of that particular onset. Once a *destination* key is reached, it is considered to be the new *departure* key, and the process repeats.

In Table 2, the departure keys are indicated by a key spelling followed by the token "=>:". Every modulation annotation is considered in that key until a new one is indicated.

[2]The slash is a conventional notation in roman numeral analysis [20, 46]. The roman numeral before the slash symbol denotes a chord, and the roman numeral after the slash symbol denotes a tonicized key.

Table 2: Computer representation of the roman numeral annotations of the excerpt in Figure 1. The modulation and tonicization columns are auto-generated based on the roman numeral annotations. Each row is an onset in the score. The position of the onset is indicated at the first column of the table, as quarter notes from the beginning of the score.

Position	Annotation	Modulation	Tonicization
0	C=>:I	C major	C major
2	viio7/ii	C major	D minor
4	ii	C major	C major
6	IV/IV	C major	F major
8	V/IV	C major	F major
10	V7/IV	C major	F major
12	F=>:I6	F major	F major
14	V43	F major	F major
16	I	F major	F major
18	V2/V	F major	C major
20	V6	F major	F major
22	V	F major	F major
24	I6	F major	F major
26	V7/V	F major	C major
28	C=>:I	C major	C major

2.2 Encoding Tonicizations

We derive the ground-truth keys of the tonicization column based on the keys implied by the roman numeral annotations. When there is no tonicization indication, the key is copied from the modulation column. When the roman numeral implies a different—tonicized—key, the ground-truth label is the key implied by the roman numeral annotation. Using this encoding strategy, we are able to compare local-key predictions to modulation and tonicization annotations.

2.3 Key and Chord Ambiguity

Although we have centered the discussion around changes of key, the line between an analysis of key changes and harmonic analysis may be blurry. This is especially true for tonicizations, which have a shorter temporal scope and often emphasize a scale degree or even a specific harmony. Our decision to encode tonicizations as roman numeral annotations reflects this relationship.

There are datasets available with encoded roman numeral analysis annotations [14, 34, 46], which could be used for studying changes of key in the manner that we have presented here. However, it is important to acknowledge that roman numeral annotations are subject to issues such as ambiguity and disagreement [2, 12, 22, 42], which may have implications for determining *where* the changes of key occur. For example, the dashed regions in Figure 1 show the areas where the key is ambiguous. The exact position of the "arrival" key within an ambiguous zone could, potentially, vary from one analyst to another. This may have implications for the modulation and tonicization annotations.

In this work, we have tried to reduce the implications related to the complexity of harmonic analysis by utilizing a collection of

scores that have been written or displayed specifically to demonstrate modulations, mostly in the manner of instructional or "cherry-picked" examples by the authors of five music theory textbooks.

Most of the examples in the dataset are: (1) very short (4–8 measures), (2) including at least one modulation, and (3) often accompanied by text explanations written by the theorists, which describe the modulation thoroughly.

We consider that these additional characteristics make this dataset slightly more robust to the issues related to disagreement and ambiguity. Therefore, they could be more suitable for studying changes of key than existing roman numeral analysis datasets.

3 DATASET

All the labels in the dataset have been obtained from the modulation excerpts of five music theory textbooks, written by different music theorists and/or composers (whom we simply refer to as "theorists" for the rest of this paper).

The dataset contains, in total, 201 excerpts of music with annotated modulations and tonicizations. The annotations are encoded in the form of roman numerals of all the chords in the dataset, which can be helpful for utilizing the dataset for other purposes beside the one presented here (e.g., chord labeling or cadence detection). Each file has been encoded in Humdrum (**kern) symbolic music representations [19]. As mentioned, the roman numeral annotations have been digitally encoded [33] within the scores.

When the theorists provided roman numeral annotations, those have been preserved in our digital transcriptions. Otherwise, we furnished them. All the annotations related to modulations have been obtained exclusively from the textbooks. Tonicizations rely on the roman numeral annotations of the chords and these were not always provided in the textbooks, therefore, we supplied some of them.

The issues of ambiguity discussed in Section 2.3 might have implications mostly for tonicizations (the ones we sometimes contributed ourselves). However, modulations have always been provided by the theorists, and we expect this to reduce the impact of these issues. For some onsets, multiple key annotations were provided by the theorists. For these excerpts, we decided to encode the keys in chronological order. An example is shown in Figure 2.[3]

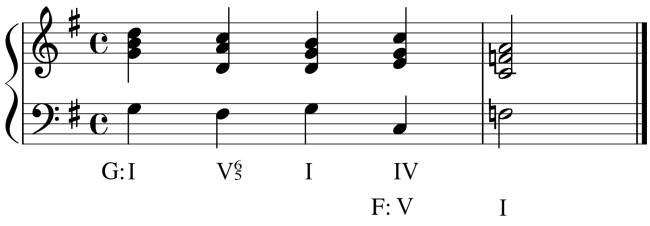


Figure 2: Example 18-3 in Kostka-Payne's Tonal Harmony [24]. Two concurrent keys (G Major and F Major) are annotated in the fourth beat of measure 1. We considered the key label of this onset to be "F Major" for the modulation column.

[3]Whenever a new key was established according to the theorists' annotations, we considered that to be the only key in the modulation column, until a new one appeared to replace it. We applied this process systematically throughout the dataset.

We describe the five textbooks and their abbreviations, which we use throughout the experiment section.

3.1 Sources of the Annotations

3.1.1 Aldwell, Schachter, and Cadwallader (ASC) [USA, contemporary]. The modulation excerpts are taken from chapter 27 *Diatonic Modulations* of the *Harmony and Voice Leading* [1]. This textbook provided seven excerpts to the dataset, the smallest amount among all textbooks. These excerpts are extracts from Bach Chorales (4), Mozart's Trio for Clarinet (1), and two original examples.

3.1.2 Kostka and Payne (KP) [USA, contemporary]. The modulation excerpts are taken from the 18th and 19th chapters of *Tonal Harmony* [24]. We took fifteen excerpts from this book, which are fragments of pieces written by Classical and Romantic composers. Previously, the annotated audio excerpts of the accompanying workbook were used for another local-key estimation study [21], however, we encoded the score excerpts of the main textbook.

3.1.3 Reger (Reg) [Germany, 1904]. A hundred modulation excerpts are taken from *On the Theory of Modulation*[4] [38]. The excerpts are very short and they are all written by Reger himself. Reger's goal was to provide cadence-like examples of modulation from two keys (C major and A minor) to almost every other possible key.

Seventeen of the examples had two terminations: one in a major key and one in a minor key.[5] We separated these examples into two files, one for each of the terminations. This increased the total number of examples from 100 to 117.

3.1.4 Rimsky-Korsakov (Rim) [Russia, 1886]. The modulation excerpts are taken from the third section of the *Practical Manual of Harmony* [39]. As with Reger, all the thirty-seven examples in this textbook are written by the author himself. Some of the examples, however, are more detailed and longer in duration than the ones by Reger.

3.1.5 Tchaikovsky (Tch) [Russia, 1872]. The modulation examples considered are taken from the third section of the *Guide to the Practical Study of Harmony* [43]. All twenty-five examples were written by Tchaikovsky himself.

3.2 Statistics About the Dataset

Some statistics about the dataset are presented in Table 3. We report, for each of the textbooks: the number of files (excerpts), the number of modulations, the number of tonicizations, and the number of labels (which is equivalent to the number of onsets, as we supplied one label per onset).

The *Reg* textbook is by far the one that contributed the largest number of excerpts. However, the ones providing a higher ratio of labels per number of files are *ASC* (26.42) and *KP* (36.93). This may be due to the use of musical examples taken from the literature, where modulations occur within a musical context and, therefore, span longer regions.

Rim and *Tch* are the textbooks that provided the highest number of tonicizations. They show tonicizations in 41.63% and 15.97% of

[4]The book we used is the republication by *Dover* with the title *Modulation* (2007).
[5]Usually parallel keys that shared a closing dominant harmony and resolved to either the major or minor mode, with the same tonic.

Table 3: Summary of the dataset. Each value indicates the number of occurrences in the corresponding textbook.

Sample	Files	Modulations	Tonicizations	Labels
ASC	7	8	7	185
KP	15	21	11	554
Reg	117	220	40	768
Rim	37	44	107	257
Tch	25	60	38	238
Total	201	555	203	2002

the onsets, respectively. In terms of investigating the relationship between predicted local keys and modulations/tonicizations, these textbooks contributed the most interesting examples.

Rim and *Tch* also tend to set the annotations of the *destination* key of a modulation in the tonic degree, considering any preceding dominant chords as secondary dominants and, consequently, part of the *departure* key (as shown on Figure 1). Other theorists, on the other hand, often set the *destination* key already in the dominant chords that precede the tonic. Therefore, they do not annotate (or imply)[6] a tonicization for the preceding dominant chords.

4 EXPERIMENT

In our experiment, we investigate whether the predictions of three local-key-estimation computational models coincide with the modulation and tonicization annotations of the music theory textbooks.

4.1 Evaluation Procedure

Even if two predicted keys do not match the ground-truth label, one of the predictions may still be better than the other, due to the *close* or *far* relationship that a predicted key may have to the ground truth. For this reason, in addition to accuracy, we propose to also use a weighted score to evaluate each onset's key.

Table 4 shows the two sets of weights we utilized to evaluate the key predictions, based on the relationship that the predicted key has to the ground truth. The MIREX score has been used in the annual Music Information Retrieval Evaluation eXchange (MIREX) evaluation of global-key-estimation algorithms since 2005 [15].

Table 4: Evaluation weights for the key predictions.

Key Relationship (Reference, Predicted)	Accuracy	MIREX
Same key	1.0	1.0
Dominant / SubDominant	0.0	0.5
Relative Major / Relative Minor	0.0	0.3
Parallel Major / Parallel Minor	0.0	0.2
Other	0.0	0.0

We apply both evaluations to the key of every onset in the score. The annotations are evaluated according to Equation 1.

[6]For the cases in which we provided the roman numeral annotations because they were not provided by the theorists.

$$score = \frac{\sum_{i=0}^{N} w(k_i, l_i) d_i}{\sum_{i=0}^{N} d_i} \quad (1)$$

N represents the number of onsets in the input score, k is a vector of ground-truth key annotations for each onset, l is the vector of local-key predictions provided by the model for each onset, and d is a vector with the durations in quarter notes of every onset. Note that the vector k corresponds to either the labels in the *Modulation* or *Tonicization* columns of Table 2, depending on the task.

The w function is a piece-wise function that evaluates either of the weighted scores shown in Table 4, given a ground-truth key and a prediction. The scalar value *score* is in the range $[0, 1]$. A value of 1.0 is obtained if and only if the model predicts the key correctly at every onset. A value of 0.0 is obtained if and only if the model makes incorrect predictions (and for the MIREX weights, also without any partial value) at each onset.

Using this methodology, we evaluate four baseline models and three local-key-estimation models from the literature. In total, we perform four evaluations for each model:

(1) Modulation ($k = Modulation$) by accuracy ($w = Accuracy$).
(2) Tonicization ($k = Tonicization$) by accuracy ($w = Accuracy$).
(3) Modulation ($k = Modulation$) by MIREX ($w = MIREX$).
(4) Tonicization ($k = Tonicization$) by MIREX ($w = MIREX$).

These evaluations coincide with the results shown in Figure 3.

4.2 Baseline Models

We describe two baseline models (*B1* and *B2*) that—we expect—will perform worse than existing local-key-estimation algorithms. Similarly, we propose two theoretical "models" that set the maximum performance that can be achieved when designing a modulation or tonicization model (*B3* and *B4*). These artificial models consist simply of the ground truth annotations for modulations and tonicizations, but evaluated in the opposite task (i.e., as if they were local-key predictions coming from a model). We expect that these baselines will set a reasonable frame for inspecting the performance of the *real* local-key-estimation models.

4.2.1 Random guess (B1). This model randomly chooses a key label for every onset in the piece.[7] We expect this model (*B1*) to be the worst-performing model in our experiment.

4.2.2 Global-key guess (B2). Given the large body of work that exists in global-key-estimation algorithms, it would be reasonable to assume that using the predictions of a global-key-estimation model in every onset would deliver reasonable results. We incorporate these global-key predictions as a baseline model, to compare it against the more specialized local-key-estimation models. The global-key-estimation model that we considered is the default key-estimation model in music21 [13].

4.2.3 Modulation (B3) and tonicization (B4) ground truth. Due to the overlap that exists between the modulation and tonicization key

[7] The key label is generated by choosing randomly between 24 possible keys ([0–23]), collapsing all enharmonic spellings into the same class. This may occlude important hints given in the note spellings of music scores, however it guarantees that this method can be applied to MIDI and audio files, which lack pitch-spelling information.

labels,[8] it is expected that a good-performing modulation model would also achieve a good performance in predicting tonicizations (and vice versa). In order to observe this, we consider two additional models: the ground truth of modulation employed as a "model" that predicts tonicization (*B3*), and the ground truth of tonicization employed as a "model" that predicts modulation (*B4*). For simplicity, we refer to these as baseline models, although they represent ground-truth annotations and not computational models per se.

We evaluate the performance that each of these theoretical models, *B3* and *B4*, achieves in its own task and the opposite one. More specifically, we expect the following: (1) these models should obtain a perfect score in their own task (no matter which weights are utilized) and (2) both models should obtain an identical performance when evaluated in the opposite task.[9]

4.3 Local-Key Models from the Literature

Three recent models of local-key estimation are evaluated. As part of the results in this paper, we intend to investigate whether these three models are better suited for finding "modulations" or "tonicizations". We consider that this methodology could equally be applied to other symbolic and audio local-key-estimation models.

4.3.1 Nápoles López et al. (M1). This model computes two stages of output: local keys and a global key [32]. Both stages are computed through an HMM. The main parameters of the HMM are a set of *key profiles* and a table of key distances. We compute both stages but only evaluate the local-key predictions. The model is applied with its default parameters.

4.3.2 Feisthauer et al. (M2). This model estimates the succession of keys and can be configured by adjusting the weights associated to three proximity measures [17]. Two sets of weights are used. The first set consists of the default weights when the model is not trained (M2a). The second set consists of the optimal weights after training the model on Mozart's string quartets (M2b).

4.3.3 Micchi et al. (M3). This model was introduced by Micchi et al. [31] and utilizes an LSTM network to provide the harmonic analysis of a musical excerpt, which is described through six features: key, degree 1, degree 2, quality, inversion, and root. We evaluate the key predictions of the model when it is used without any post-processing. A web application is also provided by Micchi et al. to facilitate the process of generating these or other annotations.[10]

All of the local-key-estimation models have been trained on different datasets (see Table 1) and have been used with their default parameters. The dataset of modulations and tonicizations introduced in Section 3 has been utilized only as a *test* set to these models. It is likely that the models would achieve better results if they were trained on a sample of the dataset. However, the current size of the dataset is not sufficient to provide training, validation, and test splits. Therefore, we decided to limit its use as a test set. With this, we investigate the generalization capabilities of these pre-trained models.

[8] This is because we duplicate the modulation label in the tonicization column unless there is a tonicization (see Table 2).

[9] This is also true for the MIREX weights, because the evaluation is symmetrical. That is, $w(gt, pred) = w(pred, gt)$; when $w = MIREX$

[10] http://roman.algomus.fr/

5 RESULTS AND DISCUSSION

Figure 3 shows the four evaluations of the baseline and local-key-estimation models. The MIREX scores are generally higher because they reward a key prediction that is *nearby* the correct class. This has increased the results of virtually all models. However, the global-key baseline model (*B2*) has benefited the most from the MIREX evaluation, followed by *M2b*. This suggests that these models predict keys that are nearby the ground truth, while other models tend to "hit-or-miss".

For the reverse ground truth "models", *B3* and *B4*, the results are symmetric, as expected. When they are used in the opposite task, they have a quasi-perfect evaluation on *ASC*, *KP* and *Reg*. Incidentally, this shows the relatively low number of tonicizations in these textbooks (see Section 2.2). However, in *Tch* and especially *Rim*, where there is a heavier use of tonicizations, the local-key-estimation models show an inclination toward the tonicization predictions, rather than the modulation ones. This is unexpected, as most researchers do not describe their local-key-estimation models as "tonicization finders".

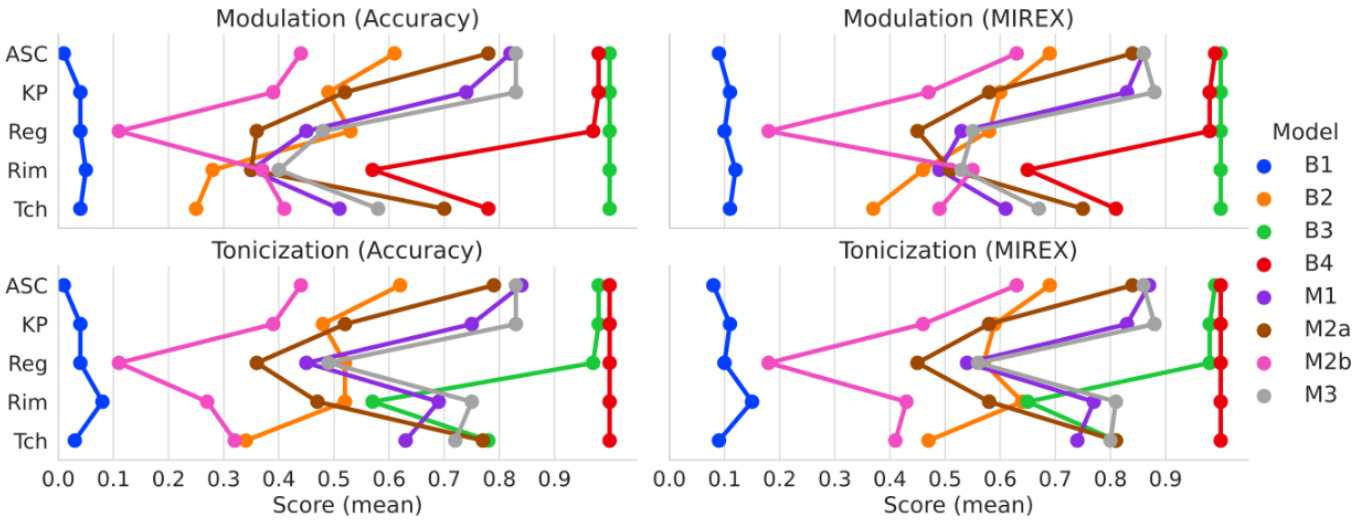
As expected, the random-guess baseline (*B1*) is the worst performing model in all evaluations. It also highlights the success of the global-key-estimation model (*B2*) compared to a random guess. This model (*B2*) achieves good performance overall; in the *Reg* examples, it even achieves better performance than all the specialized local-key-estimation models (*M1*, *M2a-b*, and *M3*).

The Nápoles López et al. model *(M1)* achieves a good performance overall, and does slightly better in predicting tonicizations than modulations. This is at least the case for *Tch*, and more evidently, for *Rim*.

The proximity measure model *M2a* gets a lower average score than *M1* and *M3*, except in the *Tch* textbook. It also performs better on the tonicization task than on the modulation one. The variant of this model, *M2b*, gets, on average, the worst results of all the models (excluding *B1*). It also seems to be the only model doing better in predicting modulations than tonicizations. The reason for this could be that the model was trained with complete musical pieces; the modulations in those pieces span several measures and, therefore, differ notably from the short excerpts used in this dataset.

M3 results are slightly better than those obtained by *M1*. It is also interesting that these two models (*M1* and *M3*) show a similar performance "shape" across all of the dataset (see Figure 3), despite their disparities in design and methodology. For example, both do poorly predicting the modulation ground truth of *Rim*, and better in predicting the tonicization ground truth. This is contrasting to the performance of, for example, *M2b*, which did better for modulation than it did for tonicization in *Rim*, and shows a different performance "shape" to these models (*M1* and *M3*).

Lastly, even within this collection of "cherry-picked" examples of modulation, the diversity in the annotations by different theorists is perceivable. Some, such as *Rim* and *Tch* make heavier use of tonicizations in their annotations, while *ASC*, *KP*, and *Reg* do not. This might be related to the issues of ambiguity and disagreement mentioned in Section 2.3. Our evaluation methodology is not doing a great deal in compensating for these issues. Future revisions of the methods for evaluating local-key-estimation algorithms and more data could certainly be of help toward addressing these problems.



Model	Task	ASC	KP	Reg	Rim	Tch
B1	Mod	0.05	0.03	0.02	0.04	0.03
	Ton	0.05	0.03	0.02	0.07	0.05
B2	Mod	0.61	0.49	**0.53**	0.28	0.25
	Ton	0.62	0.48	**0.52**	0.52	0.34
M1	Mod	0.82	0.74	0.45	0.35	0.51
	Ton	**0.84**	0.75	0.45	0.69	0.63
M2a	Mod	0.78	0.52	0.36	0.35	**0.70**
	Ton	0.79	0.52	0.36	0.47	**0.77**
M2b	Mod	0.44	0.39	0.11	0.37	0.41
	Ton	0.44	0.39	0.11	0.27	0.32
M3	Mod	**0.83**	**0.83**	0.48	**0.40**	0.58
	Ton	0.83	**0.83**	0.49	**0.75**	0.72
B3	Mod	1.00	1.00	1.00	1.00	1.00
	Ton	0.98	0.98	0.97	0.57	0.78
B4	Mod	0.98	0.98	0.97	0.57	0.78
	Ton	1.00	1.00	1.00	1.00	1.00

Model	Task	ASC	KP	Reg	Rim	Tch
B1	Mod	0.09	0.10	0.08	0.11	0.10
	Ton	0.10	0.10	0.08	0.13	0.13
B2	Mod	0.69	0.60	**0.58**	0.46	0.37
	Ton	0.69	0.59	**0.57**	0.64	0.47
M1	Mod	**0.86**	0.83	0.53	0.49	0.61
	Ton	**0.87**	0.83	0.54	0.77	0.74
M2a	Mod	0.84	0.58	0.45	0.51	**0.75**
	Ton	0.84	0.58	0.45	0.58	**0.81**
M2b	Mod	0.63	0.47	0.18	**0.55**	0.49
	Ton	0.63	0.46	0.18	0.43	0.41
M3	Mod	**0.86**	**0.88**	0.55	0.53	0.67
	Ton	0.86	**0.88**	0.56	**0.81**	0.80
B3	Mod	1.00	1.00	1.00	1.00	1.00
	Ton	0.99	0.98	0.98	0.65	0.81
B4	Mod	0.99	0.98	0.98	0.65	0.81
	Ton	1.00	1.00	1.00	1.00	1.00

Figure 3: Evaluation scores for each model on each textbook of our dataset. The models predict modulations and tonicizations. Furthermore, they are evaluated using accuracy and MIREX weights, as described in Section 4.1. A plot is shown for each of the four evaluations. The y axis shows the evaluations on different textbooks of our dataset, as the performance varies from one to another. The x axis shows the mean score obtained by a model across all the files in the textbook. Bold scores indicate the best-performing model in a given textbook and task, excluding *B3* and *B4* (see Section 4.2.3).

6 CONCLUSION

In this paper, we discussed the need to further investigate the notion of a *local key*, common in computational musicology and music information retrieval, and its relationship to the music-theoretical concepts of *modulation* and *tonicization*. We provided a small dataset of modulations and tonicizations that we collected from five music theory textbooks. With this dataset and a proposed methodology, we evaluated four baseline models and three local-key-estimation algorithms from the literature. We consider that this methodology could be applied to other algorithms in the symbolic and audio domain, and may contribute to overcome some of the semantic gaps between the terminology of MIR research and music theory. The dataset introduced in Section 3 has been made publicly available under a Creative Commons Attribution 4.0 International License at the following location: https://github.com/DDMAL/key_modulation_dataset.

ACKNOWLEDGMENTS

This research has been supported by the Social Sciences and Humanities Research Council of Canada (SSHRC) and the Fonds de recherche du Québec–Société et culture (FRQSC).

We would like to thank Micchi et al. for facilitating us with the predictions of their algorithm [31] using our dataset. Similarly, we would like to thank the anonymous reviewers of a previous version of this manuscript, who provided precious feedback for this version.

REFERENCES

[1] Edward Aldwell, Carl Schachter, and Allen Cadwallader. 2019. *Harmony and Voice Leading*. Cengage Learning. https://books.google.fr/books?id=T69EDwAAQBAJ

[2] Claire Arthur. 2017. Taking Harmony Into Account: The Effect of Harmony on Melodic Probability. *Music Perception* 34, 4 (04 2017), 405–423. https://doi.org/10.1525/mp.2017.34.4.405

[3] Beatport. 2020. *Tracks :: Beatport*. Retrieved September 7, 2020 from https://www.beatport.com/tracks/all

[4] Juan Bello, Laurent Daudet, Samer Abdallah, Chris Duxbury, Mike Davies, and Mark Sandler. 2005. A Tutorial on Onset Detection in Music Signals. *IEEE Transactions on Speech and Audio Processing* 13 (2005), 1035–1047. https://doi.org/10.1109/TSA.2005.851998

[5] Spencer Campbell. 2010. *Automatic Key Detection of Musical Excerpts from Audio*. Master's thesis. McGill University, Montréal, QC.

[6] Benoit Catteau, Jean-Pierre Martens, and Marc Leman. 2007. A Probabilistic Framework for Audio-Based Tonal Key and Chord Recognition. In *Advances in Data Analysis*, Reinhold Decker and Hans J. Lenz (Eds.). Springer, Berlin, Heidelberg, 637–644.

[7] Wei Chai and Barry Vercoe. 2005. Detection of Key Change in Classical Piano Music. In *Proceedings of the 6th International Conference on Music Information Retrieval*, Bernard Meltzer and Donald Michie (Eds.). ISMIR, London, UK, 468–473. https://doi.org/10.5281/zenodo.1415538

[8] Tsung-Ping Chen and Li Su. 2018. 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, Paris, France, 90–97. https://doi.org/10.5281/zenodo.1492351

[9] Tsung-Ping Chen and Li Su. 2019. Harmony Transformer: Incorporating Chord Segmentation into Harmony Recognition. In *Proceedings of the 20th International Society for Music Information Retrieval Conference*. ISMIR, Delft, The Netherlands, 259–267. https://doi.org/10.5281/zenodo.3527794

[10] Elaine Chew. 2000. *Towards a Mathematical Model of Tonality*. Ph.D. Dissertation. MIT, Cambridge, MA.

[11] Elaine Chew. 2002. The Spiral Array: An Algorithm for Determining Key Boundaries. In *Music and Artificial Intelligence*. Springer, Berlin, Heidelberg, 18–31.

[12] Nathaniel Condit-Schultz, Yaolong Ju, and Ichiro Fujinaga. 2018. A Flexible Approach to Automated Harmonic Analysis: Multiple Annotations of Chorales by Bach and Prætorius. In *Proceedings of the 19th International Society for Music Information Retrieval Conference*. ISMIR, Paris, France, 66–73. https://doi.org/10.5281/zenodo.1492345

[13] Michael Scott Cuthbert and Christopher Ariza. 2010. Music21: A Toolkit for Computer-Aided Musicology and Symbolic Music Data. In *Proceedings of the 11th International Society for Music Information Retrieval Conference*. ISMIR, Utrecht, Netherlands, 637–642. https://doi.org/10.5281/zenodo.1416114

[14] Johanna Devaney, Claire Arthur, Nathaniel Condit-Schultz, and Kirsten Nisula. 2015. 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, Málaga, Spain, 728–734. https://doi.org/10.5281/zenodo.1417497

[15] J. Stephen Downie, Kris West, Andreas F. Ehmann, and Emmanuel Vincent. 2005. The 2005 Music Information Retrieval Evaluation Exchange (MIREX 2005): Preliminary Overview. In *Proceedings of the 6th International Conference on Music Information Retrieval*. ISMIR, London, UK, 320–323. https://doi.org/10.5281/zenodo.1416044

[16] William Drabkin. 2001. Tonicization. Grove Music Online. https://doi.org/10.1093/gmo/9781561592630.article.28123

[17] Laurent Feisthauer, Louis Bigo, Mathieu Giraud, and Florence Levé. 2020. Estimating Keys and Modulations in Musical Pieces. In *Proceedings of the 17th Sound and Music Computing Conference*. Simone Spagnol and Andrea Valle, Torino, Italy. https://hal.archives-ouvertes.fr/hal-02886399

[18] Ultimate Guitar. 2020. *Explore tabs @ Ultimate-guitar.com*. Retrieved September 7, 2020 from https://www.ultimate-guitar.com/explore

[19] David Huron. 2002. Music Information Processing Using the Humdrum Toolkit: Concepts, Examples, and Lessons. *Computer Music Journal* 26, 2 (2002), 11–26.

[20] David Huron. 2020. *Representation: **harm*. Retrieved August 7, 2020 from https://www.humdrum.org/rep/harm/

[21] Özgür Izmirli. 2007. Localized Key Finding from Audio Using Nonnegative Matrix Factorization for Segmentation. In *Proceedings of the 8th International Conference on Music Information Retrieval*. ISMIR, Vienna, Austria, 195–200. https://doi.org/10.5281/zenodo.1417197

[22] Hendrik Vincent Koops, Bas de Haas, John Ashley Burgoyne, Jeroen Bransen, Anna Kent-Muller, and Anja Volk. 2019. Annotator Subjectivity in Harmony Annotations of Popular Music. *Journal of New Music Research* 48, 3 (2019), 232–252. https://doi.org/10.1080/09298215.2019.1613436

[23] Filip Korzeniowski. 2018. *Harmonic Analysis of Musical Audio using Deep Neural Networks*. Ph.D. Dissertation. Johannes Kepler University Linz, Linz, Austria.

[24] Stefan Kostka and Dorothy Payne. 2008. *Tonal Harmony*. McGraw-Hill Education, Boston.

[25] Carol L. Krumhansl. 1990. *Cognitive Foundations of Musical Pitch*. Oxford University Press, USA.

[26] Carol L. Krumhansl and Edward J. Kessler. 1982. Tracing the Dynamic Changes in Perceived Tonal Organization in a Spatial Representation of Musical Keys. *Psychological Review* 89, 4 (1982), 334–368. https://doi.org/10.1037/0033-295X.89.4.334

[27] Carol L. Krumhansl and Roger N. Shepard. 1979. Quantification of the Hierarchy of Tonal Functions Within a Diatonic Context. *Journal of Experimental Psychology: Human Perception and Performance* 5, 4 (1979), 579–594. https://doi.org/10.1037/0096-1523.5.4.579

[28] Fred Lerdahl. 1988. Tonal Pitch Space. *Music Perception* 5 (January 1988), 315–350.

[29] Hugh C. Longuet-Higgins and Mark Steedman. 1971. *Roman: "On Interpreting Bach"*. Edinburgh University Press, 221–241.

[30] Lesley Mearns, Emmanouil Benetos, and Simon Dixon. 2011. Automatically Detecting Key Modulations in J. S. Bach Chorale Recordings. In *Proceedings of the 8th Sound and Music Computing Conference*. 25–32.

[31] Gianluca Micchi, Mark Gotham, and Mathieu Giraud. 2020. Not All Roads Lead to Rome: Pitch Representation and Model Architecture for Automatic Harmonic Analysis. *Transactions of the International Society for Music Information Retrieval* 3, 1 (May 2020), 42–54. https://doi.org/10.5334/tismir.45

[32] Néstor Nápoles López, Claire Arthur, and Ichiro Fujinaga. 2019. Key-Finding Based on a Hidden Markov Model and Key Profiles. In *Proceedings of the 6th International Conference on Digital Libraries for Musicology*. ACM, The Hague, Netherlands, 33–37. https://doi.org/10.1145/3358664.3358675

[33] Néstor Nápoles López and Ichiro Fujinaga. 2020. Harmalysis: A Language for the Annotation of Roman Numerals in Symbolic Music Representations. In *Proceedings of the Music Encoding Conference*. MEC, Boston, MA, 83–85. http://dx.doi.org/10.17613/380x-dd98

[34] Markus Neuwirth, Daniel Harasim, Fabian C. Moss, and Martin Rohrmeier. 2018. The Annotated Beethoven Corpus (ABC): A Dataset of Harmonic Analyses of All Beethoven String Quartets. *Frontiers in Digital Humanities* 5 (July 2018). https://doi.org/10.3389/fdigh.2018.00016

[35] Hélène Papadopoulos and Geoffroy Peeters. 2009. Local Key Estimation Based on Harmonic and Metric Structures. In *Proceedings of the 12th International Conference on Digital Audio Effects (DAFx-09)*. Como, Italy, 408–415. https://hal.archives-ouvertes.fr/hal-00511452

[36] Johan Pauwels and Jean-Pierre Martens. 2014. Combining Musicological Knowledge About Chords and Keys in a Simultaneous Chord and Local Key Estimation System. *Journal of New Music Research* 43, 3 (July 2014), 318–330. https://doi.org/10.1080/09298215.2014.917684

[37] Hendrik Purwins, Benjamin Blankertz, and Klaus Obermayer. 2000. A New Method for Tracking Modulations in Tonal Music in Audio Data Format. In *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium*, Vol. 6. 270–275. https://doi.org/10.1109/IJCNN.2000.859408

[38] Max Reger. 1904. *Supplement to the Theory of Modulation*. C. F. Kahnt Nachfolger, Leipzig.

[39] Nikolay Rimski-Korsakov. 1886. *Practical Manual of Harmony*. A. Büttner, St. Petersburg.

[40] Thomas Rocher, Matthias Robine, Pierre Hanna, and Laurent Oudre. 2010. Concurrent Estimation of Chords and Keys from Audio. In *Proceedings of the 11th International Society for Music Information Retrieval Conference*. ISMIR, Utrecht, Netherlands, 141–146. https://doi.org/10.5281/zenodo.1417485

[41] Janna Saslaw. 2001. *Modulation (i). Grove Music Online*. https://doi.org/10.1093/gmo/9781561592630.article.18843

[42] Anna Selway, Hendrik Vincent Koops, Anja Volk, David Bretherton, Nicholas Gibbins, and Richard Polfreman. 2020. Explaining Harmonic Inter-Annotator Disagreement Using Hugo Riemann's Theory of 'Harmonic Function'. *Journal of New Music Research* 49, 2 (January 2020), 136–150. https://doi.org/10.1080/09298215.2020.1716811

[43] Peter I. Tchaikovsky. 1872. *Guide to the Practical Study of Harmony*. P. Jurgenson, Moscow.

[44] David Temperley. 1999. What's Key for Key? The Krumhansl-Schmuckler Key-Finding Algorithm Reconsidered. *Music Perception* 17, 1 (October 1999), 65–100.

[45] David Temperley and Elizabeth West Marvin. 2008. Pitch-Class Distribution and the Identification of Key. *Music Perception: An Interdisciplinary Journal* 25, 3 (2008), 193–212. https://doi.org/10.1525/mp.2008.25.3.193

[46] Dmitri Tymoczko, Mark Gotham, Michael Scott Cuthbert, and Christopher Ariza. 2019. 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*. Delft, Netherlands, 123–129.

[47] Piet G. Vos and Erwin W. Van Geenen. 1996. A Parallel-Processing Key-Finding Model. *Music Perception: An Interdisciplinary Journal* 14, 2 (Winter 1996), 185–223.

[48] Christof Weiß and Julian Habryka. 2014. Chroma-based Scale Matching for Audio Tonality Analysis. In *Proceedings of the 9th Conference on Interdisciplinary Musicology (CIM)*. 168–173.