Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Computer Science

Astrid Kongelf Jensen

# Minor Variations

## Improving Computational Key Finding for Minor Keys

Master's thesis in Informatics
Supervisor: Björn Gambäk
January 2023

NTNU
Norwegian University of
Science and Technology

Astrid Kongelf Jensen

# Minor Variations

## Improving Computational Key Finding for Minor Keys

Master's thesis in Informatics
Supervisor: Björn Gambäk
January 2023

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Computer Science

NTNU
Norwegian University of
Science and Technology

# Abstract

Key finding is an integral part of the harmonic structure of western music. Accurate and consistent key classification is an ongoing challenge in Music Information Retrieval and is necessary to facilitate future work in the field. This thesis proposes a new key finding approach that adapts existing methods and improves the classification of minor keys.

By utilizing ideas from western music theory, the proposed model splits its key profile into three minor key variations to facilitate music in natural, harmonic, and melodic minor, respectively. Three experiments were conducted using different parameters. These experiments were based on existing key finding models that were adapted to process the expanded key profile format. The results include accuracy measurements equal to the state-of-the-art methods and indicates a superior ability to process minor keys.

There is still work left to do in order to increase the model's overall accuracy while retaining the powerful minor key processing. It is an exciting approach that still has the potential to achieve great results. The results contribute to the ongoing effort to accurately classify the key of a piece of music using computational methods and can be used to increase the understanding of key finding and tonality in future work.

# Sammendrag

Tonearter er en viktig del av den harmoniske strukturen til vestlig musikk. Nøyaktig og konsekvent toneartklassifisering er fortsatt en utfordring i Music Information Retrieval og er en nødvendig del av videre arbeid innen feltet. Denne oppgaven foreslår en ny tilnærming til automatisk toneartgjenkjenning som tilpasser eksisterende metoder for å forbedre klassifisering av moll-tonearter.

Ved å bruke ideer fra vestlig musikkteori deler den nye modellen moll-tonearter inn i tre varianter som forbedrer måten toneartsprofilene tilpasses moll-variasjoner. Tre eksperimenter ble gjennomført med forskjellige parametere. Disse eksperimentene var basert på eksisterende metoder, og tilpasset hver algoritme for å legge til et nytt format for toneartprofiler. Resultatene inkluderer nøyaktighetsmålinger som er like gode som state-of-the-art metodene testet i eksperimentene og som indikerer en bedre prosessering av moll-tonearter enn andre toneartprofiler. Denne oppgaven bidrar til den pågående innsatsen for å nøyaktig klassifisere tonearten i et musikkstykke ved hjelp av algoritmer. Det kan brukes til å øke forståelsen av toneartsklassifisering og tonalitet i fremtidig arbeid.

# Preface

This thesis is the final part of a Master of Science (MSc) degree at the Department of Computer and Information Science (IDI) at the Norwegian University of Science and Technology (NTNU). The thesis was written in the fall of 2022 and supervised by Björn Gambäck. The work was conducted within the Data and Artificial Intelligence Group at the Department of Computer Science at NTNU in Trondheim.

In preparation for this thesis, a project was completed in the spring of 2022. This included a literature review and a discussion about computational key finding. Some parts of this thesis have been adapted from the corresponding sections in the preparatory work, especially in Chapter 2 and 3.

I want to thank Nápoles López et al., Albrecht and Shanahan, and Bill for permission so use figures and tables, and give a special thanks to Néstor Nápoles López for sharing the dataset used in their work.

Finally, I would like to thank my supervisor Björn for great advice throughout this journey, and friends and family for supporting me throughout my time at NTNU.

Astrid Kongelf Jensen<br>
Trondheim, January 23, 2023

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

The rules of western music have been based on harmony for centuries, and the musical key is the harmonic backbone that defines the structure of a piece [Rameau, 1779]. Key finding has been researched extensively and is considered a primary obstacle in Music Information Retrieval (MIR) [Albrecht and Shanahan, 2013]. A reliable and highly accurate key finding algorithm would make new autonomous analysis, labeling, and organizational techniques for music possible, multiplying the possibilities within MIR research. Despite the extensive work done in key finding, the accuracy of automatic key finding methods has yet to reach the level of a human with knowledge of music theory, especially for detecting minor keys.

This thesis investigates the challenge of classifying minor keys using computational methods compared to the more straightforward classification of major keys. An important topic in the thesis is incorporating music-theoretical approaches to improve minor key finding accuracy. By applying insights from the various types of minor keys to existing foundational and state-of-the-art key finding methods, the aim is to enhance the classification of minor keys in classical music. Incorrect classification of keys can often be attributed to the close similarities between keys or the need for more high-quality data to train key profile values or key finding models. Adding music-theoretical nuances to current models and using high-quality data makes it possible to improve the results of key finding methods, as presented in Chapter 5. To thoroughly examine the topic of key finding, there is a literature review and three experiments focused on adapting and improving existing key finding methods to classify minor keys better.

## 1.1 Background and Motivation

Because keys are such a fundamental part of musical structure, the ability to automatically label the keys of a piece would make it possible for a computer to perform a harmonic analysis of most western music. This knowledge of harmony is necessary in order to be able to search vast amounts of music to find examples of chord progressions, key changes, and other harmonic markers. Because of the vital importance of keys in music, a reliable key finding tool would affect most computational analysis and processing of music. Such a tool could, for example, help music historians by using its harmonic analyses to find historically significant musical events, such as the first use of a particular chord or a look into the evolution of different harmonic techniques over time. In the world of popular music and modern music tools, a key finding algorithm could label extensive music collections so that remixing tools and playlist generators can quickly assemble compatible pieces of music and identify covers of the same song [Temperley and de Clercq, 2013]. The understanding of harmonic structure could also be of great use to research computational creativity and automatic music performance by informing the phrasing and the best way to accompany it based on the functional structure of the music.

As an important marker of harmonic structure in western music, humans can easily use contextual and musical clues to determine the key of a piece. Despite the relative ease trained humans experience while determining the key in most music, computers still cannot determine musical keys at nearly the same level of accuracy. The tonal shifts, modulations, and complex harmonic transitions are some of the aspects of music that make automatic key finding more complicated.

The work done by Krumhansl [1990] serves as a foundation for a lot of later research into key finding. By developing profiles for each possible key, the proposed algorithm correlates each set of total pitch durations with the profile for each key and selects the key with the best match. Several articles have proposed changes to the key profiles to improve this work based on different ideas of what defines a key. Whereas Krumhansl [1990] used psychological experiments to determine which pitches sound right in a given key, Temperley [1999] tweaked the values based on theoretical reasoning and trial-and-error testing. Other researchers, such as Albrecht and Shanahan [2013], used large data sets to determine the relative frequencies of pitches in different keys, and Sapp [2011] presented key profile values based solely on music theory and the theory of keys.

More recently, the methods of key finding have included hidden Markov models (HMM), convolutional neural networks (CNN), matrix factorization, expectation maximization, and Bayesian structures in determining the local and global keys of a musical piece [Nápoles López et al., 2019; Temperley, 2002a; Korzeniowski and Widmer, 2018; Özgúr Izmirli, 2007]. These structures allow the computer to consider more music-related factors than pitches and durations. The key finding approaches make decisions by introducing probabilities, previous key findings in the music, metrical structure, and other musical features.

## 1.2 Goals and Research Questions

**Goal** *Improve the accuracy of minor key classification by combining foundational and state-of-the-art global key classification algorithms with a new key profile and improved data sets.*

As automatic key finding models continue to improve, minor keys still have lower detection accuracy rates than major keys. Despite this persistent problem, approaches have yet to consider the variation within minor keys in a significant way. This thesis attempts to improve foundational and state-of-the-art key finding methods by applying the characteristics of natural, harmonic, and melodic minor key variations. The improvements are made by enhancing the key profile values and using high-quality data to make decisions. Four questions must be answered to reach the goal; these are listed below.

**Research question 1** *How should key profiles be adapted to accommodate minor key variations?*

Key profiles show how a key is defined in a particular model. The possible ways to derive updated key profile values range from psychological experiments or historical data to values based on music theory. A good set of key profiles should yield high-accuracy results for a varied data set and be justifiable from a music theory perspective.

**Research question 2** *What state-of-the-art method(s) are most likely to improve with the minor mode modifications?*

After identifying state-of-the-art methods through a literature review, one or more methods should be chosen for the experiments. The results from Research Question 1 can be used if the methods include key profiles.

**Research question 3** *What data works best for musical key finding, and should it be changed to accommodate minor modes?*

This research question explores how a large data set optimized for key finding should be built. It also looks into the possibilities and the necessity of a new category of labeled data for experimenting with minor key variations.

**Research question 4** *What is the best way to measure success?*

The goal of improving current key finding methods must be quantifiable. By investigating the different methods used to measure success, the results from the experiments should be comparable to existing models in meaningful ways.

## 1.3 Research Method

A combination of theoretical and experimental approaches was used to answer the research questions and achieve the goal set in Section 1.2. Theoretical knowledge was accumulated through the literary review described in Section 3.1, and a combination of theories from relevant articles answered the research questions. The experiments used the knowledge from the research questions to choose the parameters and data input.

To start answering the first research question, the approach was to map out the various key profiles already in use in related work and how they relate to the definition of keys. The answer to Research Question 1 was derived and justified by combining reasoning from articles found in the literature review and domain knowledge from music theory.

The second research question is also closely related to the literature review. The process started with an overview of the available key finding methods and comparing their results based on quantity and quality of data, data types, local vs. global key finding, and experimental results. Then, a three methods were chosen for the experiments by looking at their results for minor keys and whether they used key profiles.

By answering Research Question 3, the process of finding data for the experiments started. The aim was to find a varied and large data set based on the information from the related work. In addition to already existing data, the need for a separate labeled data set with minor key variations was investigated by evaluating what data was necessary to gain insight into the experimental results.

Research Question 4 was answered by looking at the measuring methods used in related work, their strengths and weaknesses, and how they relate to each other.

The overarching goal of the thesis was evaluated using a set of experiments designed to compare the existing methods to the updated versions developed from the answers to the research questions. The experimental results are a basis for evaluating if the goal was achieved. This is discussed further in Section 7.1.

## 1.4 Contributions

1. *A new method for handling key profiles that accommodates minor key variations.*

2. *A comparison between several key finding profiles in different experimental setups.*

3. *An experimental comparison between analyses done on whole pieces of music and excerpts.*

4. *A thorough overview of computational key finding literature.*

## 1.5 Thesis Structure

The remainder of the thesis is structured as follows:

- **Chapter 2** presents the theory behind the musical concepts and computational methods used in the experiments and mentioned in the discussion.

- **Chapter 3** includes an in-depth review of literature related to automatic key finding and a description of how the literature review was constructed. The chapter also includes an introduction to frequently used data sets and methods used to measure success in key finding.

- **Chapter 4** explains the architecture of models used in the experiments conducted in connection to this thesis.

- **Chapter 5** presents the plan, parameters, and results of all the experiments.

- **Chapter 6** discusses the results of the experiments, why they are significant, and what they tell us about the proposed model.

- **Chapter 7** revisits and evaluates the research questions and the overarching goal of the thesis. Conclusions are presented, and future work is proposed.

# Chapter 2

# Background Theory

This chapter presents the theory behind musical keys, an introduction to the computational methods used in the experiments in Chapter 5, and how computational algorithms apply music theory to classify keys. The chapter also contains different methods for measuring success, information about data formats, and information about the software tools used in the thesis.

## 2.1 The Musical Theory of Keys

Music theory and the rules of harmony have defined western music since the end of the middle ages. As the music has evolved, the basic structures of keys and their harmonic progression have remained the same. Western classical music theory is strongly influenced by mathematical structures first documented by Pythagoras (ca. 570-497 B.C.) [Fauvel et al., 2003]. His research into musical ratios, specifically octaves, fourths, and fifths, formed the basis for modern western music theory. Work done by Rameau [1779] confirmed the relationship between music theory and mathematics by defining a harmonic methodology based on the ratios researched by Pythagoras. This harmonic system depends on musical keys that determine the harmonic choices available to the composer at any time. However, rules are made to be broken, and rules in music theory are no different. This variability necessitates subjective analyses by music theorists and makes the task of computational key finding non-trivial. The following subsections detail the formal definition of a musical key, key changes within a piece, and the factors contributing to the ambiguity of key finding.

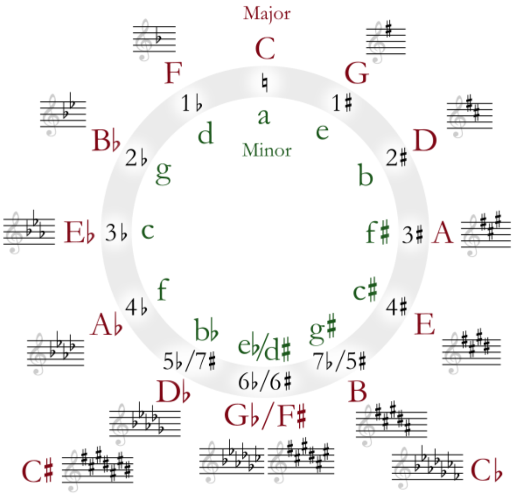


Figure 2.1: The 24 keys organized in the circle of fifths. Parallel keys share the same pitch name; relative keys are placed across from each other, and keys a fifth apart are next to each other.[1]

### 2.1.1 Technical Aspects of Musical Keys

A key is defined by its key-note (tonic) and its mode [Rameau, 1779]. The mode can be any of the twelve modes, but Ionian or Aeolian are the most commonly used, often referred to as major and minor. The mode defines the distances between scale degrees in a key. For example, the distance is four semitones between the first scale degree (tonic) and the third in major, and three semitones in minor. These distances are the most critical relationships for establishing the 'sound' of the key.

Because there are 12 pitches available and two commonly used modes, most pieces are in one of 24 keys at any time. The keys are arranged by their degree of similarity in a structure called the circle of fifths. These relationships can be seen in Figure 2.1. Keys close to each other on the circle have similar modes or tonal centers and are easier to mistake for one another.

[1] [Bill, 2021] CC BY-SA 3.0. https://creativecommons.org/licenses/by/3.0

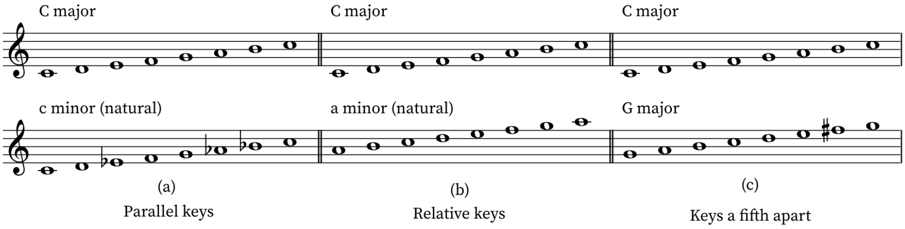


Figure 2.2: Scales showing the relationship between parallel keys (a), relative keys (b), and keys with a tonal center a fifth apart (c).

### Closely Related Keys

*Parallel keys* are keys that share the same key-note but have different modes. In Figure 2.1, this is represented by the key having the same pitch name but with different capitalization. The different modes provide different pitches for the scale degrees in each key. However, they still share a tonal center and some important pitches, for example, the fifth (dominant) and the tonic. This relationship is shown in Figure 2.2 (a).

*Relative keys* are opposite each other in the circle of fifths and share all the same pitches. The tonal center is not at the same place for both keys, but since they share all the same pitches, it can be hard to differentiate between them. An example of the C major and a minor (natural) is shown in Figure 2.2 (b).

*Dominant keys* are defined by key-notes that are a fifth apart. They are neighbors on the circle of fifths and can also be hard to differentiate. As shown in Figure 2.2 (c), they share all but one pitch. Because the dominant (fifth) is essential in the harmonic system, pieces frequently use the dominant chord and often modulate between the two keys. Therefore there can be much ambiguity and small nuances when trying to differentiate between them in analysis.

### Minor Key Variations

Minor keys are also different based on the context in which they are played. Three common variations are used: *natural*, *melodic*, and *harmonic* minor (shown in Figure 2.3). Each variation has a distinct sound, but they all share the important lowered third scale degree, making them sound minor. The *natural* variation is the original minor mode. The *harmonic* has been changed to include a raised seventh scale degree, also called a leading tone, that leads to the tonic. The leading tone is a harmonic tool to make the piece sound more at home in its key. *Melodic* minor is special because it includes different notes based on whether it is ascending or descending. The ascending scale includes a raised sixth and seventh

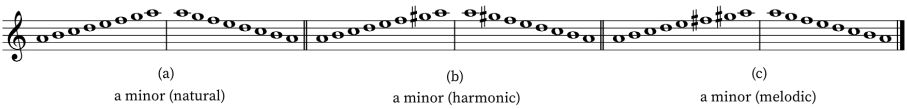


Figure 2.3: Three variations of the minor scale. Natural (a), Harmonic (b), and Melodic (c).

scale degree, which are lowered again for the descending scale. The different minor modes can skew pitch class value sets and can make minor modes more challenging to predict.

### 2.1.2 Modulations, Tonicizations, and Ambiguity

As mentioned previously, composers only follow the rules of harmony when it suits them. Even if they do, subjective analysis is often needed to pinpoint the key at any given time. The following section discusses some of the features of music that make key detection more complicated.

#### Modulation and Tonicization

It is common for composers to change keys throughout a piece, and there are many different ways composers can transition from one key to another. In many cases, the transitions are done using common chords found in both keys; in these cases, the modulation or tonicization is often to a closely related key. However, a key change can also be done by, for example, repeating a chord until the listener is used to the new sounds, and the piece can continue in the new key. Because the composer often wants the key change to be as smooth as possible, there is usually a period where the music could be in either of the keys that are a part of the transition. This ambiguous tonal space is where a human expert would perform a functional harmonic analysis and determine what key is the most prominent at any time, how the key change was made, and exactly where it changes if that is possible.

In order to complicate matters further, an apparent key change might not always be labeled as such by an expert. If an expert is satisfied that the piece changes its key, they would call it a modulation and set the section's key to the new one. If the key change is short, however, it would be labeled as a tonicization, and the key of the piece would remain the same as before the apparent key change.

### Music Beyond the 24 Keys

Most of western tradition's music is tonal, meaning we can assign a key to the piece. Some exceptions exist, specifically in older and more modern classical music.

Music has traditionally been arranged in a system of twelve modes. Because of this, the music is technically tonal but does not fit into the 24 keys commonly used today. It is likely possible for a computational model to determine the mode of a piece by adapting it to the twelve modes, but there is not as much data available as with music written in the 24 keys. In the $20^{th}$ century, some music evolved into atonal and 12-tone styles that cannot be associated with a key.

In addition to these periods of history, a lot of music is written in ambiguous tonal spaces. This includes music inspired by folk music from certain regions, non-western music, and music inspired by medieval modes.

## 2.1.3 Measuring Distance Between Keys

There are many ways to visualize and model the relationships between keys. The Circle of Fifths (Figure 2.1) and Schoenberg grids (Table 2.1) show how keys relate. Keys can also be modeled on a plane, increasing distance-measuring possibilities. Even more measuring methods are presented in Section 3.2.4.

### Correlation Coefficient

Correlation coefficients are statistical measurements that can evaluate the relationship between two sets of variables. Related to key finding, correlation coefficients can be used to evaluate the similarity between a musical section and a set of key profiles.

The key- and pitch profiles are represented as vectors using correlation coefficients for key finding. The correlation coefficient calculated between the two can range from -1 to 1, where -1 indicates a negative correlation, 1 indicates a positive correlation, and 0 indicates no correlation. A correlation means that the values have a relationship that is in some way linear. The higher the correlation, the more related the values are. A high correlation coefficient in key finding usually means that the specific key profile likely represents the piece's key. Correlation coefficients are easy to interpret and calculate and give a good measurement of the similarity between a key- and a pitch profile.

### Euclidean Distance

Euclidean distance is a way to measure the distance between two points in a multi-dimensional space. In the case of key finding, the key- and pitch profiles

Table 2.1: Visualization of a two-dimensional Schoenberg grid. Each column follows the circle of fifths and each entry is surrounded by its relative and parallel keys on either side.[2]

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| A♯ | a♯ | C♯ | c♯ | E | e | G | g | B♭ |
| D♯ | d♯ | F♯ | f♯ | A | a | C | c | E♭ |
| G♯ | g♯ | B | b | D | d | F | f | A♭ |
| C♯ | ♯ | E | e | G | g | B♭ | b♭ | D♭ |
| F♯ | f♯ | A | a | C | c | E♭ | e♭ | G♭ |
| B | b | D | d | F | f | A♭ | a♭ | C♭ |
| E | e | G | g | B♭ | b♭ | D♭ | d♭ | F♭ |
| A | a | C | c | E♭ | e♭ | G♭ | g♭ | B♭♭ |
| D | d | F | f | A♭ | a♭ | C♭ | c♭ | E♭♭ |

are treated as points in a 12-dimensional space where each pitch value represents a coordinate. If the Euclidean distance between two points is close to 0, then the two profiles are nearly identical, and there is a strong possibility that the key profile represents the key of the piece. The Euclidean distance measurement can be more sensitive to outliers than other distance measurements.

### Schoenberg Grids

Schoenberg grids were developed by Arnold Schoenberg and are a way to visualize the relationships between keys [Schoenberg and Stein, 1969; Purwins et al., 2007]. There are several variations of Schoenberg grids. The one used by Nápoles López et al. [2019] uses a two-dimensional space where each key is spaced so that the nearest neighbors are the dominant (in either direction), the relative, and the parallel keys. As more layers get added, the keys are more distantly related to the center key. Table 2.1 shows a version of this grid. In order to use the grid as a measurement, the keys relate to each other by the degree of separation. This is explained further in Section 4.4 and shown in Table 4.3 on Page 40. Schoenberg grids are not very common in key finding algorithms but are a good way to visualize the relationships between keys.

[2] Copied from Nápoles López et al. [2019] and used with permission from ACM Press, License Number 1314330-1. The license grants permission to republish a chart/graph/table/figure in a thesis/dissertation.

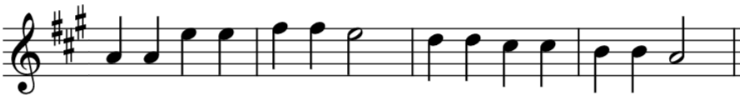


Figure 2.4: The two first measures of *Twinkle Twinkle Little Star.*

| Pitch | Duration |
|---|---|
| C | 2 |
| C♯ | 0 |
| D | 2 |
| E♭ | 0 |
| E | 4 |
| F | 0 |
| F♯ | 2 |
| G | 0 |
| A♭ | 0 |
| A | 4 |
| B♭ | 0 |
| B | 2 |

Table 2.2: Durations for the pitches in the two first bars of *Twinkle Twinkle Little Star.*

## 2.2 Computational Key Finding

Many techniques have been explored to find an accurate method for key finding in western music. The most famous early experiments were done by Krumhansl [1990], and since then, the methods used have diversified to include artificial intelligence methods and machine learning.

### 2.2.1 The Krumhansl-Schmuckler Key Finding Algorithm

The Krumhansl-Schmuckler (KS) key finding algorithm has become the basis for a lot of later research into key finding. The algorithm requires no previous knowledge of the piece and is very simple to use because of its low computational complexity and memory use [Krumhansl, 1990].

The algorithm computes a segment's global key by adding up each pitch's total duration. These durations are transformed into a 12-dimensional vector, $I = (d_1, d_2, ..., d_{12})$, and used as input for the remaining part of the algorithm. The $d_j$ values are arranged in ascending order by semitone, starting at C. This means that $d_1$ represents the total duration of the pitch C or its enharmonic

| **Key** | A | Am | E | D | Em | F♯m | C | Bm | G | C♯m | Dm | B |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Score** | 0.78 | 0.73 | 0.56 | 0.54 | 0.56 | 0.40 | 0.31 | 0.30 | 0.26 | 0.17 | 0.17 | 0.09 |

| **Key** | F | A♭m | F♯ | Gm | Cm | E♭m | B♭ | Fm | B♭m | C♯ | E♭ | A♭ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Score** | 0.23 | -0.25 | -0.29 | -0.31 | -0.34 | -0.39 | -0.41 | -0.52 | -0.52 | -0.62 | -0.63 | -0.63 |

Table 2.3: Sorted table of key value scores for the first two bars of *Twinkle Twinkle Little Star* based on the durations in Table 2.2.

equivalent B♯. $I$ is correlated with twenty-four 12-dimensional vectors containing the pitch values for each key $i$ $K_i = (P_{1i}, P_{2i}, ..., P_{12i})$. These values are based on experiments done by Krumhansl [1990] and Kessler discussed below. The correlation values are assembled in an output vector $R = (r_1, r_2, ..., r_{24})$ where $r_i$ represents the correlation between the input and the given key $i$. The key, $i$, with the highest $r_i$ is the most likely candidate.

The algorithm's results for the first two bars of *Twinkle Twinkle Little Star* (Figure 2.4) are shown in Table 2.3 along with the duration of each pitch in the sample in Table 2.2. The table follows the expected pattern. The correct key (A major) has the highest value. The other high scorers are the parallel key (a minor), the key a fifth apart (E major), the key a subdominant below (D major), the parallel key to the dominant key (e minor), and the relative key to A major, f♯ minor. The low scorers are also predictable, with A♭ major placing last, a key with no shared pitches with A major.

### The Krumhansl-Kessler Probe Tone Experiments

The Krumhansl-Kessler experiments sought to discover how humans perceive musical keys and what notes fit the best into a key [Krumhansl, 1990]. The experiments involved playing an eight note scale followed by tonic triads and three different chord sequences to solidify the key. Then a probe note was played, which the participants rated based on how well the pitch 'fit in' the context of the key. One of the important discoveries Krumhansl and Kessler made was that the probe tone values were similar for all major keys and all minor keys. The two resulting sets of values were applied to the 12 pitches, resulting in 24 12-dimensional vectors containing pitch class values.

## 2.2.2 Hidden Markov Model

Hidden Markov models (HMMs) return the most probable sequence of hidden events based on related observable events. Related to this thesis, this means

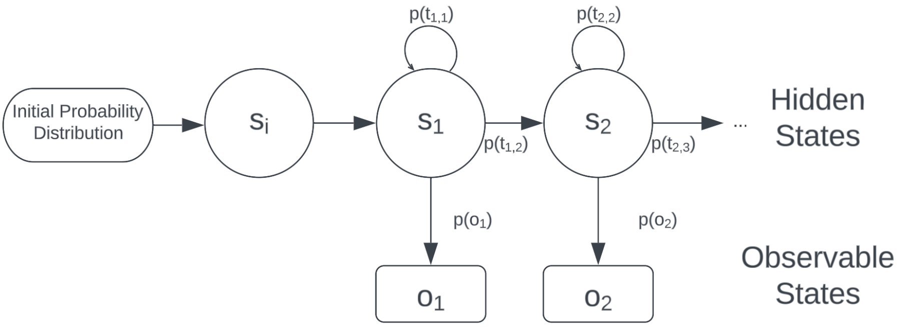


Figure 2.5: A hidden Markov Model. The initial probability distribution informs the model what states the initial hidden state can be and each of their probabilities. The observed sequence informs the hidden states via the emission probability distribution, and the transitional probability distribution informs the transitions between hidden states.

determining the sequence of keys in a piece based on the observable sequence of pitches. The model comprises five components, the initial probability distribution, observed sequence, possible hidden states, transition probability distribution, and emission probability distribution. These are described further below and shown in Figure 2.5.

The initial probability distribution denotes the likelihood of starting in any given state. The observed sequence of events is the base of the hidden sequence predictions and reflects observable events related to the hidden states. The hidden states are a collection of possible hidden states, and the transition probability distribution represents the transfers between them. The emission probability distribution shows the probability of an observable event emitting from a given hidden state.

### The Viterbi Algorithm

The Viterbi algorithm is designed to efficiently locate the most probable sequence of hidden events [Forney, 1973]. It calculates the probability of the occurrence of every state and discards the sequences that are guaranteed to be less likely than others. By applying the Viterbi formula to each possible transition and state, it returns the most probable sequence of states and the probability that it has occurred.

## 2.3 Measuring Success

Different methods of measuring success provide different insights into the strengths and weaknesses of key finding methods and key profiles. The two most common ways of determining the success of global key finding methods are the accuracy and MIREX scores.

### 2.3.1 Accuracy

The accuracy percentage is the most basic measurement of success for key finding. The accuracy score is a measure of the correct classifications over the total amount of test cases:

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Number\ of\ total\ predictions} * 100 \tag{2.1}$$

#### Measuring Accuracy for Minor Keys

Because this thesis focuses on classifying minor keys, special attention is placed on analyzing the results of the models as they pertain to minor keys. There are two ways to treat the accuracy measure to measure the success of minor key classification: the ground truth minor accuracy and the predicted label minor accuracy.

Equation 2.1 can be rewritten to apply to a subset of the data. This subset can be calculated in two ways: by using the ground truth labels or the predicted labels. Both variations are computed using the following equation:

$$Subclass\ Accuracy = \frac{Number\ of\ correct\ predictions\ within\ Subclass}{Number\ of\ total\ predictions\ in\ Subclass} * 100 \tag{2.2}$$

The ground truth variation subclass can be represented like this:

$$Subclass = Data\ entry\ whose\ ground\ truth\ mode\ is\ minor \tag{2.3}$$

This measures the overall ability of the method to classify minor keys correctly. The percentile returned is how likely the model is to predict the key correctly, given that the key is minor.

The predicted label subclass can be represented like this:

$$Subclass = The\ predicted\ label\ mode\ is\ minor \tag{2.4}$$

Which tells us that if the predicted label is a minor key, the percentage represents the probability that the prediction is correct.

These minor accuracy variations are closely related but show different behaviors from the proposed model. The difference between them is further discussed in Section 6.1.2.

### 2.3.2 MIREX

The *Music Information Retrieval Evaluation eXchange* (MIREX), introduced further in Section 3.2, hosts annual key finding competitions and has developed its own scoring system. Since many studies have similar accuracy results, many researchers use alternative scoring systems to differentiate between the results. Because most key finding models struggle with the same issues, classifying closely related keys instead of the correct one, MIREX has devised a scoring model to consider this.[3] By looking at the results through the MIREX model, it is possible to differentiate the models that make significant errors from the ones that generally classify the correct key or a closely related one. The MIREX score is calculated as follows:

> The classification belongs to one of the following categories:
> **Correct:** The classification is the same as the ground truth.
> **Dominant key:** The classification is a fifth above or below the ground truth.
> **Relative key:** The classification is the relative key to the ground truth.
> **Parallel key:** The classification is the parallel key to the ground truth.
> **Other:** The classification is none of the above.

After calculating the ratios of each classification category, the following value is computed:

$$w = r_c + 0.5 * r_d + 0.3 * r_r + 0.2 * r_p \tag{2.5}$$

Where $r_c$ is the ratio of correct classifications, $r_d$ is the ratio of dominant keys, $r_r$ is the ratio of relative keys, and $r_p$ is the ratio of parallel keys. The weights for each category were decided based on how close the relationships between the key category and the ground truth were.

The MIREX result calculation is the most widely used because it rewards models that classify keys closely related to the correct one while also demanding a completely correct set of classifications to achieve 100% accuracy.

[3] https://www.music-ir.org/mirex/wiki/2020:Audio_Key_Detection#Evaluation_Procedures

## 2.4 Data Formats

Several data formats are available for musical data labeled with the global keys. The different types of data convey different information. A suitable format may be necessary for research involving more than the basic information of pitches, pitch durations, and other essential musical markers. The formats are often interchangeable when considering the basic musical information and can be transferred to other formats using tools like music21. The two categories of data are symbolic and audio formats, each providing different challenges and problems.

### 2.4.1 Symbolic

Symbolic musical data encodes information in a notation format that does not include sound. The stored information includes pitches, onset and duration of notes, time signatures, and more. The different formats are specialized to handle different tasks, for example, information retrieval, sheet music generating, or playback, and can usually be converted to appropriate formats without much extra work.

#### **kern

The **kern[4] format is designed to represent all musical ideas from the common-practice era in western music. The important features in the format are the pitches and their durations. There is also functionality for the representation of information like accidentals, stem directions, and barlines, but these only serve as supporting information. The intended primary use for the **kern format is Music Information Retrieval (MIR) and is not necessarily useful for making a visual representation of the music.

#### MIDI

The Musical Instrument Digital Interface (MIDI)[5] format carries information about the sound of music using less memory space than audio files. The recorded information is based on pitches and their durations. Specialized programs can interpret the MIDI format and create an audio version of the music based on the information in the file.

[4] https://www.humdrum.org/rep/kern/

[5] https://www.midi.org/specifications/midi1-specifications/general-midi-specifications

#### MusicXML

Extensible Markup Language for music (MusicXML)[6] is a markup language designed to store and edit visual depictions of music, such as sheet music. The language supports standard features from western classical Music. MusicXML allows sheet music to be transferred between different software programs.

### 2.4.2 Audio

Audio data is essential for imitating real life for MIR. As mentioned in Section 1.1, several use cases involve the analysis of audio data. The data includes a set of frequencies that can be processed and analyzed to extract the pitches and durations used in the key analysis. While it is possible to transform symbolic data into audio files, for example, by processing a MIDI file, it is much more complicated to convert audio files to symbolic formats because of the non-trivial task of determining pitches, onset times, and the duration of notes.

#### Chroma

The chroma structure is used as a symbolic representation of audio data. An array of the durations of pitches is created by processing the frequencies in the sound excerpt. The result is often similar to Table 2.2. However, because the information has been found through processing frequencies, the accuracy of the data may be less precise than the distribution derived from symbolic data.

## 2.5 Tools

Because the MIR field is growing, more software tools are becoming available for processing music information. The following tools have built-in key finding methods that were adapted in the experiments presented in Chapter 5 to make the new key profile format fit with the analysis tool. The customization is explained further in Section 5.2.

### 2.5.1 Music21

For all the experiments in Chapter 5, the data is processed by the tools in the music21 library [Cuthbert and Ariza, 2010]. The toolkit is an open-source project focusing on tools for music information retrieval and research in Python. The

[6] https://www.musicxml.com/

toolkit contains functions for composing, analyzing, and transcribing music and an extensive library of musical data that can be used for research.[7]

### 2.5.2 Justkeydding

Justkeydding is a tool developed by Nápoles López et al. to perform a key analysis using hidden Markov models and ensemble training as explained in *Key-Finding Based on a Hidden Markov Model and Key Profiles* [Nápoles López et al., 2019]. The tool accepts MusicXML, MIDI, and **kern files as input; its output is the predicted key. The code is publicly available through GitHub.[8]

[7] Music21 is Copyright ©2006-2021, Michael Scott Cuthbert and cuthbertLab. Music21 code (excluding content encoded in the corpus) is free and open-source software, licensed under the BSD License.

[8] https://github.com/napulen/justkeydding/tree/pythononly/justkeydding,
MIT License
Copyright (c) 2018 Nestor Napoles

Permission is hereby granted, free of charge, to any person obtaining a copy of this software and associated documentation files (the "Software"), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions:

The above copyright notice and this permission notice shall be included in all copies or substantial portions of the Software.

THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.

# Chapter 3

# Related Work and Literature Review

This chapter presents the work done by others in the field of automatic key estimation and the approach taken to find and process relevant articles. There is an introduction to the state-of-the-art methods, methods referenced in the research presented in Chapter 4, and the foundational work still referenced frequently in papers.

## 3.1 Structured Literature Review

The structured literature review is based on the guide written by Kofod-Petersen [2018]. The guide outlines three important phases: (1) Planning the review, (2) Conducting the literature review, and (3) Reporting. In order to facilitate reproduction, each step of the literature review is documented in the protocol. Important papers cited in articles included in the review were added to the list of reading material found in the initial query.

### 3.1.1 Phase 1: Planning the Review

The first steps in the literature review consist of identifying the need for and commissioning a review, formulating research questions, and the creation, continuous evaluation, and finalization of the review protocol. Identifying the need and commissioning the review is an implicit of the master's thesis writing process. The research questions written as a part of Phase 1 are discussed in Section 1.2. These questions were developed as questions that need answers in order to complete the overarching goal of the thesis.

Table 3.1: Search terms used in the SLR.

| | Group 1 | Group 2 | Group 3 |
|---|---|---|---|
| Term 1 | Key | Detection | Music |
| Term 2 | | Recognition | Tonal |
| Term 3 | | Classification | Harmony |
| Term 4 | | Finding | Harmonic |
| Term 5 | | Determination | |
| Term 6 | | Extraction | |

### 3.1.2 Phase 2: Conducting the Literature Review

Five steps are completed in order to find the articles related to the research questions. Each of these are described below.

**Step 1: Identification of Research**

The query used to search for relevant data is designed to find the existing approaches to automatic key finding. By grouping terms based on synonyms, the query results include a wide range of articles that cover relevant topics. The terms used in the search are presented in Table 3.1: Group 1 ensures that the search result includes a mention of key, Group 2 includes synonyms for the detection of said key, and Group 3 ensures that the result has to do with music and harmony. The resulting query is based on the search terms and designed to include all permutations of the search terms:

> (Key) AND (Detection OR Recognition OR Classification OR Finding OR Determination OR Extraction) AND (Music OR Tonal OR Harmony OR Harmonic OR Minor)

The query was entered into the databases and search engines shown in Table 3.2. The different databases and search engines were chosen because of their thematic connection to the topic or in order to facilitate a broad search across many journals. The next step in the process narrows the field down considerably by defining important criteria for the papers to be included.

**Step 2: Selection of Primary Studies**

In order to decrease the amount of qualifying data, two criteria were applied to the results. The papers had to be published within the last 20 years and all duplicates were removed. By restricting the results to the ones published after 2002, all modern work is included and important work done previously can be

Table 3.2: Search engines and databases queried for the SLR, the number results for each published after 2002, and the number of studies left after filtering by year, duplicates, and performing a quality assessment.

| Database/Search Engine | # Results After 2002 | # Results After Quality Assessment |
|---|---|---|
| Google Scholar | 1,670,000 | 5 |
| IEEE Xplore | 2,643 | 5 |
| ISMIR | 1,380 | 7 |
| SemanticScholar | 781 | 9 |
| WebOfScience | 20,828 | 6 |
| ResearchGate | 256 | 13 |

accessed via the citations in the chosen studies. The number of results published after 2002 are presented in Table 3.2. The top 30 results from each site after adjusting for publishing date and duplicates were included in the next processing step.

### Step 3: Study Quality Assessment

After reducing the number of results to 30 per search engine and database (180 in total), a quality assessment was performed to reduce the primary studies further and quantify the study's quality and relevance to the research questions. The quality assessment included four inclusion criteria (IC) and ten quality criteria (QC). These criteria are listed in Appendix C and include, for example:

**IC1:** The main focus of the study is automatic key finding.

**IC2:** The study presents empirical results or other significant contribution to the field.

**IC3:** The study is a primary study, concerns descriptions of technical aspects of the key finding process (e.g., measuring the distance between keys), or describes methods of measuring success in key finding.

**IC4:** The study uses symbolic data (MIDI, Chroma, Sheet Music, etc.)

**QC1:** Is there a clear statement of the aim of the research?

**QC2:** Is the study put into the context of other studies and research?

The assessment was done in stages, starting with the ICs. The first step was reading each abstract and eliminating the articles that did not meet IC1 and

IC2. Then each article whose full text did not meet IC3 and IC4 was discarded. Finally, the articles were evaluated using the QCs presented in Kofod-Petersen [2018]. Each article was given a score based on how well they did with the QC's and the quality threshold was set at 6.5 points. Three articles were removed after scoring lower than the threshold.

The articles that remained after the quality assessment are listed in Table A.1. This table does not include the articles discovered independently from the SLR.

## 3.2 Related work

Computational key finding is considered an important step towards fully automated music analysis and is, therefore, a vital part of the field of music information retrieval (MIR). *Cognitive Foundations of Musical Pitch* by Carol L. Krumhansl [1990] was an early landmark publication in the field and is still frequently referenced in research. A vital driving force for new research is the annual *Music Information Retrieval Evaluation eXchange* (MIREX)[1] conference where a competition is held for symbolic and/or audio key finding. This encourages new research into the topic and rewards scientists who improve their models year after year. Following the review of significant contributions to key finding and state-of-the-art methods, there is a presentation of data types and challenges related to data.

What defines a key and how we measure it are two questions central to various key finding methods. This is because the theoretical and practical ways of determining the key of a piece differ. Music theory states that a key is defined solely by the pitches that make up its scale, while humans usually use context clues, meter, and musical form to help aid the classification. The key finding methods described in this section all make choices about the definition of keys. Every method defines whether keys are based on what people hear, the music theory, the historical use, whether only pitches matter or if contextual clues should play a part in how keys are classified.

### 3.2.1 Krumhansl-Schmuckler and its Variations

The Krumhansl-Schmuckler (KS) key finding algorithm discussed in Section 2.2.1 is the foundational work done in computational key finding. It is different from other key finding methods because of its focus on the human psychology and perception of musical key [Shmulevich and Yli-Harja, 2000]. The assumption behind the algorithm is that pitches define the sound of a key.

[1] https://www.music-ir.org/mirex/wiki/MIREX_HOME

As the foundation for future research, the work done by Krumhansl [1990] is helpful because of its low computational complexity and well-documented assumptions. Its results for music with a sKStonality is 91.4% accuracy [Temperley, 1999]. Despite these merits, other scientists have highlighted some problems with the algorithm and key profile that, when fixed, could lead to more accurate results. These shortcomings are part of why the algorithm, which can perform well on pieces with a well-established tonality like the preludes of Bach's *Well-Tempered Clavier*, only classifies with 45.8% accuracy when faced with preludes by Chopin that are more ambiguous in their tonality [Temperley, 1999].

The most commonly discussed areas in the KS algorithm are: the assumption behind the key profile weights, how it is best to measure the distance between pitches, chords, and keys, how the algorithm handles passages with more than one key, and that the algorithm only considers pitches when making its decision. Each of these topics is discussed in the sections below.

### 3.2.2 Key Profiles

After the publication of the KS algorithm, many have proposed new and improved key profiles, for example, Temperley [1999], *Aarden-Essen*[2], and Bellmann [2006]. The various types of key profiles are central to the definition of keys, and many varieties have been proposed. All key profiles mentioned are listed in the Appendix B. The following three are the most important groupings.

#### Key Profiles based on Human Psychology

The key profile derived from the Krumhansl-Kessler experiments is based on the idea that a musical key is defined by how humans perceive it. By calculating how well a pitch 'fit in' a particular key through extensive experimenting, they created key profiles that they later correlated with musical pieces to determine the key. As new key profiles have been proposed, this approach to making key profiles is not in frequent use anymore.

#### Theoretical Key Profiles

In his article, *What's Key for Key? The Krumhansl-Schmuckler Key Finding Algorithm Reconsidered*, Temperley [1999] proposed four changes to the KS algorithm, one of which was changing the values associated with the key profiles. The revised values aim to improve the predictions' stability and fix the skew towards predicting a minor key over a major key. He writes that the values were derived from theoretical reasoning and trial and error testing while keeping the values somewhat similar to the ones defined by Krumhansl and Kessler.

[2] http://kern.ccarh.org/browse?l=/essen

The first issue he addressed was the enlargement of the differences between diatonic pitches (the member pitches in the key) and chromatic pitches. The pitches that have relative importance according to harmonic music theory have higher values, especially the pitches in the basic tonic triad: the tonic, mediant (4 semitones above tonic), and dominant (7 semitones above tonic). By changing the key profiles to ones informed more by music theory than psychology, Temperley [1999] changed the original assumption made by Krumhansl [1990], that key is defined by the sound of the music rather than the theoretical relationship between the pitches. Studies like the one published by Catteau et al. [2006] use this as an argument for using the Temperley key values over the Krumhansl-Kessler values.

The key profile designed by Sapp [2011] is also frequently used. It was designed as a set of simple values that were supposed to act as a benchmark for the other key profiles to be tested against. The idea was to give 2 points for the tonic and dominant pitches, 1 for all other diatonic pitches, and 0 for those not in the key. The key profile achieved fairly good accuracy [Sapp, 2011].

### 3.2.3 Data-driven Key Profiles

To further improve key profile values, some researchers use data to define their values. Examples of data-driven key value sets are the ones generated by Bellmann [2006], Albrecht and Shanahan [2013], Aarden [2003], and Kostka et al. [2018]. They have all proposed new sets of values by adding up all the pitches used in music labeled with their keys. Because labeled data can be challenging to create and hard to find (discussed further in Section 3.3), the new profiles are based on relatively small data sets whose genres vary.

There are several benefits to using the data-driven approach, the main one being that it reflects the music written in practice. A drawback is that since there is still only a small amount of labeled data available, the key values only reflect a small subset of music and can easily be skewed.

In discussing different key profiles, Sapp [2011] argues that data-driven experiments to find profiles perform very well but should be specialized to the genre that the model will be classifying later.

### 3.2.4 Measuring Distance Between Keys

Measuring the distance between pitches, chords, and keys is central to key finding because it defines what keys are the closest relatives to the piece of music being processed. The first measurements used were correlations by Krumhansl [1990] and scalar products by Temperley [1999], achieving similar results. Krumhansl later performed experiments with a multidimensional spacing of keys that, when reduced to three dimensions, revealed the placements of notes in a key as a cone [Krumhansl, 1990]. The cone has roughly four layers and starts with the root pitch at the vertex at the bottom of the cone. The next layer has the two other pitches in the root triad, the third and fifth scale degree. The following layer has the remaining diatonic pitches (in the scale), and the top layer has the non-diatonic (chromatic) pitches. The spacing of the pitches on the cone has been calculated to represent the results of the Krumhansl-Kessler experiments best and to simplify calculations, for example, by averaging the perceived distance from the scale degrees $\hat{0}$ to $\hat{4}$ and $\hat{4}$ to $\hat{0}$ to disregard temporal differences. Improvements to the Krumhansl [1990] cone are proposed in work by Lerdahl [2001]. He extended the cone by adding a layer of diatonic fifths between the triad chord tones and the diatonic pitches. In addition, he flipped the cone upside down and repeated the notes at all underlying levels. The distance between two chords in a key or two separate keys can be calculated by the number of changed notes in the cone as done in work by Catteau et al. [2006].

Music theorists have developed complex models to represent the distance between keys that involve cones, toruses, and other mathematical structures [Lerdahl, 2001; Purwins et al., 2007]. Even though these models are more precise than vector-based distance measurements from a theoretical perspective, they do not appear in state-of-the-art articles. The measurement methods used for the methods discussed extensively in this thesis are correlation, Euclidean distance, and the Schoenberg grid. These methods are explained in Section 2.1.3.

### 3.2.5 Handling More Than One Key

The KS algorithm is designed to find the global key of any music section it analyzes. This makes it impossible for the algorithm to detect a modulation within the section, and the presence of more than one key weakens the confidence of the final prediction. Several approaches have been presented to combat this, for example, by deciding only to use the first and last eight bars of a piece that are usually firmly in the global key of the piece [Albrecht and Shanahan, 2013; Temperley, 2002b]. Some attempt to find local keys and decide what the global key is at the end like Sapp [2011] and Lee and Slaney [2008]. Other methods to find the global key are presented by Nápoles López et al. [2019] and Peeters [2006], using hidden Markov models to predict the global key.

Krumhansl [1990] recommended using a sliding window to improve the results of the KS algorithm. By having small, overlapping segments, the algorithm's results can be compared, and a key and moments of modulation can be determined. Despite the KS algorithm's simplicity, much information is contained in the correlations' results. For example, if one key profile scores much higher than all the others, the confidence behind the prediction is high. The relationships can also detect modulations if two key profiles score higher than others. Knowing where the piece modulates can be helpful if a model knows what local keys are present and can use this to infer the global key.

### 3.2.6 Incorporating More Information

Whether to incorporate more information than just the pitches is a choice all researchers have to make. As mentioned previously, a key is, in theory, only defined by what pitches are present in a music section. When humans study sheet music to classify a key, more information than the pitches is usually incorporated because they look at different contextual clues in the music. Below are a few ways researchers have decided to include more than just pitches in their calculations.

#### Handling Uncertainty

Most problems with uncertainty linked to them can be solved using logic and probabilities. The methods mentioned so far have been firmly planted in logic, using key profiles and other rules to make their decisions. Another approach to the problem is to incorporate probabilities into the problem solving as it is done in Temperley [2002a], Temperley [2002b], Papadopoulos and Peeters [2012], and Lee and Slaney [2008] among others. The methods that combine probability and pitches to determine the most likely key structure are Temperley [2002b], who uses a Bayesian approach, and Papadopoulos and Peeters [2012], Nápoles López et al. [2019] and Lee and Slaney [2008], who use an HMM. The methods using musical information other than pitches will be discussed later in the section.

#### Bayesian Logic

By applying Bayes' rule it is possible to calculate the most likely key based on the pitches, or in other words, calculate the most probable structure based on the surface information [Temperley, 2002b]. In order to find the key with the highest probability given the pitches, the algorithm needs to know the probability of a pitch appearing in a given key, which is information from key profiles. In Temperley [2002b], the key profiles used are the ones defined by the *Kostka-Payne* corpus [Kostka et al., 2018].

The model in Temperley [2002a], which is almost identical to the one described in Temperley [2002b] except for a few values, placed first in the first and only MIREX (Music Information Retrieval Evaluation eXchange) competition held for key finding from symbolic data, scoring 91.4% based on the MIREX scoring system described in Section 2.3.[3]

### Using HMM to Model Key Structure

Just as Bayes' rule attempts to determine the 'hidden' probability of a key given the pitches present, the HMMs used in Nápoles López et al. [2019], Raphael and Stoddard [2003], Peeters [2006], and Lee and Slaney [2008] find the optimal sequence of hidden states to match the observed events. By doing this, the model considers both the pitches and the music's temporal dimension. Two methods have trained one model for major and one for minor before transposing them to the 12 pitches [Peeters, 2006; Lee and Slaney, 2008]. This helps divide the available data to all the keys instead of lacking data on the less frequently used keys, which could lead to overfitting. Nápoles López et al. [2019] use existing key profiles as observational probabilities and a Schoenberg grid as the transitional model. This lowers the amount of training needed to prepare the model.

### Convolutional Neural Network as Pattern Recognition

Korzeniowski and Widmer [2018] use a convolutional neural network (CNN) to detect global keys and find overarching structure in the music. In their paper, they mention that local key estimation should be possible with a similar algorithm and that it has been saved for future work.

### Using Musical Structure as Supporting Information

An expert analyzing a musical piece would look at more information than just the pitches to determine the key. Papadopoulos and Peeters [2012] argue that metrical structure is important to find the best key results. Similar arguments are made by Lee [2008] about genres in the music and Raphael and Stoddard [2003] regarding the rhythm in the music.

## 3.3 Data

Limited access to labeled data is a central problem in music information retrieval (MIR) research. In order to be confident in the accuracy of harmonic analysis, both symbolic data and audio files need expert knowledge to be annotated. Each

[3] https://www.music-ir.org/mirex/wiki/2005:Symbolic_Key_Finding_Results

has different challenges regarding labeling keys, chords, and pitches. All studies referenced in this thesis use either symbolic or audio files processed in some way to extract pitches and other relevant information. Valuable data for testing usually consists of music represented as audio or symbolically with the key labeled globally for global key methods or in every section of the piece for local key finding. The data used for calculating keys is usually aggregated data represented by chroma diagrams or pitch profiles as described in Section 2.4. A pitch profile is shown in table 2.2.

### 3.3.1 Existing Labeled Data

There are existing data sets that scientists have used to train and test their new methods for key finding. Because music analysis requires an expert to annotate, the already labeled data sets are few and far between. Because of the different genres and types of music in the data sets, tests of the same algorithm on different data sets can yield different results. The data sets listed below are the ones that have appeared the most in the literature reviewed in this thesis.

#### CCARH

The CCARH data set used by Albrecht and Shanahan [2013] and Nápoles López et al. [2019] is a selection of 982 pieces from the common-practice era of western classical music. The labeled data was generated at the Center for Computer Assisted Research in the Humanities at Stanford, CA (CCARH).[4] The distribution of pieces is across several composers and subgenres, providing a good foundation of music that mostly follows the harmonic rules of music theory.

#### Kostka-Payne

The *Kostka-Payne* (KP) key values are derived from harmony textbooks and workbooks (as well as a teacher's companion guide) [Kostka et al., 2018]. All examples used in the books are analyzed with keys, and often chords and functional labels. The music used as data has been extracted from the teacher's guide, complete with analyzed examples. The KP key values are extracted from the pitch frequencies described in Section 3.3. The data is also helpful because it shows how chord progression function during modulation and shows overlapping key areas.

[4] http://www.ccarh.org/

### Essen Folksong Collection

The *Aarden-Essen* (A-E) key values are derived from a collection of folk music from the whole world called the *Essen Folksong Collection.*[5] The data includes transcriptions of 6255 folk songs complete with pitch histograms, keyscapes, meter, and other information relevant to the region.

## 3.3.2 Generating Data

Scientists have found several ways to create data sets to test their methods. When testing algorithms that find the global key and are not overly concerned with modulations, Albrecht and Shanahan [2013], for example, assembled a data set based on the titles of classical pieces. Many composers of classical music have included the key of the piece in the title of their music, for example, Johann Sebastian Bach's *Menuet in G.* By making some assumptions, including that the first and last movements of a piece start and end in the global key, large data sets, relative to the ones analyzed by experts, can be assembled with a fair amount of accuracy. This data is helpful for testing algorithms for global keys but is not very useful in the training and testing of algorithms designed to detect local keys and modulations.

### MIREX

The MIREX audio key finding competitions use data available on their website.[6] In 2020, the data set contained 1252 audio clips from classical music. The data set has been generated using the method described above, using the key in the title of the piece and sampling the first 30 seconds.

## 3.3.3 Data Sets Created for MIR

Some scientists have seen the need for more available data for MIR research and developed projects to assemble large quantities of data that include key information. The data sets described below have information about the global key of each data entry.

### The Million Song Data Set

The Million Song Data Set assembled by Bertin-Mahieux et al. [2011] includes a lot of information about a million pop songs. The information relevant to key finding is the tempo, beat onset time, key of the piece, and the confidence of the

[5] http://kern.ccarh.org/browse?l=/essen

[6] https://www.music-ir.org/mirex/wiki/2005:Audio_and_Symbolic_Key

key classification. In order to do experiments on this data, it is possible to filter the songs based on the confidence of the key label.

### The KUSC Data Set

The KUSC data set is assembled using data from the public classical radio channel in the US, *KUSC* [Chuan and Chew, 2014].[7] The data set consists of 3000 entries of 15-second excerpts of classical compositions. The excerpts are taken from each piece's first and last 15 seconds and analyzed by experts to confirm the key. The data set also includes information about instrumentation, spectrograms, chromagrams, key value sets, and more.

[7] `https://www.kusc.org/`

# Chapter 4

# Architecture

The following chapter describes the architecture used to test the improvement of minor key classification in existing key finding techniques. Three experiments were completed to test the efficacy of modifications made to standard and state-of-the-art methods. After an introduction to the architecture, Section 4.1 presents the data and processing techniques used for the experiments, and Section 4.2.2 presents the key profiles used.

## 4.1 Data and Feature Extraction

The data used for the experiments is a collection of western classical music from the late baroque to the 20th century. The data set is the same as the one used by Albrecht and Shanahan [2013] and Nápoles López et al. [2019] and was chosen because both studies have state-of-the-art results. Because the data and the train/test splits are the same, it is possible to compare the studies and experimental results in this thesis with high accuracy. The data and the data processing are described in the following subsections.

### 4.1.1 The CCARH Data Set

The data set used by Albrecht and Shanahan [2013] and Nápoles López et al. [2019] is derived from labeled data developed at the Center for Computer Assisted Research in the Humanities at Stanford, CA (CCARH).[1] The data is a collection of scores in the kern format that have global keys annotated by researchers at the CCARH. Albrecht and Shanahan [2013] adapted the data for key finding purposes by excluding any data that was not explicitly in major or minor and

[1] http://www.ccarh.org/

Table 4.1: List of composers and compositions used in the data set.[2]

| Composer | Composition | Number of Entries (Movements Counted Separately) |
|---|---|---|
| Bach | The Brandenburg Concertos | 20 |
| | Chorales | 323 |
| | Sinfonias | 15 |
| | Inventions | 15 |
| | *Well-Tempered Clavier* (Preludes and Fugues) | 69 |
| Beethoven | First and last movements of all string quartets | 31 |
| | First and last movements of all piano sonatas | 61 |
| Brahms | Op. 51 | 3 |
| Chopin | Op. 28 | 24 |
| | Mazurkas | 48 |
| Corelli | Trio Sonatas, Op. 1 | 41 |
| Haydn | First and last movements of each string quartet | 108 |
| Hummel | Op. 67 Préludes | 24 |
| Kabalevsky | "Happy" variations on a folksong | 6 |
| Miscallaneous | Barbershop quartet arrangements | 31 |
| Mozart | First and last movements of each string quartet | 67 |
| Telemann | 3 klavier fantasies | 3 |
| Scarlatti, D. | 58 piano sonatas | 58 |
| Vivaldi | Op. 8 | 32 |

labeling the first and last movements of Haydn and Mozart string quartets as the key in the title while discarding most of the middle movements. The data set contains 974 entries as seen in Table 4.1, making it one of the largest collections of labeled classical data in a symbolic format. 619 of the pieces are in a major key and 355 pieces in a minor key.

The data set is also attractive due to the genres and eras in which the music was written. The pieces are all part of the western classical tradition and within the time known as the common practice era, where music mostly followed the rules of harmony with fewer deviations than music written before and after.

### 4.1.2 Feature Extraction

A few changes were made to the data in order to prepare it for the experiments. The number of steps that had to be taken were limited due to compatible file formats and software tools.

[2] Copied and adapted to fit the revised data set from a table by Albrecht and Shanahan [2013] and used with permission from University of California Press, License Number 1314333-1. The license grants permission to republish a chart/graph/table/figure in a thesis/dissertation.

### Pitch Profiles

The experiments use the pitch profiles of each data entry as the primary feature. In order to extract these, each data point was parsed into the music21 framework preceded by a count of each pitch, returning a distribution like the one seen for Twinkle Twinkle Little Star in Table 2.2. The length unit for the experiments is quarter notes.

### Musical Excerpts

In addition to the pitch distribution, a separate data set with only the first and last 8 bars of each piece present was created. This was done in order to be able to compare results from the experiments with the ones completed by several other researchers. Articles using only the piece's first and last eight bars are listed in Table A.1.

# 4.2 Key Profiles

Key profiles are an essential part of these experiments. The proposed model's unique feature is its key profile that divides into three variations for its minor key dimension. In order to test the proposed model's accuracy, a set of existing key profiles are included in the test for comparison.

## 4.2.1 Key Profiles for Comparison

The comparison key profiles were chosen due to their prominence in the key finding literature or high accuracy scores.

### Sapp's Simple Key Profile

As mentioned in Section 3.2.2, the Sapp key profile is based on ideas from music theory [Sapp, 2011]. The essential pitches get two points, the middle gets one point, and the non-diatonic pitches get 0. The profile is easy to understand and performs well in many scenarios. Its weakness is that it is very sensitive to non-diatonic pitches since the profile has no slack to give them.

### Krumhansl-Kessler Key Profile

The Krumhansl-Kessler (KK) key profile is included due to its status as the first significant contribution to the field of key finding Krumhansl [1990]. It is also the only key profile included in the experiments based on the human experience of keys. The key profile performs well if the tonality is even and predictable but is easily swayed to other keys in the face of non-diatonic pitches due to its generous values given to the chromatic pitches.

#### Kostka-Payne Key Profile

As presented in Section 3.3, the Kostka-Payne key values are based on the music from the Kostka-Payne textbook for music theory [Kostka et al., 2018]. The values usually perform well but rarely achieve the highest accuracy scores.

#### Aarden-Essen Key Profile

The Aarden-Essen key profile is based on the Essen collection of folksongs described in Section 3.3 [Aarden, 2003]. The profile is one of the models that achieves the highest accuracy scores. However, sometimes it can be hindered by its limited scope due to the data it is trained on only consisting of folk music. Since the diversity of its training data is lacking, it can be confused by more complex or different music.

#### Albrecht-Shanahan Key Profile

The Abrecht-Shanahan Key Profile is based on a subset of the CCARH data set presented in Section 4.1. The key profile performs on par with other state-of-the-art methods and performs better with the predicted label minor accuracy. Each experiment is completed once with the whole data set and once with the training set. This approach allows the key profile to be tested without the training data bias. This key profile might have an advantage despite the separate experiments because it has been trained on very similar data to the test set.

#### Bellmann-Budge Key Profile

The Bellmann-Budge key profile was developed by Bellmann based on a Ph.D. in chord frequencies in music from the 18th and 19th centuries by Budge. Bellmann used the data from Budge's work and developed the data-driven key profile for major and minor keys. The key profile works best for harmonically stable pieces and is a strong contender when analyzing common-practice classical music.

### 4.2.2 Key Profiles for Improved Minor Key Classification

The key profiles were developed in three steps, as shown in Figure 4.2. The values by Sapp [2011] were chosen as the base to create a basis for the key profiles. They are helpful because of their adaptability and basic structure rooted in music theory. The simple key profile performs well compared to other key profiles and does not require data sets labeled with minor variations or psychological experiments to tweak them. As presented in Section 3.2.2, Sapp designates the value 2 to the tonic and dominant, the most defining pitches of a key, 1 to all other diatonic scale degrees, and 0 for the remaining pitches. In order to adapt these to account for the minor variations, the three key variations were given new values for the sixth and seventh scale degrees. These key profiles are shown in Table 4.2a, where the natural and harmonic minors follow the pattern defined by Sapp, and the melodic minor has values of 0.5 where the pitches change in the ascending and descending scales.

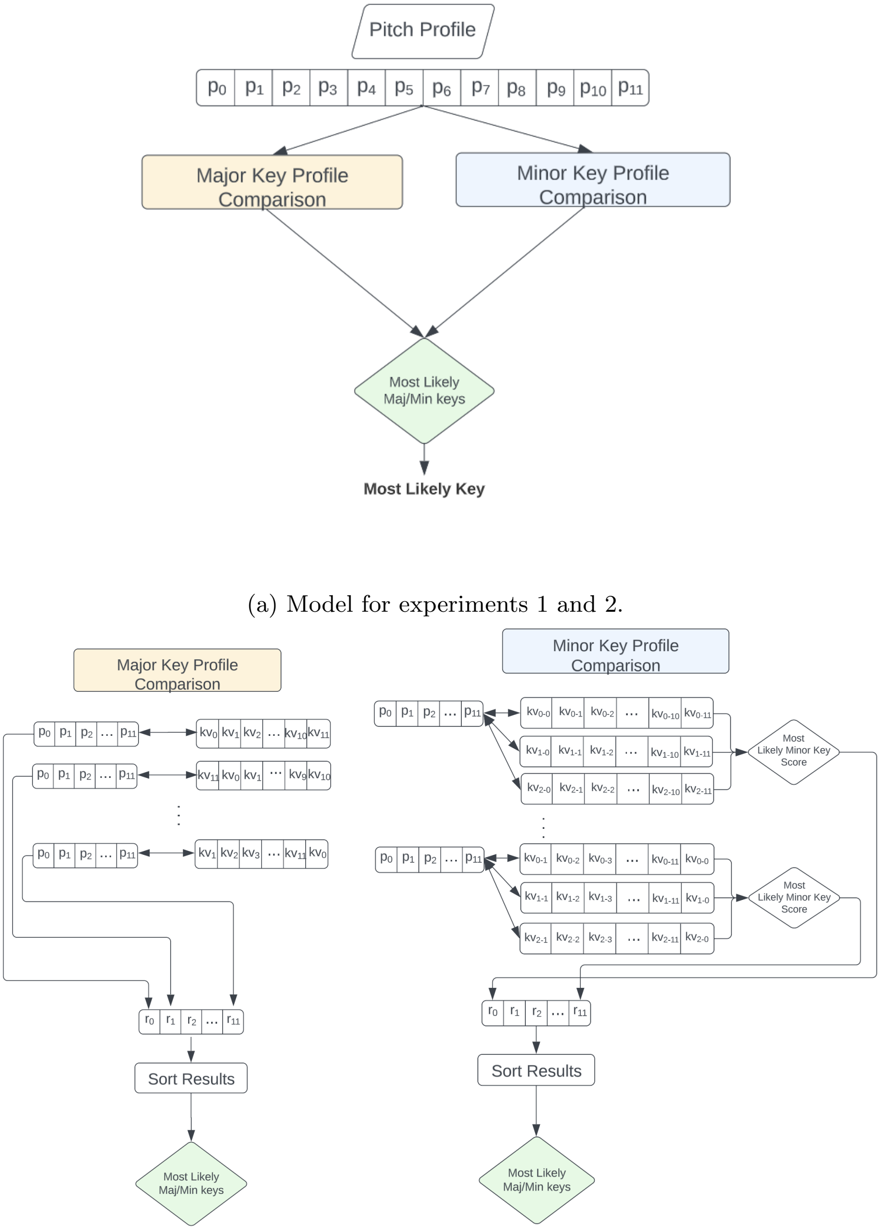
After developing a base for the new key profiles, a few adjustments were made to counteract two of the simple key profile's tendencies when detecting keys. Because all non-diatonic pitches are given a score of 0, the simple key profile has a very low tolerance for pitches not found in the key and can change its classification based on a few non-chord tones. As shown in Nápoles López et al. [2019], the simple key profile changes its key classification more times than any of the other commonly used key profiles. In order to create more tolerance for non-diatonic pitches, 0.1 was added to all non-diatonic values as shown in Table 4.2b

The last modification made to the key profile was to make the values more similar to the Aarden-Essen key profile by adding value to the third scale degree [Aarden, 2003]. This change led to the final iteration of the new and modified key profile shown in Table 4.2c.

## 4.3 Model for Key Profile Comparison Experiments

Experiments 1 and 2 have the structure of the KS algorithm, presented in Section 2.2.1, with a minor difference in the way the comparison between the pitch profile and key profile is calculated. The model for the experiments is shown in Figure 4.1.

The model compares the pitch profile to each variation of the key profile. This results in 24 comparisons for each pitch profile, one for each of the rotations of the key profiles. After storing each result, the results are sorted to find the highest correlation or shortest Euclidean distance. After comparing the scores for the top-scoring major and minor keys, a key is finally chosen and presented as the chosen key for the piece.


Pitch Profile
p0 p1 p2 p3 p4 p5 p6 p7 p8 p9 p10 p11
Major Key Profile Comparison
Minor Key Profile Comparison
Most Likely Maj/Min keys
Most Likely Key
Major Key Profile Comparison
p0 p1 p2 … p11
kv0 kv1 kv2 … kv10 kv11
kv11 kv0 kv1 … kv9 kv10
kv1 kv2 kv3 … kv11 kv0
r0 r1 r2 … r11
Sort Results
Most Likely Maj/Min keys
Minor Key Profile Comparison
kv0-0 kv0-1 kv0-2 … kv0-10 kv0-11
kv1-0 kv1-1 kv1-2 … kv1-10 kv1-11
kv2-0 kv2-1 kv2-2 … kv2-10 kv2-11
Most Likely Minor Key Score
kv0-1 kv0-2 kv0-3 … kv0-11 kv0-0
kv1-1 kv1-2 kv1-3 … kv1-11 kv1-0
kv2-1 kv2-2 kv2-3 … kv2-11 kv2-0
Most Likely Minor Key Score
r0 r1 r2 … r11
Sort Results
Most Likely Maj/Min keys

(a) Model for experiments 1 and 2.

(b) Major key comparison method.

(c) Minor key comparison method.

Figure 4.1: The model for experiments 1 and 2. The minor key profile comparison for pre-existing key profiles is equivalent to the major key comparison shown in (b).

Table 4.2: New key profile values for minor key variations.

(a) Initial values based on the Simple key profile by Sapp [2011].

| Pitch | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Natural | 2 | 0 | 1 | 1 | 0 | 1 | 0 | 2 | 1 | 0 | **1** | **0** |
| Harmonic | 2 | 0 | 1 | 1 | 0 | 1 | 0 | 2 | 1 | 0 | **0** | **1** |
| Melodic | 2 | 0 | 1 | 1 | 0 | 1 | 0 | 2 | **0.5** | **0.5** | **0.5** | **0.5** |

(b) Second iteration, adding value to non-diatonic pitches.

| Pitch | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Natural | 2 | **0.1** | 1 | 1 | **0.1** | 1 | **0.1** | 2 | 1 | **0.1** | 1 | **0.1** |
| Harmonic | 2 | **0.1** | 1 | 1 | **0.1** | 1 | **0.1** | 2 | 1 | **0.1** | **0.1** | 1 |
| Melodic | 2 | **0.1** | 1 | 1 | **0.1** | 1 | **0.1** | 2 | 0.5 | 0.5 | 0.5 | 0.5 |

(c) Final iteration, increasing the value of the third scale degree.

| Pitch | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Natural | 2 | 0.1 | 1 | **1.5** | 0.1 | 1 | 0.1 | 2 | 1 | 0.1 | 1 | 0.1 |
| Harmonic | 2 | 0.1 | 1 | **1.5** | 0.1 | 1 | 0.1 | 2 | 1 | 0.1 | 0.1 | 1 |
| Melodic | 2 | 0.1 | 1 | **1.5** | 0.1 | 1 | 0.1 | 2 | 0.5 | 0.5 | 0.5 | 0.5 |

## 4.4 Model for HMM Key Finding

Experiment 3 is based on the model designed by Nápoles López et al. [2019]. It uses hidden Markov models (HMMs) to predict both local and global keys by observing the sequence of pitches in the data. The HMM components described in Section 2.2.2 are treated as follows:

- Initial probability distribution: A uniform probability distribution across all 24 keys.

- Observed sequence of events: The sequence of pitches in the data. Each pitch is represented by a number in [0,11].

- The hidden states: Each of the possible 24 keys. Numbers in [0,11] represent major keys and [12, 23] represent minor keys.

- The transition probability distribution: A distribution based on a 2D Schoenberg grid denoting how 'close' two keys are. These values are computed from a matrix of neighboring keys and shown in Table 4.3.

- The emission probability distribution: Key profiles are used as probability distributions denoting how likely it is that a key emits a certain pitch.

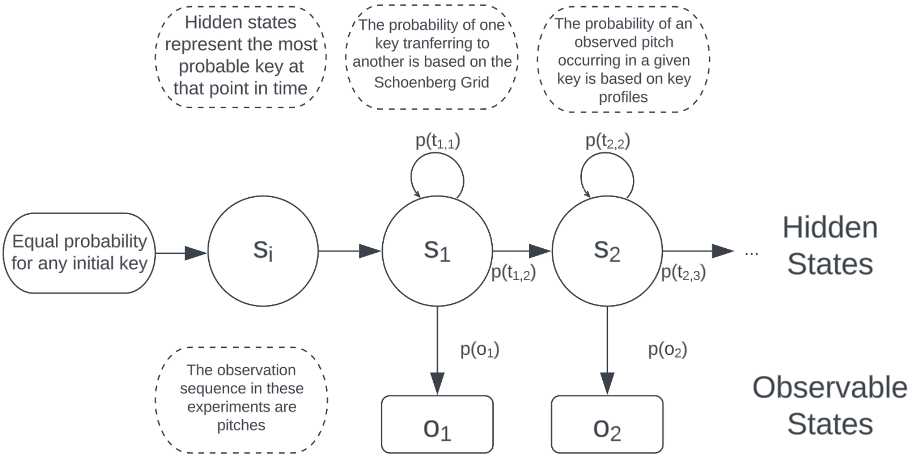

Hidden states represent the most probable key at that point in time
The probability of one key tranferring to another is based on the Schoenberg Grid
The probability of an observed pitch occurring in a given key is based on key profiles
p(t1,1)
p(t2,2)
Equal probability for any initial key
si
s1
s2
p(t1,2)
p(t2,3)
...
Hidden States
p(o1)
p(o2)
The observation sequence in these experiments are pitches
o1
o2
Observable States

Figure 4.2: The model for experiment 3.

Table 4.3: The distance to keys from C major.[3]

| Group | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| Keys | C | F | d | D | E | D♭ | e♭ | c♯ | F♯ |
| | | G | e | E♭ | A♭ | D♭ | f♯ | a♭ | |
| | | a | f | A | b♭ | B | | | |
| | | c | g | B♭ | b | | | | |

In order to include the expanded proposed key profile, each of the minor variations is tested and the highest score is chosen.

Figure 4.2 shows how the components interact with the HMM model for key finding. The proposed model includes the expanded key profile. When making its comparisons, the algorithm chooses the minor key variation that is the most similar pitch profile.

[3] Copied from Nápoles López et al. [2019] and used with permission from ACM Press, License Number 1314330-1. The license grants permission to republish a chart/graph/table/figure in a thesis/dissertation.

# Chapter 5

# Experiments and Results

This chapter contains the plans, the setup, and information about the experiments and their results. Section 5.1 explains what experiments were conducted, Section 5.2 includes all the information about the setup and technology used for each experiment, and Section 5.3 concludes the chapter with the results of each of the experiments. These results are discussed further in Chapter 6 and used in Section 7.1 to evaluate the goals and research questions established in Section 1.2.

## 5.1 Experimental Plan

The experiments aim to improve the classification of minor keys by differentiating the key profiles for natural, harmonic, and melodic minor keys. Each experiment group is based on foundational or state-of-the-art key finding algorithms.

### 5.1.1 Experiment 1: The Krumhansl-Schmuckler Algorithm

The first experiment is an adapted form of the Krumhansl-Schmuckler (KS) algorithm. The experiment results were used to see how the standard key finding method responded to the new minor key approach. They were also used as an initial confirmation that dividing the minor key profiles into ones specific to the common minor variation resulted in results similar to or better than other key profiles.

The experiment is based on the model in Figure 4.1. The pitch- and key profiles are compared by finding the correlation between the arrays. All the parameters for the experiment are presented in Section 5.2. The experiment was completed four times (1a-d) with different parameters to ensure comparability with results presented for similar experiments in literature.

### 5.1.2 Experiment 2: The Albrecht-Shanahan Model

The second experiment is closely related to the Albrecht-Shanahan (AS) model proposed by Albrecht and Shanahan [2013]. The difference between the two is that the comparisons between pitch- and key profiles are made by computing the Euclidean distance. In the article describing the model, only the key profile weights developed by Albrecht and Shanahan were included in the Euclidean distance experiments. In this experiment, several key profiles, including the ones proposed in this thesis are included in order to be able to compare the results to the ones from the other experiments.

Experiment 2 is based on the model in Figure 4.1. The parameters for the experiment are listed in Section 5.2. The experiment was completed four times (2a-d) with different parameters to ensure comparability with results presented for similar experiments in literature.

### 5.1.3 Experiment 3: Hidden Markov Model

Experiment 3 is an adapted version of the work done by Nápoles López et al. [2019]. The model is based on the code available via the Justkeydding tool presented in Section 2.5. In order to attempt to improve the classification of minor keys, the extra step of choosing the most likely minor key variation profile is added. The experiment is based on the model in Figure 4.2 and tested with several key profiles.

## 5.2 Experimental Setup

The ten experiments are divided into three groups. Each group contains several experiments with different parameters, shown in Table 5.3. All the experiments share some parameters, like the key profiles they test and the types of results they return. Table 5.1 shows the key profiles and result metrics used for all the experiments. The number of entries in each data set is listed in Table 5.2 and all the key profile values used in the experiments are listed in Appendix B.

Table 5.1: Key profiles and result metrics used for all the experiments. The key profiles are presented in Section 4.2.2 and the result metrics are presented in Section 2.3.

(a) Key profiles tested in each experiment.

| Key Profile |
|---|
| Proposed Model |
| Sapp's Simple Key Profile |
| Krumhansl-Kessler |
| Kostka-Payne |
| Aarden-Essen |
| Albrecht-Shanahan |
| Bellmann-Budge |

(b) Result metrics for each experiment.

| Result Metric |
|---|
| Ground Truth Minor Accuracy |
| Predicted Label Minor Accuracy |
| MIREX Score |
| Accuracy |
| MIREX Distribution |

Table 5.2: The number of entries in each of the data sets used for testing. The test set is included in the whole data set and is the same as used by Nápoles López et al. [2019] and Albrecht and Shanahan [2013].

| Data Set | Number of Entries |
|---|---|
| Whole Data Set | 974 |
| Test Set | 492 |

Table 5.3: An overview of the experimental setup.

| Model | Experiment | Data | Excerpt |
|---|---|---|---|
| Krumhansl-Schmuckler | 1a | Whole Set | Whole Piece |
| | 1b | Whole Set | Excerpt |
| | 1c | Test Set | Whole Piece |
| | 1d | Test Set | Excerpt |
| Albercht-Shanahan | 2a | Whole Set | Whole Piece |
| | 2b | Whole Set | Excerpt |
| | 2c | Test Set | Whole Piece |
| | 2d | Test Set | Excerpt |
| Hidden Markov Model | 3a | Whole Set | Whole Piece |
| | 3b | Test Set | Whole Piece |

## 5.3 Experimental Results

The following section summarizes the results of the experiments. Each experiment group's results are presented, mentioning interesting trends and values as they appear. Because several experiments have similar results, not all are discussed in this section. The complete set of results is listed in Appendix D, and the results are discussed further in Chapter 6.

### 5.3.1 Experiment 1

Experiment 1 consisted of running all the key profiles through the Krumhansl-Schmuckler algorithm. Table 5.4 shows that the proposed model scores highest in all categories except the predicted label minor accuracy category. The margins are small, and the differences are not statistically significant. However, it shows that the proposed model can perform as well as the existing key profiles and even contend among the best across several experiments with different data and feature-length parameters.

Almost all the key profiles have stable performances across the first group of experiments. The exceptions are the ground truth minor accuracy metric across all sub-experiments and the Krumhansl-Kessler (KK) profile's overall performance.

The KK profile shows a striking difference in how it performs when it analyzes a whole piece versus an excerpt. Tables 5.4 and 5.6 show this improvement between the experiments. The behavior is a symptom of the KK profile's tendency to prioritize non-diatonic pitches mentioned in Section 4.2.2. This jagged performance continues in various scales across all three experiment groups.

The ground truth minor accuracy metric shows a surprising variation across

Table 5.4: Results from experiment 1a

| Key Profile | Ground Truth<br>Minor Accuracy | Predicted Label<br>Minor Accuracy | MIREX | Accuracy |
|---|---|---|---|---|
| **Proposed Model** | **94.71** | 90.22 | **94.71** | **93.22** |
| **Sapp** | 94.30 | 90.80 | 94.30 | 92.40 |
| **Krumhansl-Kessler** | 85.09 | 87.50 | 85.09 | **76.39** |
| **Kostka-Payne** | **93.37** | 91.98 | 93.37 | 91.07 |
| **Aarden-Essen** | 93.96 | 87.90 | 93.96 | 91.99 |
| **Albrecht-Shanahan** | 94.59 | 91.06 | 94.59 | 92.71 |
| **Bellmann-Budge** | 85.88 | **93.25** | 94.16 | 92.30 |

Table 5.5: MIREX distribution for experiment 1a.

| | Correctly Classified | Dominant | Relative | Parallel |
|---|---|---|---|---|
| **Proposed Model** | 908 | 15 | 16 | 11 |
| **Sapp** | 900 | 18 | 23 | 13 |
| **Krumhansl-Kessler** | 744 | 157 | 13 | 12 |
| **Kostka-Payne** | 887 | 16 | 40 | 12 |
| **Aarden-Essen** | 896 | 19 | 25 | 11 |
| **Albrecht-Shanahan** | 903 | 23 | 14 | 13 |
| **Bellmann-Budge** | 899 | 9 | 36 | 14 |

Table 5.6: Results for experiment 1b.

| Key Profile | Ground Truth<br>Minor Accuracy | Predicted Label<br>Minor Accuracy | MIREX | Accuracy |
|---|---|---|---|---|
| **Proposed Model** | 90.96 | 90.96 | 94.93 | 93.53 |
| **Sapp** | 89.55 | 92.42 | 94.83 | 93.12 |
| **Krumhansl-Kessler** | 81.07 | 88.04 | 88.31 | **83.26** |
| **Kostka-Payne** | **84.75** | 92.31 | 93.89 | 92.09 |
| **Aarden-Essen** | 90.11 | 89.11 | 93.94 | 91.99 |
| **Albrecht-Shanahan** | 90.40 | 91.43 | 95.00 | 93.74 |
| **Bellmann-Budge** | 86.44 | 93.29 | 94.21 | 92.09 |

Table 5.7: MIREX distribution for experiment 1b.

| | Correctly Classified | Dominant | Relative | Parallel |
|---|---|---|---|---|
| **Proposed Model** | 911 | 11 | 17 | 15 |
| **Sapp** | 907 | 14 | 24 | 12 |
| **Krumhansl-Kessler** | 811 | 82 | 15 | 18 |
| **Kostka-Payne** | 897 | 11 | 32 | 12 |
| **Aarden-Essen** | 896 | 20 | 22 | 12 |
| **Albrecht-Shanahan** | 913 | 11 | 16 | 10 |
| **Bellmann-Budge** | 897 | 19 | 31 | 9 |

Table 5.8: Results for experiment 2a.

| Key Profile | Ground Truth Minor Accuracy | Predicted Label Minor Accuracy | MIREX | Accuracy |
|---|---|---|---|---|
| **Proposed Model** | **94.63** | 64.18 | 84.13 | **79.06** |
| **Sapp** | 87.57 | 91.45 | 94.16 | 92.20 |
| **Krumhansl-Kessler** | 87.57 | 66.81 | 78.89 | 70.53 |
| **Kostka-Payne** | 76.55 | 92.81 | 91.76 | 88.81 |
| **Aarden-Essen** | **88.14** | 92.31 | 94.10 | 91.99 |
| **Albrecht-Shanahan** | 86.16 | 93.85 | 93.83 | 91.48 |
| **Bellmann-Budge** | 87.29 | 91.42 | 94.00 | 92.20 |

Table 5.9: MIREX distribution for experiment 2a.

| | Correctly Classified | Dominant | Relative | Parallel |
|---|---|---|---|---|
| **Proposed Model** | **770** | 14 | **126** | 23 |
| **Sapp** | 898 | 18 | 25 | 13 |
| **Krumhansl-Kessler** | 687 | 126 | 54 | 11 |
| **Kostka-Payne** | 865 | 16 | 61 | 12 |
| **Aarden-Essen** | 896 | 21 | 26 | 11 |
| **Albrecht-Shanahan** | 891 | 25 | 26 | 13 |
| **Bellmann-Budge** | 898 | 9 | 31 | 19 |

the sub-experiments, given that every other metric remains relatively stable for almost all the key profiles. One example is how the results for the Kostka-Payne profile change from experiment 1a to 1b (Tables 5.4 and 5.6). The minor accuracy measurement dropped by nine percentage points when it transferred from analyzing whole pieces to excerpts. This occurred while the overall accuracy of the key profile increased. There is no indication of the reason for this in the MIREX distributions (Table 5.5 and 5.7), so further testing would have to be done to find the key profile's weakness.

### 5.3.2 Experiment 2

The second experiment was similar to experiment 1 but resulted in very different behavior from the proposed model. Table 5.8 and 5.10 show that all measurements are much lower than previously and that it mimics the KK profile's patterns discussed in the previous section. The proposed model performs worse on all metrics except the ground truth minor accuracy, where it scores six percentage points higher than the second-best.

The drop in accuracy coincides with the significant increase in relative key predictions. The MIREX distributions shown in Table 5.9, 5.11, 5.13, and 5.15 show the persistence of the skew toward relative keys and show that the correct

Table 5.10: Results for experiment 2b.

| Key Profile | Ground Truth Minor Accuracy | Predicted Label Minor Accuracy | MIREX | Accuracy |
|---|---|---|---|---|
| **Proposed Model** | 94.35 | 74.39 | 89.80 | **86.34** |
| **Sapp** | 88.14 | 93.41 | 94.72 | 92.92 |
| **Krumhansl-Kessler** | 86.16 | 78.61 | 87.04 | 82.24 |
| **Kostka-Payne** | 79.94 | 94.02 | 92.94 | 90.66 |
| **Aarden-Essen** | 88.98 | 88.24 | 93.23 | 91.17 |
| **Albrecht-Shanahan** | 82.77 | 94.52 | 93.73 | 91.89 |
| **Bellmann-Budge** | 86.44 | 93.29 | 94.18 | 92.09 |

Table 5.11: MIREX distribution for experiment 2b.

| | Correctly Classified | Dominant | Relative | Parallel |
|---|---|---|---|---|
| **Proposed Model** | 841 | 13 | **84** | 10 |
| **Sapp** | 905 | 15 | 25 | 13 |
| **Krumhansl-Kessler** | 801 | 74 | 26 | 10 |
| **Kostka-Payne** | 883 | 11 | 45 | 16 |
| **Aarden-Essen** | 888 | 21 | 22 | 15 |
| **Albrecht-Shanahan** | 895 | 12 | 27 | 19 |
| **Bellmann-Budge** | 897 | 19 | 30 | 9 |

classifications improve when they are for experiments with excerpts to analyze.

As predicted, the Albrecht-Shanahan profile performs as well as other state-of-the-art methods. It is especially consistent and high-scoring in the predicted label minor accuracy metric and the experiment contexts with excerpts and test data that it was explicitly designed to be used for (Table 5.12 and 5.14).

Table 5.12: Results for experiment 2c.

| Key Profile | Ground Truth Minor Accuracy | Predicted Label Minor Accuracy | MIREX | Accuracy |
|---|---|---|---|---|
| **Proposed Model** | 93.53 | 61.15 | 83.38 | 78.82 |
| **Sapp** | 86.47 | 90.18 | 94.30 | 92.67 |
| **Krumhansl-Kessler** | 87.65 | 63.40 | 77.74 | 69.45 |
| **Kostka-Payne** | 77.06 | 91.61 | 92.36 | 89.82 |
| **Aarden-Essen** | 87.06 | 92.50 | 94.44 | 92.46 |
| **Albrecht-Shanahan** | 84.71 | **92.90** | 94.22 | 92.46 |
| **Bellmann-Budge** | 87.06 | 89.16 | 93.87 | 92.26 |

Table 5.13: MIREX distribution for experiment 2c.

| | **Correctly Classified** | **Dominant** | **Relative** | **Parallel** |
|---|---|---|---|---|
| **Proposed Model** | 387 | 2 | **64** | 11 |
| **Sapp** | 455 | 6 | 12 | 7 |
| **Krumhansl-Kessler** | 341 | 60 | 31 | 7 |
| **Kostka-Payne** | 441 | 5 | 28 | 8 |
| **Aarden-Essen** | 454 | 9 | 12 | 8 |
| **Albrecht-Shanahan** | 454 | 7 | 13 | 6 |
| **Bellmann-Budge** | 453 | 3 | 14 | 11 |

Table 5.14: Results for experiment 2d.

| **Key Profile** | **Ground Truth<br>Minor Accuracy** | **Predicted Label<br>Minor Accuracy** | **MIREX** | **Accuracy** |
|---|---|---|---|---|
| **Proposed Model** | 93.53 | 74.65 | 90.24 | 86.97 |
| **Sapp** | 8765 | 93.13 | 94.79 | 93.08 |
| **Krumhansl-Kessler** | 85.29 | 80.11 | 87.72 | 83.10 |
| **Kostka-Payne** | 80.00 | 93.15 | 92.85 | 90.43 |
| **Aarden-Essen** | 87.06 | 87.57 | 92.83 | 90.63 |
| **Albrecht-Shanahan** | 83.53 | **94.04** | 94.03 | 92.46 |
| **Bellmann-Budge** | 86.47 | 92.45 | 94.42 | 92.46 |

Table 5.15: MIREX distribution for experiment 2d.

| | **Correctly Classified** | **Dominant** | **Relative** | **Parallel** |
|---|---|---|---|---|
| **Proposed Model** | 427 | 7 | **38** | 6 |
| **Sapp** | 457 | 8 | 10 | 7 |
| **Krumhansl-Kessler** | 408 | 37 | 10 | 6 |
| **Kostka-Payne** | 444 | 8 | 21 | 8 |
| **Aarden-Essen** | 445 | 12 | 10 | 9 |
| **Albrecht-Shanahan** | 454 | 6 | 11 | 7 |
| **Bellmann-Budge** | 454 | 9 | 13 | 6 |

Table 5.16: Results for experiment 3a.

| **Key Profile** | **Ground Truth Minor Accuracy** | **Predicted Label Minor Accuracy** | **MIREX** | **Accuracy** |
|---|---|---|---|---|
| **Proposed Model** | 86.44 | **90.53** | 92.74 | **90.04** |
| **Sapp** | 83.33 | **93.95** | 93.05 | 90.35 |
| **Krumhansl-Kessler** | 73.73 | **89.69** | 84.15 | **76.28** |
| **Kostka-Payne** | 85.88 | **92.97** | 92.90 | 90.04 |
| **Aarden-Essen** | 87.01 | **89.80** | 92.63 | 89.94 |
| **Albrecht-Shanahan** | 85.31 | **93.79** | 93.12 | 90.25 |
| **Bellmann-Budge** | 85.03 | **94.06** | 93.45 | 90.86 |

Table 5.17: Results for experiment 3b.

| **Key Profile** | **Ground Truth Minor Accuracy** | **Predicted Label Minor Accuracy** | **MIREX** | **Accuracy** |
|---|---|---|---|---|
| **Proposed Model** | 86.47 | **91.88** | 93.32 | 90.84 |
| **Sapp** | 81.76 | **93.92** | 93.22 | 90.43 |
| **Krumhansl-Kessler** | 74.71 | **88.81** | 84.18 | 76.17 |
| **Kostka-Payne** | 84.71 | **91.72** | 92.51 | 89.61 |
| **Aarden-Essen** | 87.06 | **89.70** | 93.22 | 90.63 |
| **Albrecht-Shanahan** | 86.47 | **93.63** | 93.91 | 91.45 |
| **Bellmann-Budge** | 85.29 | **92.95** | 93.71 | 91.24 |

### 5.3.3 Experiment 3

The two sub-experiments in experiment three had few surprising results. The proposed model performed as well as the state-of-the-art methods, and the only key profile that scored lower than the leading group was the KK profile. It is also worth noting that every key profile scored similarly on the predicted label minor accuracy metric for experiments 3a and 3b as shown in Table 5.16 and 5.17. This is the only metric that features a single group with no outliers across all experiments.

# Chapter 6

# Discussion

The results from the experiments shown in Chapter 5 show that the proposed method performs as well as state-of-the-art methods in several scenarios. The following chapter includes discussions about the results, what the results can tell us about the architecture used in the experiments, and how the proposed model compares to existing key finding algorithms.

## 6.1 Discussion of Results

To learn as much as possible from the results from Section 5.3, the data from each experiment must be seen in relation to other key profiles and the other experiments. Key takeaways are discussed below.

### 6.1.1 Result Comparisons

The proposed model returns results that are, in most cases, equal to the state-of-the-art methods included in the experiments for comparison. Figure 6.1 shows the accuracy metric across all the experiments. Except for the Krumhansl-Kessler key profile, the other pre-existing key profiles follow the same trajectory with only slight deviations. The proposed model also achieves state-of-the-art results for experiments one and three. The drop in accuracy for experiment two and how it relates to the KK accuracy is most likely rooted in how easily the key profile algorithms change their prediction in the face of non-diatonic pitches. This assumption is discussed further later in this and the following subsections.

The MIREX scores shown in Figure 6.2 are similar to the accuracy graph in Figure 6.1. This similarity is not surprising, given that the most important factor in the MIREX score calculation considers how many keys were correctly classified.

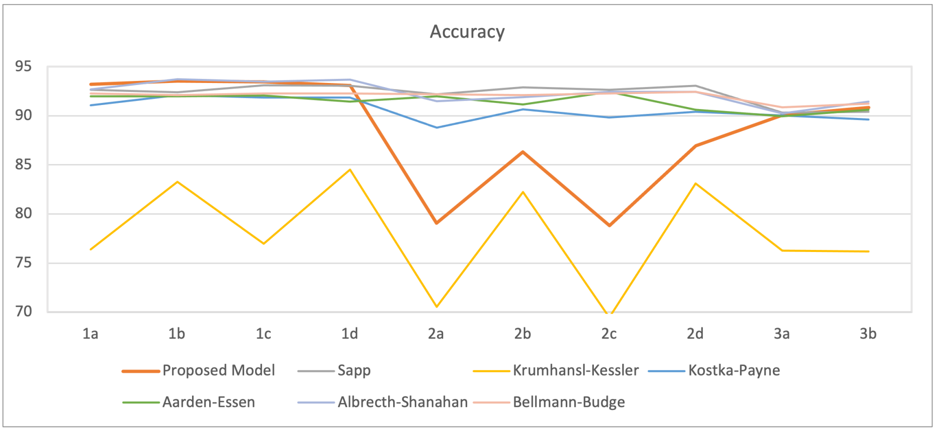


Figure 6.1: The accuracy of each key profile across all the experiments

Some differences between the key profile MIREX scores change compared to the accuracy metric in Figure 6.1. Since the MIREX score is a more holistic measurement, if the score improved relative to the other key profiles in MIREX vs. accuracy, the misclassified instances most likely belong to closely related keys. These small changes do not significantly impact the relationships between the key profile results.

The reason for these low scores can be found in the second minor graph in Figure 6.3 and the MIREX distributions listed in Section 5.3. The MIREX distributions show a prominent skew toward relative keys in the second experiment. The method in the second experiment utilizes Euclidean distance to measure key and pitch profiles from the data, which is discussed more in Section 3.2.4. Still, when talking about the results, we can assume that this is the primary reason the accuracy rates drop for the proposed model in experiment 2. The MIREX distribution shows that the skew toward relative keys grows significantly in the second experiment, which sheds light on the minor accuracy graphs. The graph in Figure 6.4 shows that when the ground truth mode is minor, the proposed model has a high probability of classifying the correct key. The corresponding decrease in the first minor measurement indicates that the minor key profiles get a higher score than the corresponding major key profile and chooses to classify the key as the relative minor of the pair. This is bad because it means that the relative weighting of the major and minor key profiles is skewed, but the good news is that the model prioritizes the correct set of pitches. The specifics of the skew towards the relative keys are discussed below.

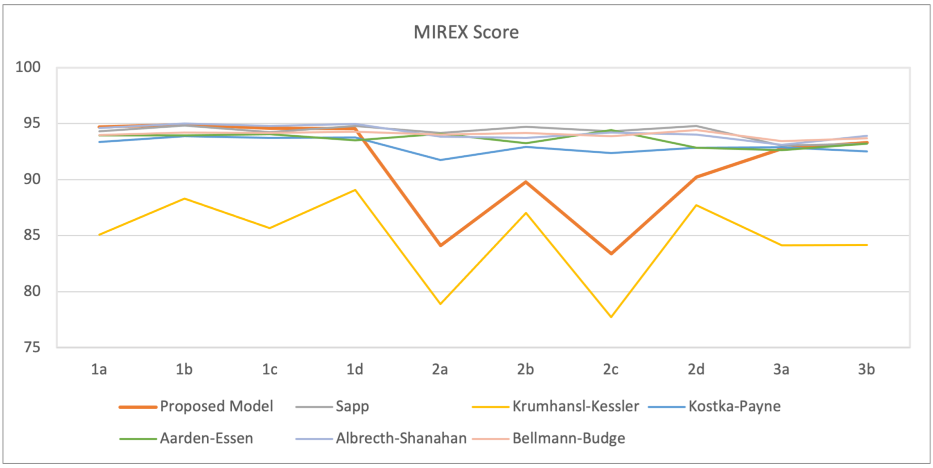


Figure 6.2: The MIREX score for each key profile across all experiments.

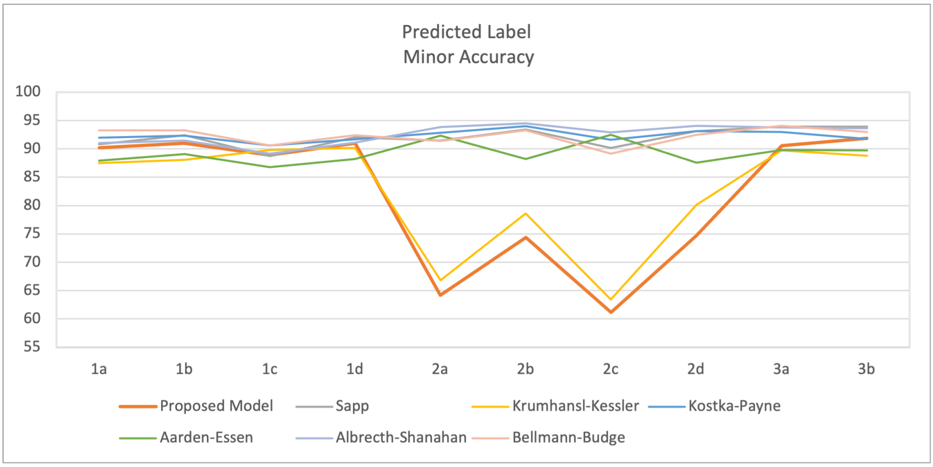


Figure 6.3: The predicted label minor accuracy for each key profile across all experiments.

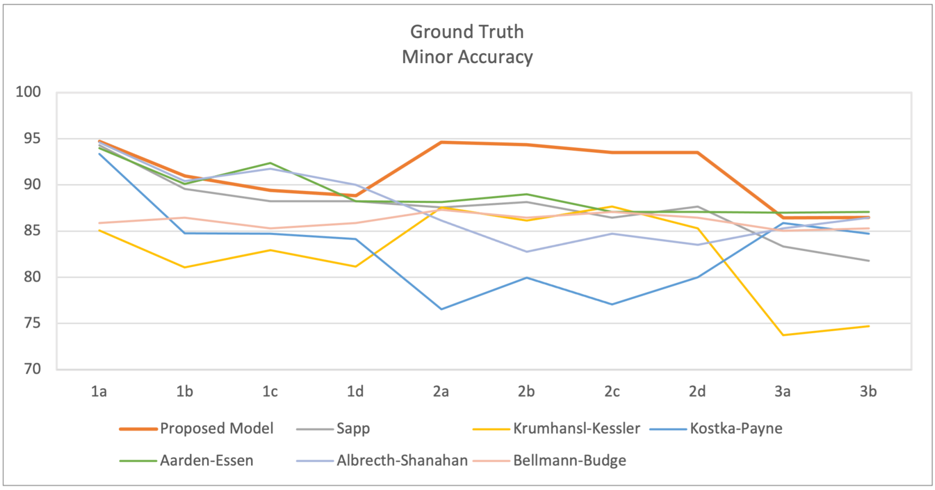


Figure 6.4: The accuracy for pieces with a minor ground truth mode for each key profile across all experiments.

### 6.1.2 The Skew Towards Relative Keys

The high accuracy scores for pieces with a ground truth mode of minor displayed in Figure 6.4 show us that the method works better than all other state-of-the-art methods at classifying minor keys in experiment 2. This is partly because it classifies many major keys as their relative minor and because other key profiles tend to give higher scores to major keys. This tendency to promote major keys exists because of a few factors. First, there are many more pieces in the data used by researchers to develop data-driven key profiles. This leads to minor key profiles that are less developed and less specific than their major counterparts. Second, because there are many variations of minor keys, most key profiles tend to simplify or combine variations to make the minor key profiles. The simplified version used in most key profiles smooths out the differences between or disregards the minor variations and therefore creates vague key profiles that are not preferred by the algorithm when placed next to the major profiles.

### 6.1.3 Measuring Adaptability

To investigate the adaptability of the proposed model, the results from sub-experiments with whole pieces versus excerpts are interesting. In the context of these experiments, excerpts are easier to analyze because they are taken from the start and end of pieces. The beginning and ends are often written in the global key without many non-diatonic pitches. In Figure 6.5, every key profile performs about the same or better when analyzing excerpts. Because the proposed model aims to predict the global key, the goal is to have results that are as stable and accurate for the whole piece as when only excerpts are analyzed. The model performs well in the first experiment shown in Figure 6.5a but exhibits much greater variety in the second experiment shown in Figure 6.5b.

## 6.2 Discussion of Architecture

The experiments revealed strengths and weaknesses in the architecture of the models tested. By discussing the aspects that worked well and the parts that should be reconsidered and improved, researchers can build future work on the knowledge gained from this thesis.

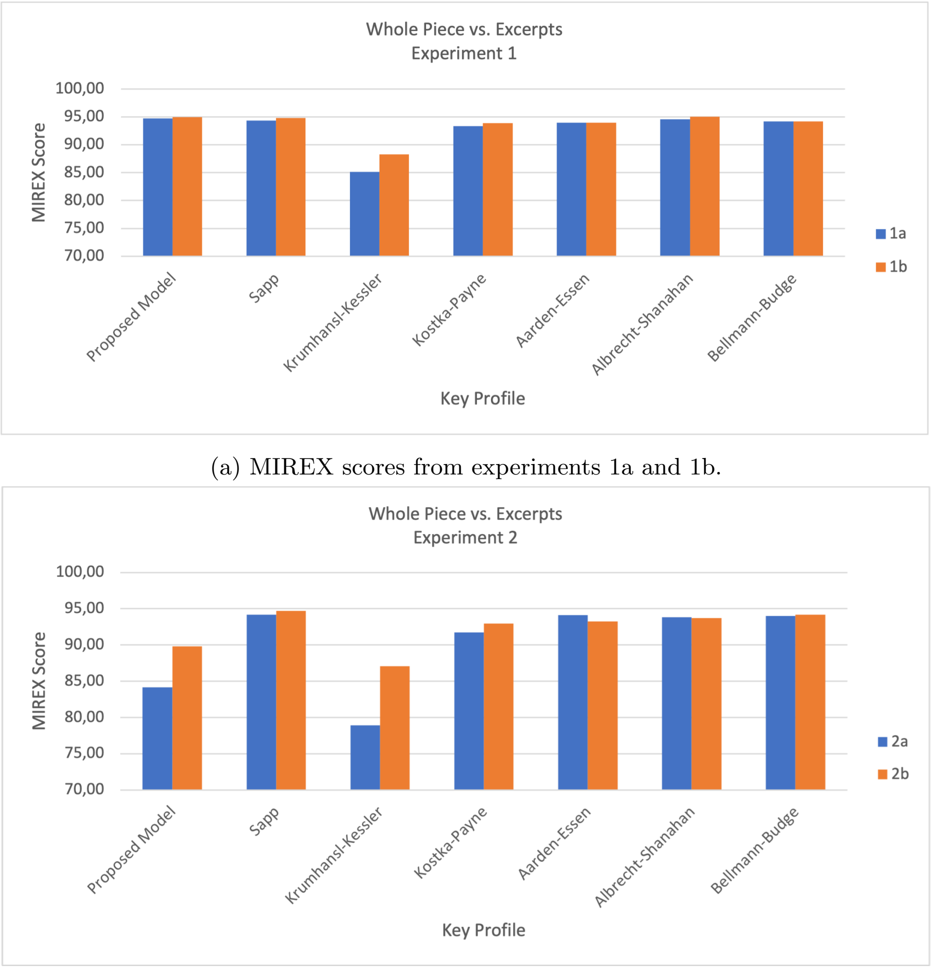
### 6.2.1 Key profiles

The main difference between the proposed model and the pre-existing key finding methods is the division of minor keys into three separate profiles. The minor key accuracy results from the second experiment show that in a context where the margins between key profiles are small, the proposed model is more successful at differentiating the minor keys than existing key profiles. As discussed in Section 6.1.2, there are two important reasons why the key profile performs better at classifying minor keys than other profiles: It has a skew towards minor keys and adds specificity to the key profile that traditional key profiles lack. It might be possible to minimize the skew by rebalancing the weightings of the major/minor key profiles in the proposed model to even them out. Once the number of relative key classifications drops, it is possible to see the effect of the three-parted minor key profile.

The idea behind dividing the minor key profile into three derives from the music theory behind the minor key variations. While it is promising to see the minor key accuracy scores from experiment 2 (Figure 6.1), the results mean little if the model is not contributing to more accurate classifications overall. By completing more experiments on the topic, it might be possible to retune the values in the model to perform better with the same basic three-parted idea.

### 6.2.2 Distance Measurement

The difference between the results from experiments 1 and 2 shows how critical the way distance is measured is. The architecture and key profiles remained the same, but the success of the proposed model drastically declined. The theory discussed in previous paragraphs is that the Euclidean distance measurement



(a) MIREX scores from experiments 1a and 1b.

(b) MIREX scores from experiments 2a and 2b.

Figure 6.5: A comparison of MIREX scores from experiments with whole pieces as its input (1a, 2a) and experiments analyzing excerpts (1b, 2b).

placed the major and minor key profiles in the proposed model much closer to each other than the correlation measurement and introduced a skew toward the minor keys. The maximization of the minor key scores worsens the skew. Future experiments might address this by rebalancing the weightings between major and minor.

The distance measurement used in experiment three does not have enough similar experiments to compare the distance measurement, but it works reasonably well with the proposed model. In the future, this distance measurement can be tested by using it as a key profile in experiments similar to 1 and 2 and having key profiles included as transition matrices in experiments equivalent to 3.

### 6.2.3 Analyzing Excerpts

An essential aspect of the experiments was the difference between the analysis of whole pieces and excerpts taken from the beginning and end of pieces. The results were as predicted, almost equal or better when analyzing excerpts. The different parameters were good for gaining insight into the proposed model and showing how models with diverse goals can use different inputs to achieve their goals. For global key finding, it is possible to argue that only excerpts should be used because they provide slightly better results. On the other hand, global key finding, where the method can extract the most prominent key from the whole piece, is more resistant to opening or closing passages in another key. The results from the experiments are mostly very similar, whether they analyze whole pieces or excerpts, so it might be better to use whole pieces when classifying global keys.

## 6.3 Discussion of Experiments

The experiments executed as a part of this thesis provide a good look into the behaviors of the proposed model in different contexts. Even though the first experiment was added as a test to see if the proposed model was viable, it was the set of experiments that performed the best overall. The closely related second experiment was beneficial in comparing the results and the methods, making it possible to learn more about the proposed model. The third experiment, which had the highest scores of any state-of-the-art method with the same general parameters, performed surprisingly poorly with the proposed model. All three experiments have in common that they can be optimized even further by tweaking parameters, key profiles, and probability distributions.

## 6.4 Data

The data used in the three experiments is optimized for western classical music key finding. Because the music in the data is from the common-practice era, most pieces follow the harmonic rules from music theory and should be relatively easy to classify. Because the results from experiments 1 and 3 were reasonably good, adding additional music to the data set might be beneficial to test the proposed model further.

Several experiments only contained a subset of the data to directly compare to the work by Albrecht and Shanahan [2013]. It was essential to have the correct Albrecht-Shanahan (AS) values to compare to (not running the AS algorithm on its training data), but it gave little insight beyond this comparison.

# Chapter 7

# Evaluation and Conclusion

The final chapter concludes the thesis, and the work is evaluated by revisiting the goals and research questions presented in Section 1.2. The chapter ends with suggestions for future work as extensions to this work and key finding in general.

## 7.1 Evaluation

This thesis was shaped around the goals and research questions presented in Section 1.2. Each research question is reviewed and discussed to evaluate the work, and finally, the primary goal is revisited.

**Research question 1** *How should key profiles be adapted to accommodate minor key variations?*

This question was explored through music theory and discussed as an integral part of the experimental architecture. The adapted key profile materialized when answering this question and is at the center of the proposed model. The results show that it differentiated the minor keys in specific contexts much better than other established key profiles but still needs to be worked on to achieve higher overall accuracy.

**Research question 2** *What state-of-the-art method(s) are most likely to improve with the minor mode modifications?*

In order to answer this question, a structured literature review was completed as described in Section 3.1. The review focused on research topic, experiment parameters, data, and results to find the most important methods and state-of-the-art approaches. Several articles were discussed in Section 3.2, and three methods were chosen as the foundations for the three experiments. The three selected papers all reported either abnormally good or bad results for minor key classification. While none of the methods were significantly improved, they provided a good canvas for learning more about the proposed algorithm.

**Research question 3** *What data works best for musical key finding, and should it be changed to accommodate minor modes?*

The data used in the experiments is a collection of classical pieces from the common-practice era. The data was chosen for two reasons. The data is the same as the set used by both Nápoles López et al. [2019] and Albrecht and Shanahan [2013], and the research done while exploring research question 2 showed that western classical music is the most precise data to use for key finding [Temperley and de Clercq, 2013]. No changes were made to the data to accommodate minor key finding. While it would be great to have a data set with labels for each minor key variation present, generating such a data set is a time-consuming task and has to be done by experts in the field of music theory. One adaptation that could be done was to increase the percentage of minor key pieces in the set. This was not done for these experiments so that it would be possible to directly compare the proposed model to the state-of-the-art methods using the same data set.

**Research question 4** *What is the best way to measure success?*

Different ways of measuring success in computational key finding were discussed in Section 2.3. The metrics ultimately chosen for the experiments were accuracy, MIREX scoring, minor accuracy, and minortest. Accuracy and the MIREX score were important markers of how well the model performs and as comparison values to other results. These two metrics, combined with the results for minor keys, gave a well-rounded picture of the proposed model's nature.

**Goal** *Improve the accuracy of minor key classification by combining foundational and state-of-the-art global key classification algorithms with a new key profile and improved data sets.*

The results of experiment 2a also show that when the margins are small enough, the model differentiates minor keys better than other key profiles. This element has to be tested more to make a factual claim. Hopefully, the state-of-the-art results in experiments 1 and 3, combined with the improved handling of minor keys in experiment 2a, indicate that it might be possible to improve the method to balance major and minor keys correctly and still keep its edge for classifying minor keys.

## 7.2 Conclusion

Computational key finding is still a significant obstacle in music information retrieval. The focus on minor keys in this thesis's research and experiments emphasizes a challenging part of the key finding field. The main contributions include a new model for handling minor keys that performs as well as state-of-the-art methods, a comparison of the significant key profiles used in articles from the past 20 years ans how they perform in different experimental contexts, and a structured literature review of the field of computational key finding.

In conclusion, this thesis explored the challenge of computational key finding by adapting existing key finding methods to improve their classification of minor keys. The proposed model shows promising results, for example, a 93.3% accuracy (Table 5.4) and a ground truth minor accuracy of 94.35% (Table 5.8). By splitting the minor key profile into three distinct key profiles correlating to the significant minor variations, the model brings more specificity to the individual minor pitch profiles than existing key profiles. These promising results show that the model still needs work to improve the overall accuracy while taking full advantage of the improved handling of minor keys.

## 7.3 Future Work

There are many ways the ideas of this work and computational key finding, in general, can be improved. As discussed in Section 6.2.1, the proposed model in this thesis shows promising results but needs more work and tests to improve its overall accuracy and quantify the improvements made to minor key detection more rigorously. One way to do this is to reexamine the key profile values used in the proposed model and rebalance them so they can perform better. This might be done by applying musical theory and trial, and error testing as Temperley [1999] or by training the key profile on data like Albrecht and Shanahan [2013].

Because the model performs reasonably well relative to the state-of-the-art methods it was tested against, another dimension to consider is the type of data it was tested on. The data used in this thesis was from the common-practice era of western classical music. Adding data from other genres will challenge the algorithm by introducing new harmonic structures and patterns that are more relaxed about following the rules set in place by music theory. As mentioned in Section 7.1, another option to improve the data is to use data labeled with each minor key variation. A final data adaptation is to increase the percentage of minor key pieces in the set to increase the amount of training data and provide more information about minor key characteristics.

Future work in the more general area of computational key finding could take many directions. As different research groups produce more labeled data, new possibilities open up. With larger data sets and more high-quality data, the efficacy of many machine learning algorithms can increase drastically. With the advent of labeled data that includes more features than the pitches and their durations, researchers could focus on developing algorithms that review more complex musical information, such as timing, meter, and the musical context within the piece. With more information available and more data to train on, the expansion into algorithms that can detect all twelve modes and other complex musical structures can be developed.

Because global key finding algorithms have improved dramatically in the past decade, localized key finding is a natural next area to research. As shown by Sapp [2011], the fundamental global key finding algorithms can be adapted and fine-tuned to classify local keys. This work will be more accessible once more data is available that is labeled with local keys. Localized key finding is helpful for many applications, such as music transcription, where it is important to accurately identify the tonality of different sections of a piece to generate correct notation. It could also be used in music analysis and musicological research to enable a more detailed examination of the tonal structure of a piece of music. Additionally, localized key finding algorithms could be used in music education to help students learn about the structure of music and how tonality changes within a piece.

# Bibliography

Aarden, B. (2003). *Dynamic Melodic Expectancy.* Ohio State University.

Albrecht, J. and Shanahan, D. (2013). The use of large corpora to train a new type of key-finding algorithm. *Music Perception: An Interdisciplinary Journal*, 31:59–67.

Bellmann, H. (2006). About the determination of key of a musical excerpt. In Kronland-Martinet, R., Voinier, T., and Ystad, S., editors, *Computer Music Modeling and Retrieval*, pages 76–91, Berlin, Heidelberg. Springer Berlin Heidelberg.

Bertin-Mahieux, T., Ellis, D. P., Whitman, B., and Lamere, P. (2011). The million song dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*, Miami, Florida.

Bill, J. P. (2021). File:circle of fifths deluxe 4.svg. [Online; accessed 29-June-2022].

Budge, H. (1943). *A Study of Chord Frequencies Based on the Music of Representative Composers of the Eighteenth and Nineteenth Centuries.* PhD thesis, Columbia University.

Catteau, B., Martens, J.-P., and Leman, M. (2006). A probabilistic framework for audio-based tonal key and chord recognition. In *Advances in Data Analysis, Proceedings of the 30th Annual Conference of the Gesellschaft für Klassifikation e.V.*, pages 637–644, Freie Universität Berlin.

Chuan, C.-H. and Chew, E. (2014). The KUSC classical music dataset for audio key finding. *The International Journal of Multimedia & Its Applications*, 6:1–18.

Cuthbert, M. and Ariza, C. (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 2010*, pages 637–642, Utrecht, Netherlands.

Fauvel, J., Flood, R., and Wilson, R. (2003). *Music and Mathematics.* Oxford University Press.

Forney, G. (1973). The Viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–278.

Kofod-Petersen, A. (2018). How to do a structured literature review in computer science. Technical report, Department of Computer and Information Science, Faculty of Information Technology, Mathematics and Electrical Engineering, Norwegian University of Technology and Science (NTNU), Trondheim, Norway.

Korzeniowski, F. and Widmer, G. (2018). Genre-agnostic key classification with convolutional neural networks. In Gómez, E., Hu, X., Humphrey, E., and Benetos, E., editors, *Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR 2018, Paris, France, September 23-27, 2018*, pages 264–270.

Kostka, S., Payne, D., and Almén, B. (2018). *Tonal Harmony: With an Introduction to Post-tonal Music.* College Ie Overruns. McGraw-Hill Education.

Krumhansl, C. L. (1990). *Cognitive Foundations of Musical Pitch.* Oxford University Press, 1 edition.

Lee, K. (2008). A system for automatic chord transcription from audio using genre-specific hidden Markov models. In Boujemaa, N., Detyniecki, M., and Nürnberger, A., editors, *Adaptive Multimedia Retrieval: Retrieval, User, and Semantics*, pages 134–146, Berlin, Heidelberg. Springer Berlin Heidelberg.

Lee, K. and Slaney, M. (2008). Acoustic chord transcription and key extraction from audio using key-dependent hmms trained on synthesized audio. *Audio, Speech, and Language Processing, IEEE Transactions on*, 16:291 – 301.

Lerdahl, F. (2001). *Tonal Pitch Space.* Oxford University Press, 1 edition.

Nápoles López, N., Arthur, C., and Fujinaga, I. (2019). Key-finding based on a hidden Markov model and key profiles. In *Proceedings of the 6th International Conference on Digital Libraries for Musicology*, DLfM '19, New York, NY, USA. ACM.

Özgúr Izmirli (2007). Localized Key Finding from Audio Using Nonnegative Matrix Factorization for Segmentation. In *Proceedings of the 8th International Conference on Music Information Retrieval*, pages 195–200, Vienna, Austria. ISMIR.

Papadopoulos, H. and Peeters, G. (2012). Local key estimation from an audio signal relying on harmonic and metrical structures. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(4):1297–1312.

Peeters, G. (2006). Musical key estimation of audio signal based hidden Markov modeling of chroma vectors. In *International Conference on Digital Audio Effects (DAFx-06)*, Montreal, Canada.

Purwins, H., Blankertz, B., and Obermayer, K. (2007). 5 toroidal models in tonal theory and pitch- class analysis. *Computing in Musicology (Tonal Theory for the Digital Age )*, 15:73–98.

Rameau, J.-P. (1779). *A Treatise of Music, Containing the Principles of Composition.* Luke White, Dublin, 2 edition. Translated to English from the original in the French language.

Raphael, C. and Stoddard, J. (2003). Harmonic analysis with probabilistic graphical models. In *ISMIR 2003 : Proceedings of the fourth International Conference on Music Information Retrieval : October 26-30, 2003, Baltimore, Maryland, USA*.

Sapp, C. S. (2011). *Computational Methods for the Analysis of Musical Structure.* PhD thesis, Stanford University.

Schoenberg, A. and Stein, L. (1969). *Structural functions of harmony.* Benn London, 2nd (revised) ed.; edited by Leonard Stein. edition.

Shmulevich, I. and Yli-Harja, O. (2000). Localized key-finding: Algorithms and applications. *Music Perception: An Interdisciplinary Journal*, 17(1):531–544.

Temperley, D. (1999). What's key for key? The Krumhansl-Schmuckler key-finding algorithm reconsidered. *Music Perception: An Interdisciplinary Journal*, 17(1):65–100.

Temperley, D. (2002a). A bayesian approach to key-finding. In Anagnostopoulou, C., Ferrand, M., and Smaill, A., editors, *Music and Artificial Intelligence*, pages 195–206, Berlin, Heidelberg. Springer Berlin Heidelberg.

Temperley, D. (2002b). A bayesian approach to key-finding. In Anagnostopoulou, C., Ferrand, M., and Smaill, A., editors, *Music and Artificial Intelligence*, pages 195–206. Springer Berlin Heidelberg, Berlin, Heidelberg.

Temperley, D. and de Clercq, T. (2013). Statistical analysis of harmony and melody in rock music. *Journal of New Music Research*, 42(3):187–204.

# Appendix A

# Articles from the SLR

Table A.1: Articles from the Structured Literature Review. Articles found outside of the review are not included in this table.

| Reference | Name | Style | # Excerpts | Length | Metric | Max Score | Scope |
|---|---|---|---|---|---|---|---|
| Albrecht and Shanahan (2013) | The use of large corpora to train a new type of key-finding algorithm: an improved treatment of the minor mode | Classical | 982 | Whole Piece, Excerpts | Accuracy | 93.1 | Global |
| Ariza and Cuthbert (2010) | music21: A Toolkit for Computer-Aided Musicology and Symbolic Music Data | NA | NA | NA | NA | NA | NA |
| Ariza and Cuthbert (2011) | Analytical and Compositional Applications of a Network-Based Scale Model in music21 | NA | NA | NA | NA | NA | NA |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Baumann (2021) | Deeper convolutional neural networks and broad augmentation policies improve performance in key estimation | Popular | 5400 | Excerpts | MIREX | 74.5 | Global |
| Bernardes, Davies, and Guedes (2017) | Automatic Musical key estimation with adaptive mode bias | Beatles, Popular | 879 | Excerpts | Accuracy | 67.9 | Global |
| Catteau, Martens, and Leman (2006) | A Probabilistic Framework for Audio-Based Tonal Key and Chord Recognition | Classical, Popular | 270 | Whole Piece | Accuracy | 83 | Local |
| Chen and Su (2019) | Harmony transformer: incorporating chord segmentation into harmony recognition | Popular | 64 | Whole Piece | Accuracy | 78.35 | Global |
| Chuan (2013) | A temporal multi-view approach for audio key finding using adaboost | Classical | 2785 | Excerpts | Accuracy | 80 | Global |
| Chuan and Chew (2014) | The KUSC classical music dataset for audio key finding | Classical | 3000 | Excerpts | NA | NA | Global |
| Dixon, Mauch, and Anglade (2011) | Probabilistic and logic-based modelling of harmony | Beatles, Jazz | 424 | Whole Piece | Accuracy | 81 | Global |
| Finley and Razi (2019) | Musical key estimation with unsupervised pattern recognition | Popular | 10000 | Whole Piece | Accuracy | 85 | Global |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Foscarin, Audebert, and S'niehotta (2021) | PKSPELL: data driven pitch spelling and key signature estimation | Classical | 222 | Whole Piece | Accuracy | 90.3 | Global |
| Gebhardt and Margraf (2017) | Applying Psychoacoustics to Key Detection and Root Note Extraction in EDM | EDM | 68 | Excerpts | Accuracy | 57.35 | Global |
| Gebhardt, Lykartsis, and Stein (2018) | A confidence measure for key labelling | Popular, Rock, Classical, EDM | 834 | Whole Piece | Accuracy | 75.18 | Local |
| George, Mary, and George (2022) | Development of an intelligent model for musical key estimation using machine learning techniques | Classical, Folk Music | 3243 | Whole Piece | Accuracy | 91.49 | Global |
| Giorgi, Zanoni, Sarti, and Tubaro (2013) | Automatic chord recognition based on the probabilistic modeling of diatonic modal harmony | Popular, Rock | 62 | Whole Piece | Accuracy | 70.5 | Global |
| Hu and Saul (2009) | A probabilistic topic model for unsupervised learning of musical key-profiles. | Classical | 235 | Whole Piece | Accuracy | 79 | Global |
| Izmirli (2006) | Audio Key Finding Using Low-Dimensional Spaces | Classical | 152 | Excerpts | MIREX | 88.9 | Global |
| Izmirli (2007) | Localized key finding from audio using negative matric factorization for segmentation. | Classical | 169 | Excerpts | Accuracy, MIREX | 88.3, 92.8 | Local |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Kania and Kania (2019) | A key-finding method based on music signature | Classical | 72 | Excerpts | Accuracy | 91.67 | Global |
| Kania, Kania, and Lukaszewicz (2021) | Trajectory of fifths in music data mining | Classical, Popular | 242 | Excerpts | NA | NA | Global |
| Kania, Kania, and Lukaszewicz (2021) | A Hardware-oriented Algorithm for Real-time Music Key Signature Recognition | Classical | 122 | Excerpts | Accuracy | 96 | Global |
| Korzeniowski and Widmer (2018) | Genre-agnostic key classification with convolutional neural networks | EDM, Popular, Rock, Classical | 3732 | Excerpts | Accuracy, MIREX | 67.9, 74.6 | Global |
| Lee and Slaney (2007) | A unified system for chord transcription and key extraction using hidden Markov models | Beatles | 28 | Whole Piece | Accuracy | 84.62 | Global |
| Lee and Slaney (2008) | Acoustic chord transcription and key extraction from audio using key dependent Hmms trained on synthesized audio | Classical, Beatles | 923 | Whole Piece | Accuracy | 94.69 | Local |
| Lerdahl (1988) | Tonal Pitch Space | NA | NA | NA | NA | NA | NA |
| Lin and Yeh (2017) | Automatic Chord Arrangement with Key Detection for Monophonic Music | Hymns | 195 | Whole Piece | Accuracy | 60.38 | Local |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Lindenbaum, Yeredor, and Cohen (2014) | Musical Key Extraction using diffusion maps | Beatles | 179 | Whole Piece | Accuracy, MIREX | 66.5, 75.6 | Global |
| Marquez (2019) | A chord distance metric based on the Tonal Pitch Space and a key-finding method for chord annotation sequences | Popular | 240 | Whole Piece | Accuracy | ca 81 | Global |
| Mauch and Dixon (2010) | Simultaneous estimation of chords and musical context from audio | Beatles | 176 | Whole Piece | MIREX | 71 | Local |
| McLeod and Rohrmeier (2021) | A modular system for the harmonic analysis of musical scores using a large vocabulary. | Classical | 742 | Whole Piece | Accuracy | 70.2, 69.4 | Local |
| Lopez, Arthur, and Fijinaga (2019) | Key-finding based on a hidden Markov model and key profiles | Classical | 982 | Whole Piece | Accuracy | 94.4 | Local and Global |
| Noland and Sandler (2006) | Key Estimation Using a Hidden Markov Model | Beatles | 110 | Whole Piece | Accuracy | 91 | Global |
| Papadopolous and Peeters (2012) | Local key estimation from an audio signal relying on harmonic and metrical structures | Classical, Popular | 444 | Whole Piece | Accuracy, MIREX | 80.2, 93.6 | Local |
| Pauws (2004) | Musical Key Extraction From Audio | Classical | 237 | Whole Piece | Accuracy | 75.1 | Global |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Quinn and White (2017) | Corpus-derived Key Profiles Are Not Transpositionally Equivalent | Classical | 1591 | Excerpts | NA | NA | Global |
| Schuller and Gollan (2012) | Music Theoretic and Perception-based Features for Audio Key Determination | Popular, Jazz, Classical, Rock, Hip Hop, EDM, Blues | 520 | Whole Piece | Accuracy | 77.3 | Global |
| Sun, Li, and Lei (2009) | Key detection through pitch class distribution model and ANN | Classical, Popular, Rock, Jazz, New Age, Folk | 228 | Whole Piece | Accuracy | 66.1 | Local and Global |
| Temperley (2002) | A Bayesian Approach to Key-Finding | Classical | 896 | Excerpts | Accuracy | 83.6 | Global |
| Temperley and de Clercq (2013) | Statistical analysis of harmony and melody in rock music | Rock | 200 | Whole Piece | Accuracy | 97 | Global (disregards maj/min) |
| Weiss, Cano, and Lukashevich (2014) | A mid-level approach to local tonality analysis: extracting key signatures from audio | Popular | 30 | Whole Piece | Accuracy, MIREX | 86.8, 90.8 | Local |
| Weiß (2013) | Global Key Extraction from Classical music Audio Recordings Based on the Final Chord | Classical | 478 | Whole Piece | Accuracy | 97 | Global |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| White (2018) | Feedback and Feedforward Models of musical key | Classical | 41 | Excerpts | Accuracy | 87.8 | Global |
| Zenz and Rauber (2007) | Automatic Chord Detection Incorporating Beat and Key Detection | Popular, Classical | 35 | Whole Piece | Accuracy | 65 | Local |

# Appendix B

# Key Profiles

Table B.1: Proposed key profile values for major and three variations of minor.

| **Scale Degree** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Value Major** | 2 | 0 | 1 | 0 | 1 | 1 | 0 | 2 | 0 | 1 | 0 | 1 |
| **Value natural minor** | 2 | 0.1 | 1 | 1.5 | 0.1 | 1 | 0.1 | 2 | 1 | 0.1 | 1 | 0.1 |
| **Value harmonic minor** | 2 | 0.1 | 1 | 1.5 | 0.1 | 1 | 0.1 | 2 | 1 | 0.1 | 0.1 | 1 |
| **Value melodic minor** | 2 | 0.1 | 1 | 1.5 | 0.1 | 1 | 0.1 | 2 | 0.5 | 0.5 | 0.5 | 0.5 |

Table B.2: Sapp's Simple key profile values for major and minor.

| **Scale Degree** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Value Major** | 2 | 0 | 1 | 0 | 1 | 1 | 0 | 2 | 0 | 1 | 0 | 1 |
| **Value minor** | 2 | 0 | 1 | 1 | 0 | 1 | 0 | 2 | 1 | 0 | 0.5 | 0.5 |

Table B.3: Krumhansl-Kessler key profile values for major and minor.

| **Scale Degree** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Value Major** | 6.35 | 2.23 | 3.48 | 2.33 | 4.38 | 4.09 | 2.52 | 5.19 | 2.39 | 3.66 | 2.29 | 2.88 |
| **Value minor** | 6.33 | 2.68 | 3.52 | 5.38 | 2.60 | 3.53 | 2.54 | 4.75 | 3.98 | 2.69 | 3.34 | 3.27 |

Table B.4: Temperley key profile values for major and minor.

| **Scale Degree** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Value Major** | 5.0 | 2.0 | 3.5 | 2.0 | 4.5 | 4.0 | 2.0 | 4.5 | 2.0 | 3.5 | 1.5 | 4.0 |
| **Value minor** | 5.0 | 2.0 | 3.5 | 4.5 | 2.0 | 4.0 | 2.0 | 4.5 | 3.5 | 2.0 | 1.5 | 4.0 |

Table B.5: Bellman-Budge key profile values for major and minor.

| **Scale Degree** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Value Major** | 16.80 | 0.86 | 12.95 | 1.41 | 13.49 | 11.93 | 1.25 | 20.28 | 1.80 | 8.04 | 0.62 | 10.57 |
| **Value minor** | 18.16 | 0.69 | 12.99 | 13.34 | 1.07 | 11.15 | 1.38 | 21.07 | 7.49 | 1.53 | 0.92 | 10.21 |

Table B.6: Kostka-Payne key profile values for major and minor.

| **Scale Degree** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Value Major** | 0.748 | 0.060 | 0.488 | 0.082 | 0.670 | 0.460 | 0.096 | 0.715 | 0.104 | 0.366 | 0.057 | 0.400 |
| **Value minor** | 0.712 | 0.084 | 0.474 | 0.618 | 0.049 | 0.460 | 0.105 | 0.747 | 0.404 | 0.067 | 0.133 | 0.330 |

Table B.7: Aarden-Essen key profile values for major and minor.

| **Scale Degree** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Value Major** | 17.77 | 0.15 | 14.93 | 0.16 | 19.80 | 11.36 | 0.29 | 22.06 | 0.15 | 8.15 | 0.23 | 4.95 |
| **Value minor** | 18.26 | 0.74 | 14.05 | 16.86 | 0.70 | 14.44 | 0.70 | 18.62 | 4.57 | 1.93 | 7.38 | 1.76 |

Table B.8: Albrecht-Shanahan key profile values for major and minor.

| **Scale Degree** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Value Major** | 0.24 | 0.01 | 0.11 | 0.01 | 0.14 | 0.09 | 0.02 | 0.21 | 0.01 | 0.08 | 0.01 | 0.08 |
| **Value minor** | 0.22 | 0.01 | 0.10 | 0.12 | 0.02 | 0.10 | 0.01 | 0.21 | 0.06 | 0.02 | 0.06 | 0.05 |

# Appendix C

# SLR Inclusion and Quality Criteria

**IC1:** The main focus of the study is automatic key finding.

**IC2:** The study presents empirical results or other significant contribution to the field.

**IC3:** The study is a primary study, concerns descriptions of technical aspects of the key finding process (e.g. measuring the distance between keys), or describes methods of measuring success in key finding.

**IC4:** The study uses symbolic data (MIDI, Chroma, Sheet Music etc.)

Quality criteria from the SLR guide Kofod-Petersen [2018]:

**QC1:** Is there is a clear statement of the aim of the research?

**QC2:** Is the study is put into context of other studies and research?

**QC3:** Are system or algorithmic design decisions justified?

**QC4:** Is the test data set reproducible?

**QC5:** Is the study algorithm reproducible?

**QC6:** Is the experimental procedure thoroughly explained and reproducible?

**QC7:** Is it clearly stated in the study which other algorithms the study's algorithm(s) have been compared with?

**QC8:** Are the performance metrics used in the study explained and justified?

**QC9:** Are the test results thoroughly analyzed?

**QC10:** Does the test evidence support the findings presented?

# Appendix D

# Complete Set of Experimental Results

Table D.1: Complete set of experimental results: 1a

| Key Profile | Ground Truth Minor Accuracy | Predicted Label Minor Accuracy | MIREX | Accuracy |
|---|---|---|---|---|
| **Proposed Model** | **94.71** | 90.22 | **94.71** | **93.22** |
| **Sapp** | 94.30 | 90.80 | 94.30 | 92.40 |
| **Krumhansl-Kessler** | 85.09 | 87.50 | 85.09 | 76.39 |
| **Kostka-Payne** | 93.37 | 91.98 | 93.37 | 91.07 |
| **Aarden-Essen** | 93.96 | 87.90 | 93.96 | 91.99 |
| **Albrecht-Shanahan** | 94.59 | 91.06 | 94.59 | 92.71 |
| **Bellmann-Budge** | 85.88 | **93.25** | 94.16 | 92.30 |

Table D.2: Complete set of experimental results: MIREX distribution 1a.

| | Correctly Classified | Dominant | Relative | Parallel |
|---|---|---|---|---|
| **Proposed Model** | 908 | 15 | 16 | 11 |
| **Sapp** | 900 | 18 | 23 | 13 |
| **Krumhansl-Kessler** | 744 | 157 | 13 | 12 |
| **Kostka-Payne** | 887 | 16 | 40 | 12 |
| **Aarden-Essen** | 896 | 19 | 25 | 11 |
| **Albrecht-Shanahan** | 903 | 23 | 14 | 13 |
| **Bellmann-Budge** | 899 | 9 | 36 | 14 |

Table D.3: Complete set of experimental results: 1b

| Key Profile | Ground Truth Minor Accuracy | Predicted Label Minor Accuracy | MIREX | Accuracy |
|---|---|---|---|---|
| **Proposed Model** | 90.96 | 90.96 | 94.93 | 93.53 |
| **Sapp** | 89.55 | 92.42 | 94.83 | 93.12 |
| **Krumhansl-Kessler** | 81.07 | 88.04 | 88.31 | 83.26 |
| **Kostka-Payne** | 84.75 | 92.31 | 93.89 | 92.09 |
| **Aarden-Essen** | 90.11 | 89.11 | 93.94 | 91.99 |
| **Albrecht-Shanahan** | 90.40 | 91.43 | 95.00 | 93.74 |
| **Bellmann-Budge** | 86.44 | 93.29 | 94.21 | 92.09 |

Table D.4: Complete set of experimental results: MIREX distribution 1b.

| | Correctly Classified | Dominant | Relative | Parallel |
|---|---|---|---|---|
| **Proposed Model** | 911 | 11 | 17 | 15 |
| **Sapp** | 907 | 14 | 24 | 12 |
| **Krumhansl-Kessler** | 811 | 82 | 15 | 18 |
| **Kostka-Payne** | 897 | 11 | 32 | 12 |
| **Aarden-Essen** | 896 | 20 | 22 | 12 |
| **Albrecht-Shanahan** | 913 | 11 | 16 | 10 |
| **Bellmann-Budge** | 897 | 19 | 31 | 9 |

Table D.5: Complete set of experimental results: 1c

| Key Profile | Ground Truth Minor Accuracy | Predicted Label Minor Accuracy | MIREX | Accuracy |
|---|---|---|---|---|
| **Proposed Model** | 89.41 | 88.89 | 94.60 | 93.48 |
| **Sapp** | 88.24 | 88.76 | 94.24 | 92.67 |
| **Krumhansl-Kessler** | 82.94 | 89.81 | 85.68 | 76.99 |
| **Kostka-Payne** | 84.71 | 90.57 | 93.73 | 91.85 |
| **Aarden-Essen** | 92.35 | 86.74 | 94.07 | 92.06 |
| **Albrecht-Shanahan** | 91.76 | 89.14 | 94.79 | 93.48 |
| **Bellmann-Budge** | 85.29 | 90.63 | 93.97 | 92.26 |

Table D.6: Complete set of experimental results: MIREX distribution 1c.

| | **Correctly Classified** | **Dominant** | **Relative** | **Parallel** |
|---|---|---|---|---|
| **Proposed Model** | 459 | 5 | 6 | 6 |
| **Sapp** | 455 | 6 | 11 | 7 |
| **Krumhansl-Kessler** | 378 | 80 | 5 | 6 |
| **Kostka-Payne** | 451 | 5 | 17 | 8 |
| **Aarden-Essen** | 452 | 8 | 15 | 7 |
| **Albrecht-Shanahan** | 459 | 6 | 6 | 8 |
| **Bellmann-Budge** | 453 | 3 | 17 | 9 |

Table D.7: Complete set of experimental results: 1d

| **Key Profile** | **Ground Truth Minor Accuracy** | **Predicted Label Minor Accuracy** | **MIREX** | **Accuracy** |
|---|---|---|---|---|
| **Proposed Model** | 88,82 | 90,96 | 94,54 | 93,08 |
| **Sapp** | 88,24 | 92,02 | 94,79 | 93,08 |
| **Krumhansl-Kessler** | 81,18 | 90,20 | 89,08 | 84,52 |
| **Kostka-Payne** | 84,12 | 91,67 | 93,75 | 91,85 |
| **Aarden-Essen** | 88,24 | 88,24 | 93,52 | 91,45 |
| **Albrecht-Shanahan** | 90,00 | 91,07 | 94,97 | 93,69 |
| **Bellmann-Budge** | 85,88 | 92,41 | 94,28 | 92,26 |

Table D.8: Complete set of experimental results: MIREX distribution 1d.

| | **Correctly Classified** | **Dominant** | **Relative** | **Parallel** |
|---|---|---|---|---|
| **Proposed Model** | 457 | 7 | 7 | 8 |
| **Sapp** | 457 | 8 | 10 | 7 |
| **Krumhansl-Kessler** | 415 | 38 | 6 | 8 |
| **Kostka-Payne** | 451 | 8 | 13 | 7 |
| **Aarden-Essen** | 449 | 11 | 11 | 7 |
| **Albrecht-Shanahan** | 460 | 6 | 7 | 6 |
| **Bellmann-Budge** | 453 | 9 | 14 | 6 |

Table D.9: Complete set of experimental results: 2a

| Key Profile | Ground Truth Minor Accuracy | Predicted Label Minor Accuracy | MIREX | Accuracy |
|---|---|---|---|---|
| **Proposed Model** | 94.63 | 64.18 | 84.13 | 79.06 |
| **Sapp** | 87.57 | 91.45 | 94.16 | 92.20 |
| **Krumhansl-Kessler** | 87.57 | 66.81 | 78.89 | 70.53 |
| **Kostka-Payne** | 76.55 | 92.81 | 91.76 | 88.81 |
| **Aarden-Essen** | 88.14 | 92.31 | 94.10 | 91.99 |
| **Albrecht-Shanahan** | 86.16 | 93.85 | 93.83 | 91.48 |
| **Bellmann-Budge** | 87.29 | 91.42 | 94.00 | 92.20 |

Table D.10: Complete set of experimental results: MIREX distribution 2a.

| | Correctly Classified | Dominant | Relative | Parallel |
|---|---|---|---|---|
| **Proposed Model** | 770 | 14 | 126 | 23 |
| **Sapp** | 898 | 18 | 25 | 13 |
| **Krumhansl-Kessler** | 687 | 126 | 54 | 11 |
| **Kostka-Payne** | 865 | 16 | 61 | 12 |
| **Aarden-Essen** | 896 | 21 | 26 | 11 |
| **Albrecht-Shanahan** | 891 | 25 | 26 | 13 |
| **Bellmann-Budge** | 898 | 9 | 31 | 19 |

Table D.11: Complete set of experimental results: 2b

| Key Profile | Ground Truth Minor Accuracy | Predicted Label Minor Accuracy | MIREX | Accuracy |
|---|---|---|---|---|
| **Proposed Model** | 94.35 | 74.39 | 89.80 | 86.34 |
| **Sapp** | 88.14 | 93.41 | 94.72 | 92.92 |
| **Krumhansl-Kessler** | 86.16 | 78.61 | 87.04 | 82.24 |
| **Kostka-Payne** | 79.94 | 94.02 | 92.94 | 90.66 |
| **Aarden-Essen** | 88.98 | 88.24 | 93.23 | 91.17 |
| **Albrecht-Shanahan** | 82.77 | 94.52 | 93.73 | 91.89 |
| **Bellmann-Budge** | 86.44 | 93.29 | 94.18 | 92.09 |

Table D.12: Complete set of experimental results: MIREX distribution 2b.

| | **Correctly Classified** | **Dominant** | **Relative** | **Parallel** |
|---|---|---|---|---|
| **Proposed Model** | 841 | 13 | 84 | 10 |
| **Sapp** | 905 | 15 | 25 | 13 |
| **Krumhansl-Kessler** | 801 | 74 | 26 | 10 |
| **Kostka-Payne** | 883 | 11 | 45 | 16 |
| **Aarden-Essen** | 888 | 21 | 22 | 15 |
| **Albrecht-Shanahan** | 895 | 12 | 27 | 19 |
| **Bellmann-Budge** | 897 | 19 | 30 | 9 |

Table D.13: Complete set of experimental results: 2c

| | **Ground Truth** | **Predicted Label** | | |
|---|---|---|---|---|
| **Key Profile** | **Minor Accuracy** | **Minor Accuracy** | **MIREX** | **Accuracy** |
| **Proposed Model** | 93.53 | 61.15 | 83.38 | 78.82 |
| **Sapp** | 86.47 | 90.18 | 94.30 | 92.67 |
| **Krumhansl-Kessler** | 87.65 | 63.40 | 77.74 | 69.45 |
| **Kostka-Payne** | 77.06 | 91.61 | 92.36 | 89.82 |
| **Aarden-Essen** | 87.06 | 92.50 | 94.44 | 92.46 |
| **Albrecht-Shanahan** | 84.71 | 92.90 | 94.22 | 92.46 |
| **Bellmann-Budge** | 87.06 | 89.16 | 93.87 | 92.26 |

Table D.14: Complete set of experimental results: MIREX distribution 2c.

| | **Correctly Classified** | **Dominant** | **Relative** | **Parallel** |
|---|---|---|---|---|
| **Proposed Model** | 387 | 2 | 64 | 11 |
| **Sapp** | 455 | 6 | 12 | 7 |
| **Krumhansl-Kessler** | 341 | 60 | 31 | 7 |
| **Kostka-Payne** | 441 | 5 | 28 | 8 |
| **Aarden-Essen** | 454 | 9 | 12 | 8 |
| **Albrecht-Shanahan** | 454 | 7 | 13 | 6 |
| **Bellmann-Budge** | 453 | 3 | 14 | 11 |

s

Table D.15: Complete set of experimental results: 2d

| Key Profile | Ground Truth Minor Accuracy | Predicted Label Minor Accuracy | MIREX | Accuracy |
|---|---|---|---|---|
| **Proposed Model** | 9353 | 7465 | 9024 | 8697 |
| **Sapp** | 8765 | 93,13 | 94,79 | 93,08 |
| **Krumhansl-Kessler** | 85,29 | 80,11 | 87,72 | 83,10 |
| **Kostka-Payne** | 80,00 | 93,15 | 92,85 | 90,43 |
| **Aarden-Essen** | 87,06 | 87,57 | 92,83 | 90,63 |
| **Albrecht-Shanahan** | 83,53 | 94,04 | 94,03 | 92,46 |
| **Bellmann-Budge** | 86,47 | 92,45 | 94,42 | 92,46 |

Table D.16: Complete set of experimental results: MIREX distribution 2d.

| | Correctly Classified | Dominant | Relative | Parallel |
|---|---|---|---|---|
| **Proposed Model** | 427 | 7 | 38 | 6 |
| **Sapp** | 457 | 8 | 10 | 7 |
| **Krumhansl-Kessler** | 408 | 37 | 10 | 6 |
| **Kostka-Payne** | 444 | 8 | 21 | 8 |
| **Aarden-Essen** | 445 | 12 | 10 | 9 |
| **Albrecht-Shanahan** | 454 | 6 | 11 | 7 |
| **Bellmann-Budge** | 454 | 9 | 13 | 6 |

Table D.17: Complete set of experimental results: 3a

| Key Profile | Ground Truth Minor Accuracy | Predicted Label Minor Accuracy | MIREX | Accuracy |
|---|---|---|---|---|
| **Proposed Model** | 86.44 | 90.53 | 92.74 | 90.04 |
| **Sapp** | 83.33 | 93.95 | 93.05 | 90.35 |
| **Krumhansl-Kessler** | 73.73 | 89.69 | 84.15 | 76.28 |
| **Kostka-Payne** | 85.88 | 92.97 | 92.90 | 90.04 |
| **Aarden-Essen** | 87.01 | 89.80 | 92.63 | 89.94 |
| **Albrecht-Shanahan** | 85.31 | 93.79 | 93.12 | 90.25 |
| **Bellmann-Budge** | 85.03 | 94.06 | 93.45 | 90.86 |

Table D.18: Complete set of experimental results: MIREX distribution 3a.

| | Correctly Classified | Dominant | Relative | Parallel |
|---|---|---|---|---|
| **Proposed Model** | 877 | 28 | 33 | 12 |
| **Sapp** | 880 | 25 | 40 | 9 |
| **Krumhansl-Kessler** | 743 | 127 | 33 | 16 |
| **Kostka-Payne** | 877 | 30 | 36 | 10 |
| **Aarden-Essen** | 876 | 29 | 33 | 9 |
| **Albrecht-Shanahan** | 879 | 31 | 35 | 10 |
| **Bellmann-Budge** | 885 | 25 | 35 | 11 |

Table D.19: Complete set of experimental results: 3b

| Key Profile | Ground Truth Minor Accuracy | Predicted Label Minor Accuracy | MIREX | Accuracy |
|---|---|---|---|---|
| **Proposed Model** | 86.47 | 91.88 | 93.32 | 90.84 |
| **Sapp** | 81.76 | 93.92 | 93.22 | 90.43 |
| **Krumhansl-Kessler** | 74.71 | 88.81 | 84.18 | 76.17 |
| **Kostka-Payne** | 84.71 | 91.72 | 92.51 | 89.61 |
| **Aarden-Essen** | 87.06 | 89.70 | 93.22 | 90.63 |
| **Albrecht-Shanahan** | 86.47 | 93.63 | 93.91 | 91.45 |
| **Bellmann-Budge** | 85.29 | 92.95 | 93.71 | 91.24 |

Table D.20: Complete set of experimental results: MIREX distribution 3b.

| | Correctly Classified | Dominant | Relative | Parallel |
|---|---|---|---|---|
| **Proposed Model** | 446 | 14 | 14 | 5 |
| **Sapp** | 444 | 14 | 21 | 2 |
| **Krumhansl-Kessler** | 374 | 67 | 14 | 8 |
| **Kostka-Payne** | 440 | 16 | 18 | 4 |
| **Aarden-Essen** | 445 | 14 | 17 | 3 |
| **Albrecht-Shanahan** | 449 | 13 | 16 | 4 |
| **Bellmann-Budge** | 448 | 11 | 18 | 6 |

NTNU

Norwegian University of
Science and Technology