THE KUSC CLASSICAL MUSIC DATASET FOR AUDIO KEY FINDING

Ching-Hua Chuan[1] and Elaine Chew[2]

[1]School of Computing, University of North Florida, Jacksonville, Florida, USA
[2]School of Engineering and Computer Science, Queen Mary, University of London, London, UK

ABSTRACT

In this paper, we present a benchmark dataset based on the KUSC classical music collection and provide baseline key-finding comparison results. Audio key finding is a basic music information retrieval task; it forms an essential component of systems for music segmentation, similarity assessment, and mood detection. Due to copyright restrictions and a labor-intensive annotation process, audio key finding algorithms have only been evaluated using small proprietary datasets to date. To create a common base for systematic comparisons, we have constructed a dataset comprising of more than 3,000 excerpts of classical music. The excerpts are made publicly accessible via commonly used acoustic features such as pitch-based spectrograms and chromagrams. We introduce a hybrid annotation scheme that combines the use of title keys with expert validation and correction of only the challenging cases. The expert musicians also provide ratings of key recognition difficulty. Other meta-data include instrumentation. As demonstration of use of the dataset, and to provide initial benchmark comparisons for evaluating new algorithms, we conduct a series of experiments reporting key determination accuracy of four state-of-the-art algorithms. We further show the importance of considering factors such as estimated tuning frequency, key strength or confidence value, and key recognition difficulty in key finding. In the future, we plan to expand the dataset to include meta-data for other music information retrieval tasks.

KEYWORDS

Audio Key Finding, Ground Truth, Tuning, Dataset, Evaluation

1. INTRODUCTION

In this paper, we present a publicly available dataset for audio key finding. In tonal music, the key establishes the context in which the roles of pitches and chords are defined. Determining key from audio recordings is one of the most important music information retrieval (MIR) tasks; key finding forms an essential component of systems for music segmentation, similarity assessment, and mood detection. Due to issues such as copyright restrictions and the lack of annotated ground truth, algorithms are often evaluated using relatively small datasets from private collections. In addition, tonal complexity varies significantly in different genres and stylistic periods. Therefore, simply using correct rates and dataset size to judge the performance of an audio key finding system is a far from ideal approach. A commonly accessible dataset consisting of adequate examples with sufficient variety and accurate annotations is thus needed.

The KUSC dataset presented in this paper consists of 15-second excerpts of compositions by Bach, Mozart, and Schubert, including symphonies, concertos, preludes, fugues, sonatas, quartets and more. We focus on the global key by examining only the first and last 15 seconds of the

DOI : 10.5121/ijma.2014.6401

recording—typically when the key of the piece is established and when the piece returns to the original key, respectively—to reduce the possibility of key modulations. A common key labelling method simply uses the key in the title, such as Symphony in G major, as ground truth.

While highly efficient, we show that this method can sometimes be prone to error. We introduce a hybrid key annotation scheme that combines the use of title keys with selective manual validation and correction of challenging cases. The manual annotations by expert musicians are accompanied by ratings of key recognition difficulty, on a 5-point Likert scale, that provide valuable information for evaluation of key-finding algorithms. Meta-data related to instrumentation are also listed with the excerpt as tags.

For copyright reasons, the excerpts are shared in the form of commonly used and irreversible acoustic features instead of audio waves. The acoustic features include the constant-Q spectrogram, chromagram, and other mid-level features proposed by existing algorithms.

With the goal to develop the dataset as a benchmark for audio key finding, we conducted a series of experiments using existing algorithms to provide an initial baseline for performance comparisons. We use the dataset to test four existing audio key finding algorithms and report their individual and combined scores for key finding accuracy. We further demonstrate that it is important to consider in the evaluations factors such as tuning frequency, key strength/confidence, and key recognition difficulty.

The recorded ensemble's or instrument's tuning directly impacts an algorithm's chance of success. Instrumentalists and ensembles may tune to reference frequencies different from A440; for example, early music ensembles may tune to an A of 415 Hz, which would currently be heard as an A^b. We first apply two tuning estimation methods and compare the difference between their tuning frequency outputs. We also study the difference in estimated tuning between the first and last 15 seconds of the same piece.

Using straightforward methods for determining key strength or confidence from the respective models, we assessed the relation between the algorithm's accuracy and the confidence or key strength value. Based on the combined score, we investigate the relationship between key recognition difficulty and instruments/music styles, and the algorithms' accuracy.

Although the dataset is originally designed for audio key finding, we plan to expand the dataset with more data for other MIR tasks such as automatic chord recognition and instrument identification.

The remainder of the paper is organized as follows: Section 2 provides a general introduction to audio key finding systems and current methods for evaluating them; Section 3 describes the KUSC Classical music dataset, and the meta-data provided with the dataset; Section 4 introduces the hybrid approach to key annotation and the rating of key-finding difficulty; Section 5 presents the benchmark evaluations of four state-of-the-art audio key finding algorithms using the KUSC dataset, followed by the conclusions.

2. AUDIO KEY FINDING AND RELATED WORK

2.1. Problem Definition and Background

The task of audio key finding is to identify the key of a music composition from its audio recording. Polyphonic music comprise of multiple streams of notes sounded simultaneously, such as music with multiple parts played simultaneously by different instruments as shown in the score

in Figure 1. Each part is a sequence of note and silent events, and each note event can be described by a vector of properties, including <pitch, onset time, and duration>. For example, the first three note events in the part that shown on the top of the score in Figure 1 are <b_4^b, 1, 0.5>, <b_4^b, 1.5, 0.125> and <a_4^b, 1.625, 0.125>, in which onset time and duration are presented in beats. A pitch (b_4^b) consists of pitch class (b^b) and register (4) that indicates the height of the pitch. There exists 12 pitch classes including {c, $c^\#/d^b$, d, $d^\#/e^b$, e, f, $f^\#/g^b$, g, $g^\#/a^b$, a, $a^\#/b^b$, b}.

In Western tonal music, the key refers to the pitch context of a piece of music, and is represented by the tonal center (a.k.a. the tonic), the most stable pitch class (for example, A or B) in the key. The tonal center provides a reference for the identity and function of all other pitches in the piece. For example, in a piece of E^b major key as shown in Figure 1, pitch e^b is the tonic and pitch b^b is the fifth (five scale steps up from the pitch e^b), the two most stable pitches in the piece. In contrast, pitches such as a and b are less likely to appear in a piece of E^b major. Therefore, the occurrence of the 12 pitch classes is an important indicator for the key. There are 24 keys in Western tonal music, 12 major keys and 12 minor keys. In this study we concentrate on the global key, the one key that operates over the entire length of a piece of music.

2.2. A General Architecture of Audio Key Finding Systems

A general architecture of audio key finding systems is illustrated in Figure 1. The input of the system is an audio recording of a music composition or improvisation, and the output is the key. An audio key finding system generally consists of two components: the first component focuses on audio signal processing to extract features related to the pitch classes content of the sample, while the second component determines the key based on the extracted pitch class information.

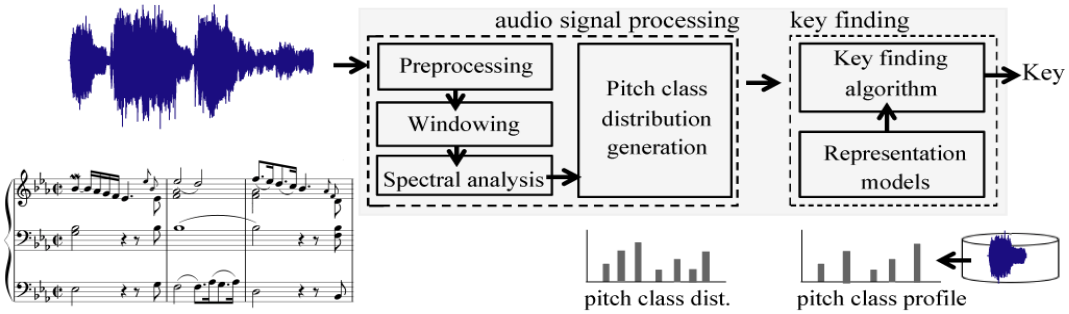


Figure 1. A general architecture of audio key finding systems

In the audio signal-processing component, the musical signal is first pre-processed by removing silence and reducing noise. It is then divided into overlapped frames for spectral analysis. The purpose of spectral analysis on short-duration frames is to obtain local information about note events, particularly pitch information. Pitch information is related to the frequency of the signal. In the equal temperament system, the interval between any two pitches corresponds to a simple ratio with whole numbers indicating the relation between their frequencies. Using the standard tuning 440Hz for pitch A_4, the frequency of the pitch p that is n semitones away from pitch A_4 can be calculated by:

$$\text{Frequency}(p) = 440 \times 2^{\frac{n}{12}},$$

where n is a positive integer if the pitch is higher than A_4 and a negative integer if it is lower.

The result of spectral analysis is further converted into a distribution representing the occurrence of the 12 pitch classes. Various approaches have been proposed for the generation of pitch class

distributions with the goal of increasing the accuracy of pitch information. A detailed survey of existing algorithms is summarized in [1].

The key finding component in Figure 1 consists of representation models for the 24 keys and an algorithm for determining the key by comparing the generated pitch class distribution and the representation models. The representation model is usually defined as a typical pitch class distribution for a specific key. The model can be constructed based on psychoacoustic experiments [2], music theory [3], or learned from a dataset [4]. The comparison between the generated pitch class distribution and the representation model in a key finding algorithm usually involves calculating a distance function [5], correlations [6], or posterior probabilities [4]. The key that has the shortest distance, or the highest correlation or probability, is chosen as the output.

2.3. Related Work

Table 1 summarizes existing approaches for audio key finding with their reported accuracy and datasets used for evaluations.

Table 1. A summary of existing approaches for audio key finding

Approach	Accuracy	Dataset, Ground Truth, and Availability
Chai & Verco [7]	84%	10 classical piano pieces Ground truth was manually annotated by the author.
Chuan & Chew [5, 8]	75% (410 classical), 70% (Chopin)	410 classical pieces / 24 Chopin preludes Title key was used as ground truth.
Gómez & Herrera [6]	64%	878 excerpts of classical music Title key was used as ground truth.
İzmirli [9] (template-based)	86%	85 classical pieces
İzmirli [10] (local key)	84%(pop), 76.9% (classical)	17 pop / 17 classical excerpts Ground truth was manually annotated.
Peeters [4]	81%	302 classical pieces Title key was used as ground truth.
Papadopoulos & Peeters [1]	80.21% (classical), 61.31% (pop)	2 datasets: • 5 movements of Mozart piano sonatas; ground truth (key and chord progression) was manually annotated; info. about composer and composition title is available • 16 real audio songs annotated by IRCAM (QUAERO project)
Schuller & Gollan [11]	78.5% (pop), 88.8 (classical), 63.4 (jazz)	520 pieces in pop, jazz, classical and others Ground truth was labeled by three musicians without disagreements. Tracks with key modulation were removed.
Sun, Li & Ma [12]	68.7% in average	228 pieces in classical, pop, jazz, new age, and folk Ground truth was labeled either by students or from published music books.
Weiss [13]	92.2%, 93.7%, and 97%	3 datasets: • 115 tracks of 29 symphonies: title key was used as the ground truth; info. about composer and symphony no. is available • Saarland Music Data [14]: 126 tracks performed by students; ground truth was manually annotated; recordings are available in mp3 • 237 piano sonatas [15]: title key was used as the ground truth; info. about composer, composition title, and performer is available.

3. THE KUSC CLASSICAL MUSIC DATASET

This section describes the KUSC classical music dataset for audio key finding, focusing on acoustic features, meta-data, and data formats that are made available with the dataset.

3.1. Music in the Dataset

The dataset used in this study was provided by Classical KUSC, a classical public radio station. We selected compositions by Bach, Mozart and Schubert for audio key finding because they represent three different styles with distinguishable levels of tonal complexity. In addition, tonality is more clearly defined in these pieces than in more contemporary compositions. The selected compositions include symphonies, concertos, sonatas, quartets, preludes, fugues, and other forms with instruments including the recorder, violin, flute, piano, oboe, guitar, and choir. For multi-movement works, we used only the first and last movement because these segments are generally in the title key.

The entire dataset consists of 3224 different excerpts. We extracted two excerpts from each recording: one containing the first 15 seconds and the other representing the last 15 seconds of the recording. We only considered the 15-second segments of the recording because the key is more likely to remain in the global key without modulations in these parts of the piece. Silence in the beginning and the end of each recording were identified and removed using root-mean square energy as described in [16].

The audio excerpts and their titles are not available for public access for copyright reasons. However, low-level acoustic features related to audio key finding are made available for key finding research as described in the following section.

3.2. Acoustic Features

Acoustic features related to tuning frequency and to presence of pitch classes are the most important for audio key finding algorithms. Therefore, for each excerpt, we extracted frame-by-frame features—the constant-Q spectrogram, chromagram, and pitch class profiles—related to pitch classes using the fuzzy analysis center of effect generator algorithm proposed by Chuan and Chew [5], Gómez's harmonic pitch class profile [17], Lartillot and Toiviainen's Matlab implementation [16] of Gómez's method, and Noland and Sandler's hidden Markov models for tonal estimation [18]. The features are given based on the tuning frequency estimated by two existing methods [19, 20], as well as the standard tuning frequency (440 Hz for pitch A_4).

3.2.1. Estimated Tuning Frequency

Here, we describe the two methods employed for estimating the tuning frequency.

The first is based on a method described by Müller and Ewert in [19]. The global tuning of a recording is estimated by shifting the center of multi-rate filter banks to obtain an average spectrogram based on several tuning frequencies. The deviation of the center that produces the highest average value in the spectrogram is then identified as the estimated tuning frequency. In this study, six deviation classes corresponding to shifts of {-1/2,-1/3,-1/4,0,1/4,1/3} semitones from 440Hz were used.

The second is based on the technique outlined by Mauch and Dixon in [20]. Estimation of the tuning frequency corresponds to the phase angle of discrete Fourier transform using circular statistics. The frequency is wrapped onto the interval $[-\pi, \pi)$ where π represents a quartertone. Details of obtaining the estimated frequency can be found in [21].

Note that equal temperament is assumed in both tuning methods. To compare the differences between the estimated tuning frequencies from the two methods, the output of [20] is mapped to one of the categories in [19]: {427.4741, 431.6092, 433.6918, 440, 446.3999, 448.5539} Hz for A_4.

3.2.2. Pitch Class Distribution/Chromagram and Pitch-Based Spectrogram

Each pitch-based spectrogram consists of a series of vectors that represent spectral information related to pitches as defined by their frequency ranges. For example, a pitch-based spectrogram can be a matrix in which each row consists of 88 values for pitches A_0 to C_8. Pitch class distributions or chromagrams, usually computed by folding values in pitch-based spectrograms from pitches into pitch classes, are vectors of 12 values that relate to the distribution of 12 pitch classes. Four algorithms [5, 16, 17, 18], to be outlined in Section 4.1.1, were used to extract features for pitch-based spectrograms and pitch class distribution/chromagrams. In [5, 16, 17], Fast Fourier Transform or Short Time Fourier Transform is used to produce spectral information. In [18], the constant-Q transform is used. We briefly outline the concepts underlying these methods.

In [17], Gómez detected spectral peaks using 120 bins per octave, a higher number of bins than the standard 12 per octave for pitches. A triangular weighting function is used to reduce boundary errors in the adjacent frequency bins. The resulting Harmonic Pitch Class Profile (HPCP) is a vector of 120 values for 12 pitch classes. The chromagram in Lartillot and Toiviainen's MIR Toolbox [16] is another implementation of Gómez's approach with slightly different default parameter settings including frame size, hop size, pitch range and sampling frequency.

Chuan and Chew's fuzzy analysis technique in [5] also generates a 12-value Pitch Class Profile (PCP) for audio key finding. The technique uses knowledge of the harmonic series to reduce the errors in noisy low frequency pitches. Pitches in low frequency are important to key finding because they usually indicate the root of the chord played by the bass. However, it is difficult to identify these pitches correctly because of the logarithmic scale in pitch frequency.

We also used Queen Mary University of London's Vamp plug-in [22] to generate a constant-Q spectrogram [23] for each excerpt. The spectrogram is used in Noland and Sandler's key finding method in [18]. The constant-Q spectrogram is the pitch-based spectrogram used to create the folded chromagram.

Figure 2 provides an example of the extracted acoustic features for a recording of Bach's Prelude BWV 552 as shown in (a). The color specifies the significance of the observed frequency or pitch, with the color of red for high values and blue or black for lower values. The resolution in the grids depends on the algorithm's default setting in parameters such as frame size and hop size. Figure 2 (b) and (d) are generated by QM Vamp plug-in and (e) and (f) are generated using Matlab, which result in similar color scheme for the figures using the same software. The y-axis represents the 12 pitch classes in (d), (e), and (f) while the y-axis in (b) lists a range of detected pitches. In Figure 2 (c), the y-axis is the 120 bins for the 12 pitch classes, which result in the different visualization from other figures in Figure 2.

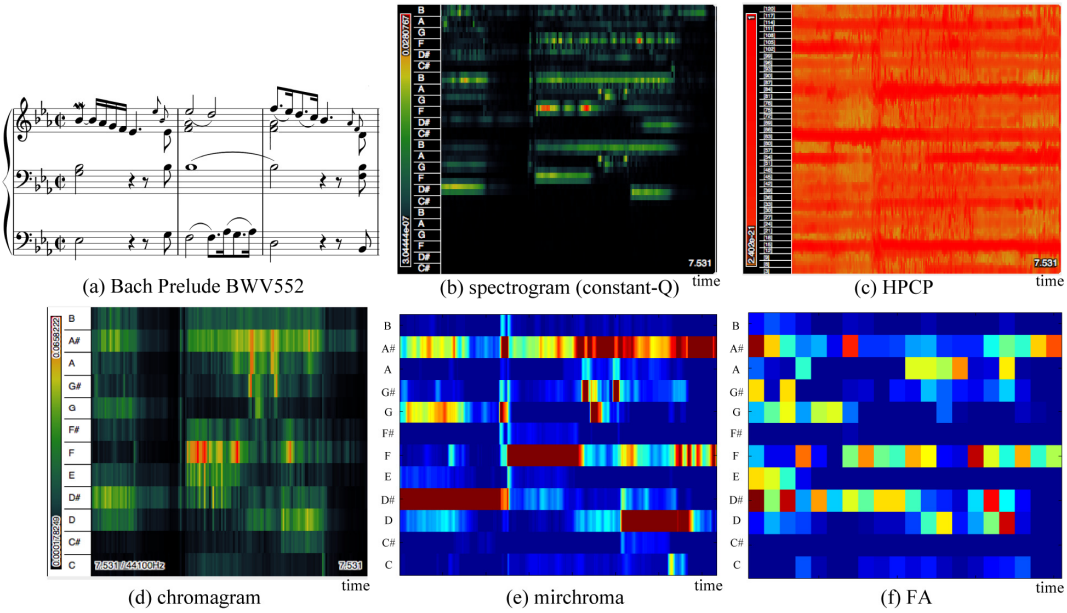


Figure 2. (a) The first three bars of Bach Prelude in E^b major BWV552, and extracted acoustic features: (b) spectrogram [22], (c) HPCP [17], (d) chromagram [23], (e) mirchroma [16] and (f) fuzzy analysis [5].

3.3. Meta-data

In addition to acoustic features and manual annotations, we also provided meta-data related to the instrumentation. The meta-data consist of the tags, listed in Table 2, associated with each piece of music.

Two types of information were manually annotated for the excerpts: ground truth key and key recognition difficulty. We took a hybrid approach to ask musicians to manually annotate the key only for the challenging excerpts without exhausting the annotators. We also noted the key recognition difficulty by asking another three musicians to indicate how difficult it was for them to determine the key using a 5-point Likert scale. The details for annotating ground truth key and key recognition difficulty are described in Section 4.

For researchers to systematically compare the performance of their newly developed algorithm with existing ones, we computed an average weighted score for each excerpt using the four audio key finding algorithms [5, 16, 17, 18]. For each excerpt, each algorithm receives a score based on the relation between their output key and the ground truth: 1 point (a full score) if the output is identical to the ground truth, 0.5 if they are a perfect fifth apart, 0.3 if one is the relative major/minor of the other, and 0.2 if one is the parallel major/minor of the other. An average is then calculated using the four scores obtained by the algorithms. The details of the score calculation and experiments are described in Section 5.

Table 2. Tags and their associated number

1	Soprano	7	Choir	13	Symphony	19	Harpsichord	25	Clarinet
2	Alto	8	Cello	14	Orchestra	20	Fortepiano	26	Oboe
3	Tenor	9	Violin	15	Concerto	21	Piano	27	Trumpet
4	Bass	10	Viola	16	Lute	22	Sonata	28	Bassoon
5	Baritone	11	String quartet	17	Guitar	23	Recorder	29	Horn
6	Chorus	12	Organ	18	Harp	24	Flute	30	Posthorn

3.4. Data Formats

The dataset is available via *www.unf.edu/~c.chuan/KUSC/*. Acoustic features described in Section 3.2.2 are pre-computed using different settings for the pre-defined parameters: frame size, hop size, pitch range, sampling frequency, number of frequency bins per octave. The dataset can be downloaded as delimited text files and MySQL database files.

4. ANNOTATING GROUND TRUTH AND KEY RECOGNITION DIFFICULTY

This section introduces our procedure to obtaining accurate ground truth key information for the dataset by first identifying challenging musical fragments, then manually annotating them with key information and key finding difficulty ratings.

4.1. A Hybrid Approach to Ground Truth Annotations

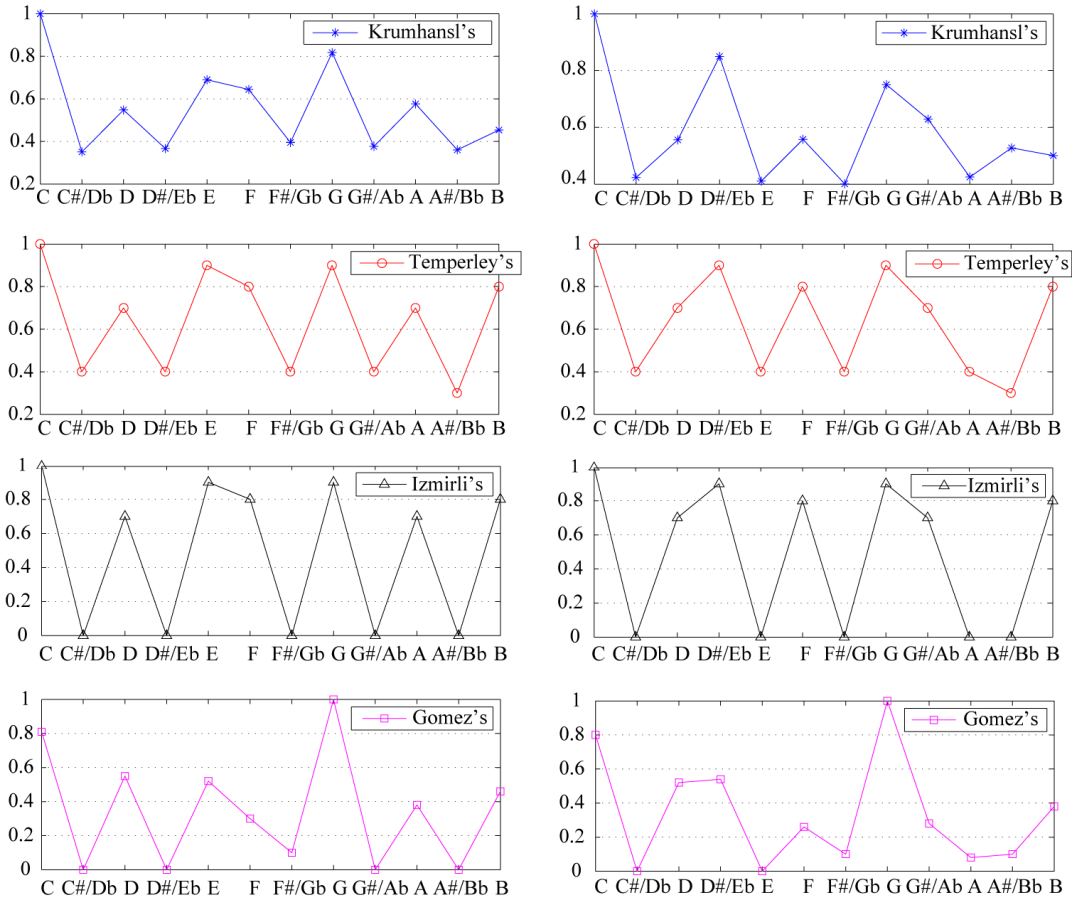
We first assigned the title key as ground truth for every excerpt in the dataset. To avoid exhaustive manual examination of the entire dataset, we implemented five audio key finding systems and used them to focus the manual examination on a subset consisting of challenging excerpts in which the title key may not be the key. Each excerpt was first supplied as the input to the five systems, which produces five key outputs. If no more than two out of the five systems reported the same answer as the title key, the excerpt was labeled as a challenging case requiring further manual examination. This challenging set was re-examined by three professional musicians and their keys manually labelled by them.

4.1.1. Key Finding Algorithms

In this section, we briefly describe the audio key finding systems implemented for the construction of the challenging set in detail, emphasizing the uniqueness of each system. More details can be found in [24]. The methods we have chosen are mainly template-based approaches, in which a pitch class distribution or chromagram is compared to a template key representation. Note that systems that rely on training data were not implemented because the title key (ground truth) may not be the key of the excerpt.

The five systems implemented include algorithms using Krumhansl and Schmuckler's (K-S) probe tone profile [2], Temperley's modified version of the K-S model [3], İzmirli's template-based correlation model [9], Gómez's harmonic pitch class profile (HPCP) method [17], and our fuzzy analysis center of effect generator (FACEG) algorithm [5]. The first four algorithms all compute a correlation coefficient between the pitch class distribution of the excerpt and the template pitch class profile for each key. They differ mostly in their template pitch class profiles as shown in Figure 3. We thus implemented a basic approach to generate pitch class distributions using the Fast Fourier Transform.

In [5], we proposed an audio key finding system called fuzzy analysis center of effect generator. A fuzzy analysis technique is used for PCP generation using the harmonic series to reduce the errors in noisy low frequency pitches. The PCP is further re-fined periodically using the current key information. The representation model used in the system is Chew's Spiral Array model [25, 26], a representation of pitches, chords, and keys in the same 3-dimensional space with distances reflecting their musical relations. The Center of Effect Generator (CEG) key finding algorithm determines key via a nearest-neighbor search in the Spiral Array space in real-time: an instantaneous key answer is generated in each window based on past information. The CEG is the only model that considers pitch spelling.



(a) C major key profiles (b) C minor key profiles

Figure 3. C major and C minor key profiles proposed by Krumhansl, Temperley, İzmirli, and Gómez.

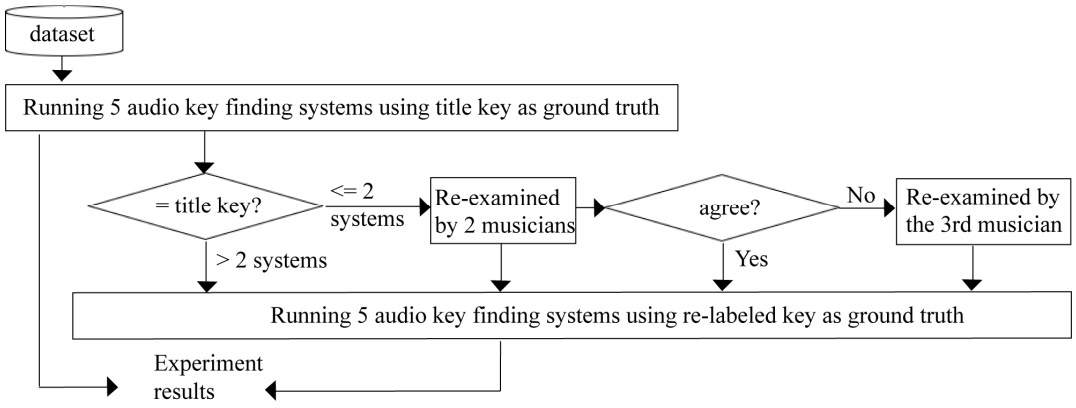


Figure 4. The flowchart of the hybrid approach for ground truth annotations

4.1.2. Experiments and Results

The process of examining the accuracy of the title key and re-labeling the ground truth is illustrated in Figure 4. An excerpt was first examined by the five implemented audio key finding systems. If more than two systems out of the five reported the key identical to the one in the title,

we saved the title key as the ground truth. If not, the excerpt was moved to the challenging set and re-examined by two musicians. If the two musicians had the same key for the excerpt, then we saved the re-labeled key as the ground truth. If the two musicians disagreed, the excerpt was examined by the third musician. Experiment results were reported by comparing the accuracy of the five audio key finding systems using the title key and the relabelled key as ground truth.

Out of the 3324 excerpts, 727 excerpts (21.87%) were moved to the challenging set that was examined by musicians. Table 3 lists the distribution of the challenging set, in absolute numbers and as a percentage of the number of excerpts we considered by each composer.

Table 3. Details of the entire dataset and the challenging set by Bach, Mozart, and Schubert

Composer	Total # of recordings	Challenging set (first 15 sec)	Challenging set (last 15 sec)
Bach	553	245 (44.30%)	244 (44.12%)
Mozart	873	75 (8.59%)	98 (11.23%)
Schubert	236	24 (10.17%)	41 (17.37%)

Figures 5 and 6 show the experiment results using the title key and the re-labeled key respectively. The reported key are divided into nine categories based on its relation to the ground truth: correct (Cor), dominant (Dom), sub-dominant (Sub), parallel major/minor (Par), relative major/minor (Rel), same mode with the root one semitone higher (M+1), same mode with the root one semitone lower (M-1), same mode but not in the previous categories (MO), and relations not included in any of the previous categories (Others).

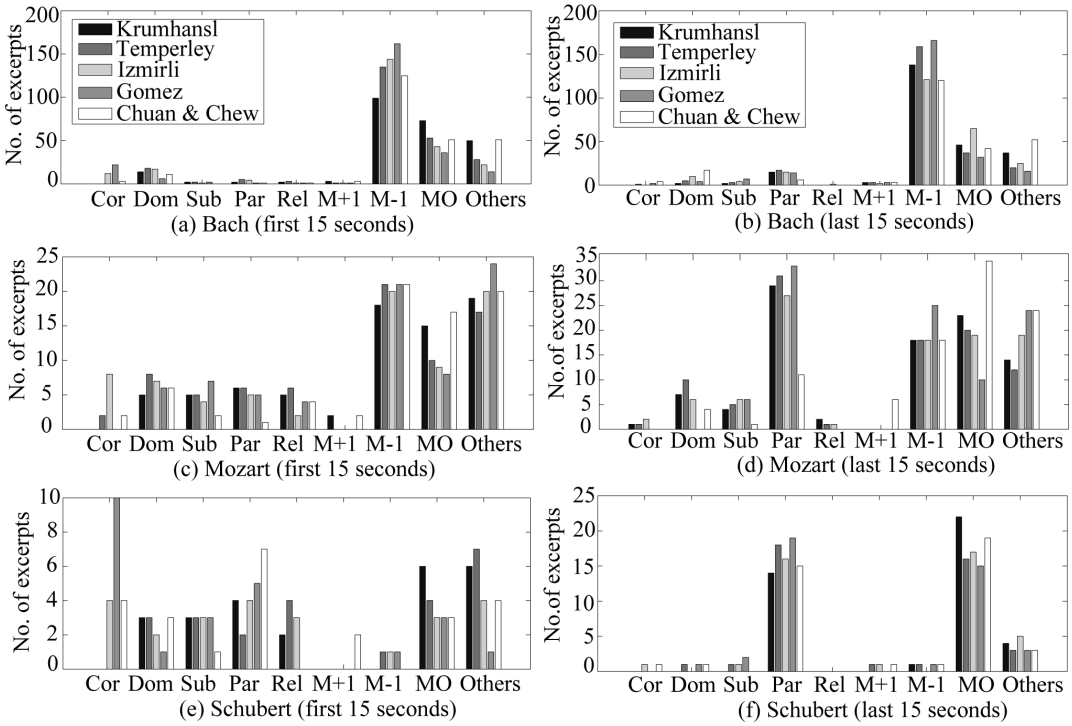


Figure 5. Key finding results for the challenging dataset using the title key as ground truth.

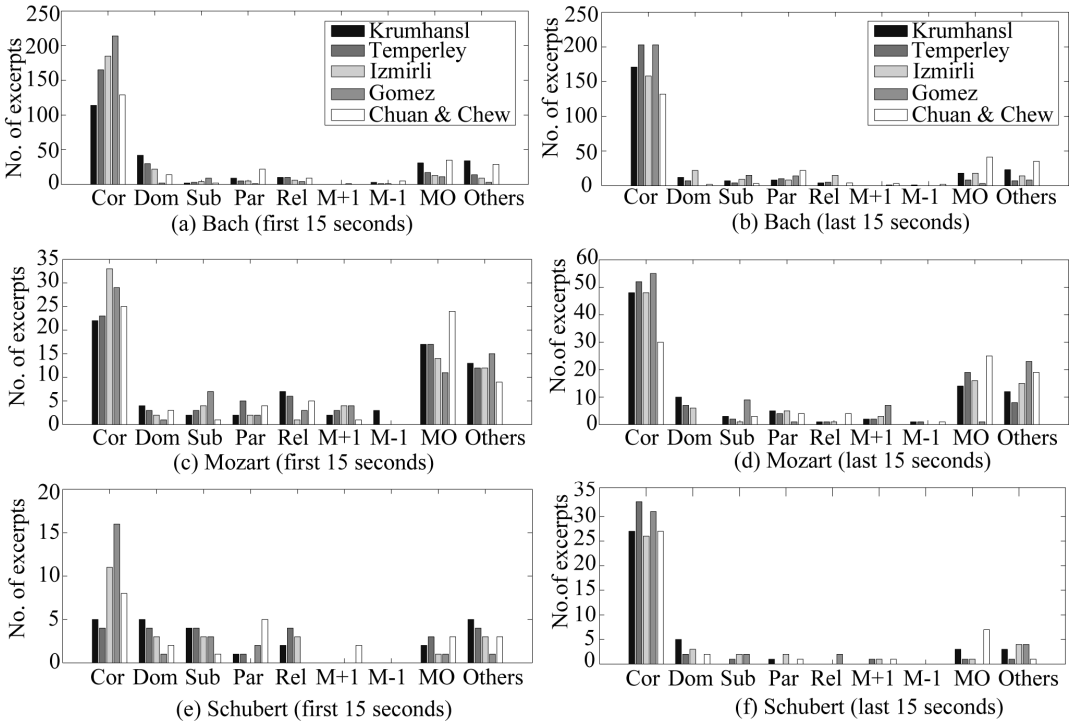


Figure 6. Key finding results for the challenging dataset using the re-labeled key as ground truth.

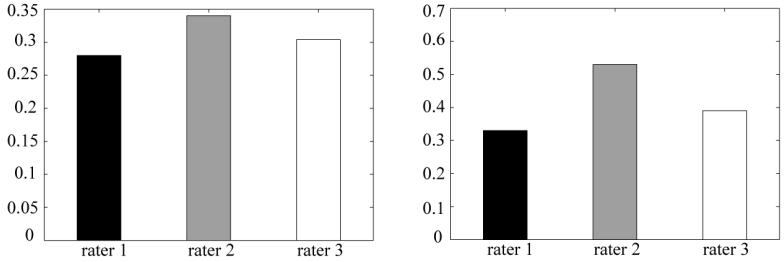
Comparing Figure 5 (a), (b) with Figure 6 (a), (b), it can be observed that most of the incorrect answers using the title key fall into the last three categories, especially (M-1), which indicates that tuning may be an issue in key finding for Bach's pieces. Similar results can be observed in Mozart's pieces by comparing Figure 5 (c), (d) with Figure 6 (c), (d). However, the significant proportion in parallel major/minor (Par) category in Figure 5 (d) indicates that many Mozart pieces actually end in the parallel key with respect to the title key. In Figure 5 (e), the reported keys are more evenly distributed among the categories for Schubert's pieces using the title key as the ground truth. This implies more complex compositional strategies that the composer employed to start the piece. The result in Figure 5 (f) for the last 15-second of Schubert's pieces is more similar to Mozart's than Bach's, showing that the pieces end in the parallel major/minor key or a key that has the same mode as the title key. More detailed analysis can be found in [24].

4.2. Key Recognition Difficulty Annotations

In addition to reporting the key of the excerpts, we also collected data about key recognition difficulty. Three musicians were asked to listen to the excerpts in the challenging dataset, and use the 5-point Likert scale, with 1 indicating the easiest level and 5 the most difficult, to rate the difficulty that they experienced in recognizing the key. One objective of this data collection step was to observe how individuals perceive this difficulty and the difference between individuals' responses. Therefore, no formal definition of key recognition difficulty, nor examples of each level, were given to the musicians.

We also studied the consistency of individual rater's annotations by examining the rater's responses to multiple versions of the same compositions, i.e., recordings of the same piece by different performers. A consistent rater should indicate similar difficulty levels and identical, or closely related, keys for the different recordings of the same piece.

Figure 7 shows the rater's consistency according to two measures: (a) the weighted average of standard deviations of rated difficulty levels; and, (b) the ratio of number of distinct keys to total number of duplicate excerpts. The first measure is the average of the standard deviations, computed for each set of multiple recordings of the same piece, weighted by the number of recordings in the set. This measure is bounded between zero and 2.5, when two scores are at the extreme ends. The second measure serves as a confusion indicator for determining the key for different versions of the same piece. For the given dataset, this measure is bounded between 0.248 and 1, when every duplicated excerpt is assigned a different key. According to both measures, the raters' labels are found to be consistent as the results are closer to the lower bounds of each measure. We also proposed ways to automatically predict key recognition difficulty based on acoustic features as described in [27].



(a) Weighted average of standard deviations of rated difficulty levels

(b) Ratio of number of distinct keys to total number of repeated excerpts

Figure 7. Consistency of individual rater's labels

5. EVALUATIONS USING THE KUSC DATASET

After we collected data for the ground truth and key recognition difficulty, we used the dataset and the annotated data to test several state-of-the-art audio key-finding algorithms as described in this section. The test focuses on five aspects: tuning, audio key finding accuracy against re-labeled ground truth, the relation between an algorithm's accuracy and its confidence value, the relation between an algorithm's accuracy and human perceived key recognition difficulty, and the relation between an algorithm's accuracy and instruments/musical styles. The experiment's results not only provide a detail examination of the dataset, but also offer a baseline to which any newly developed algorithm can be compared in the future.

5.1. Tuning

Figure 8 shows the difference, measured in number of semitones, between the estimated tuning frequencies generated by the Chroma Toolbox [19] and non-negative least squares (NNLS) [20]. It can be observed that the outputs of the two tuning estimation methods are mostly the same (2961 out of 3224 excerpts). However, the two methods reported different estimated tuning frequencies, in some cases more than 0.8 semitones apart.

In addition to examining the difference between the outputs of the two tuning methods, we also compared the estimated tuning frequencies of the first and last 15 seconds of the same composition. Table 4 lists the percentage of excerpts in which the estimated tuning frequencies are different for the beginning and ending segments. The percentage in Table 4 is relatively high based given that the tuning usually stays the same throughout a recording. More studies will be conducted to examine the reasons for the observed discrepancies.

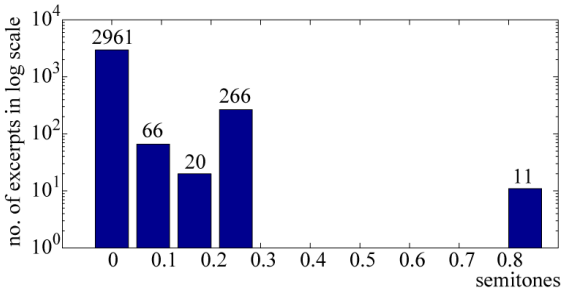


Figure 8. The distance between the tuning frequencies estimated using the Chroma Toolbox [19] and NNLS [20]

Table 4. The percentage of the estimated tuning frequencies that differ in the first and last 15 seconds

Tuning method	Bach	Mozart	Schubert
Chroma toolbox [19]	12.45%	23.22%	17.45%
NNLS [20]	10.65%	21.49%	17.02%

5.2. Audio Key Finding Accuracy

We tested four audio key finding algorithms [5, 16, 17, 18] using the dataset. Since it is not the focus of this paper to optimize the algorithms, we only implemented the basic versions of these algorithms. We used their default settings for parameters such as frame size, hop size, and pitch range to generate frame-by-frame pitch class profiles or chromagrams. We then calculated the average profile over all frames.

For the algorithms described in [16, 17], the average profile was then compared to the pre-defined templates for the 24 keys, and the template with the highest correlation value was then reported as the key. For FACEG [5], the only algorithm that considered pitch spelling, we re-moved the requirement of pitch spelling by considering all spellings for the purpose of comparison. The average profile was used to generate the Center of Effect (CE) representing the tonal center in the 3-dimensional Spiral Array model [1]. The key was determined by searching for the key representation, from among the pre-defined points representing the 24 keys, closest to the CE.

The algorithm in [18] constructs a hidden Markov model with 24 major and minor key states and observations representing chord transitions. The key profiles in the implementation of the algorithm are created from analysis of Bach's Well Tempered Clavier, Book I. The implemented program available via QM Vamp plugin [22] tracks local keys and key changes. For comparison with the other approaches, we needed to select one key from the multiple local keys as the global key for the excerpt. The selection was based on the total duration of the key, i.e., the local key reported with the longest period was chosen to be the global key.

To prevent tuning frequencies that deviated from the norm from affecting the key finding result, we used the output of the tuning estimation methods to specify the reference frequency for the algorithms instead of fixing it at 440Hz for A_4. For conciseness, in this paper, we only report the results using the estimated frequency from the Chroma Toolbox [19].

Table 5 shows the key finding accuracy, overall weighted scores, for the four algorithms. Each weighted score is calculated based on the relation of the output key and the ground truth key (as described in Section 3.4), and converted to a percentage by dividing by the total number of excerpts. Note that the key finding program in the MIR Toolbox [16] reported the highest weighted score, with the HPCP [17] reporting one very close to it. The low percentage in the result of [18] may be due to the fact that the program is too sensitive to local changes. The

availability of the KUSC dataset will allow researchers to further investigate the effect of local key modulations on global key determination.

For each excerpt, we calculated the average of the four scores from the four algorithms. Figure 9 shows the distribution of average weighted scores for the dataset. A peak can be observed at 0.7, with low values on either side. This information is helpful for evaluating a new audio key finding algorithm: excerpts that receive an average score above 0.7 can be used as benchmarks for evaluating the basic performance of a key finding algorithm.

Table 5. The four algorithms' key finding accuracy scores for the KUSC classical music dataset

	MIR	HPCP	FACEG	QM
Weighted score %*	91.7%	90.34%	82.64%	59.74%

*identical = 1, dominant = 0.5, relative = 0.3, parallel = 0.2

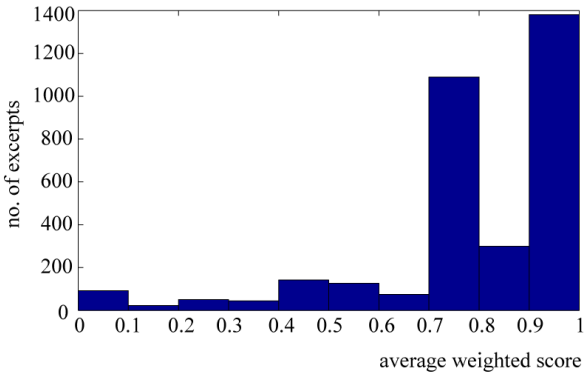


Figure 9. The histogram showing the distribution of average weighted scores of the four algorithms

5.3. Key Strength and Confidence Value

We also used the dataset to examine the relation between key strength/confidence value and key determination accuracy. In addition to the output key, audio key finding algorithms also report a numeric value corresponding to key strength or confidence, a value indicating the certainty of the output key. Such key strength/confidence values have also been used by İzmirli as the weights in a voting scheme to determine the global key from multiple local keys [9].

For the key finding algorithm in HPCP [17] and the MIR Toolbox [16], we used the correlation between the pitch class profile and the 24 keys as the key strength value. The confidence value in FACEG [5] is calculated using the distance between the CE and the two closest keys. Suppose d_1 is the distance between the CE and the closest key and d_2 is the same for the second closest key. The confidence value is calculated as $(d_2 - d_1)/ d_1$, the scaled difference between the distance to the closest and second closest key. The key detector program in the QM Vamp plugin [22] outputs key probabilities for all 24 keys for each frame. In this study, the key strength value is given by the average of the global key's probabilities across all frames.

Table 6 lists the correlation between key strength/confidence value and accuracy as represented by the weighted score. Although some algorithms use this value to determine key, the value is not strongly correlated with the accuracy of the output key.

The correlation coefficients between the key strength/confidence values of pairs of the four algorithms are shown in Table 7. Except for the higher correlation between the key finding programs in the MIR Toolbox and HPCP, both based on the Krumhansl-Schmuckler template-

matching algorithm [2], no strong correlations are observed between any other pairs of algorithms.

Table 6. Correlation between the key strength/confidence value and accuracy scores for the four algorithms.

	MIR	HPCP	FACEG	QM
Correlation	0.2666	0.2922	0.092	-0.0592

Table 7. The four algorithms' key finding accuracy scores for the KUSC classical music dataset

	MIR	HPCP	FACEG	QM
MIR	1	-	-	-
HPCP	0.5479	1	-	-
FACEG	0.1124	0.0653	1	-
QM	0.0633	-0.0144	-0.0030	1

5.4. Key Recognition Difficulty

To understand the connection between challenges faced by the algorithms and musicians, we calculated the correlation between the algorithms' average weighted accuracy score and the annotators' average key recognition difficulty ratings. Although the musicians were instructed to use the 5-point scale for their ratings, they may not use the scale the same way. Therefore, in addition to simply calculating the average of the three ratings, we also normalized and standardized their ratings before calculating the correlation coefficient.

Table 8 shows the correlation results. The values in the table show negative correlation between the difficulty ratings and the average weighted score generated by the four algorithms. However, the correlation is not strong.

Table 8. Correlation between the key recognition difficulty and weighted accuracy scores.

	Unprocessed difficulty	Normalized difficulty	Standardized difficulty
Correlation	-0.3026	-0.2817	-0.2970

Since the correlation in Table 8 was not strong, we further conducted t-tests. Based on the distribution of the average weighted scores as shown in Figure 9, we propose a null hypothesis (H_0) that excerpts with an average weighted score less than 0.7 received the same mean difficulty ratings as those with scores equal to or above 0.7. The alternative hypothesis (H_1) is that excerpts with a score less than 0.7 have a significant lower mean difficulty rating than ones with scores equal to or above 0.7. Results of the t-tests, shown in Table 9, reject the null hypothesis. Thus excerpts receiving a weighted accuracy score higher than 0.7 from the algorithms do have a mean difficulty rating lower than those with weighted accuracy scores below 0.7.

Table 9. The result of the t-tests

	Unprocessed difficulty	Normalized difficulty	Standardized difficulty
H	1	1	1
p-value	1.4808×10^{-8}	1.2505×10^{-7}	2.0356×10^{-8}

5.5. Instrumentation

For excerpts that contain instrument and music style tags as listed in Table 2, we calculated the mean of average weighted scores for each tag to examine how instrumentation impacts the algorithms' accuracy. The result is shown in Figure 10 and it indicates that wind instruments (tags no. 23 – 30), except for the oboe (tag no. 26), have higher average weighted scores than others. It is also observed that excerpts labelled as vocal tracks (tag no. 1 – 7) report higher standard errors than others. However, more data is needed in order to obtain more conclusive results.

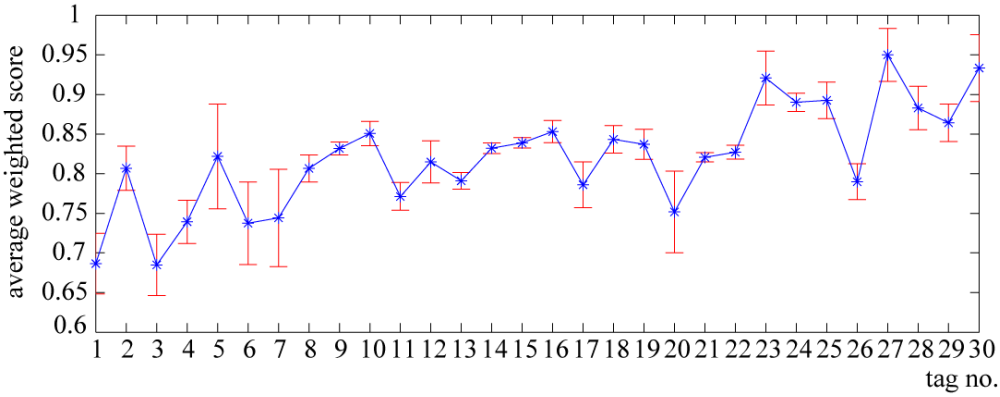


Figure 10. The average weighted scores with standard errors for the 30 instrumentation tags listed in Table 2.

6. CONCLUSIONS

A dataset consisting of 3224 classical music excerpts for audio key finding was described in the paper. To share the dataset with the research community without violating copyright, excerpts were represented as commonly used and irreversible acoustic features such as chromagrams and constant-Q spectrograms instead of audio source files. In addition, manual annotations such as ground truth for the global key and key recognition difficulty were collected. Meta-data relating to instrumentation was also made available.

We conducted a series of experiments with the dataset using several algorithms pertaining to aspects of the audio key finding process. First, for tuning frequency estimation, the two methods tested agreed on the same frequency for over 90% of the excerpts. But 10% to 23% of the beginning excerpts returned a different tuning frequency than that for the corresponding end portion of the same piece. Four audio key finding algorithms were also tested with weighted accuracy scores ranging from 60% to 92%. In addition to the estimated key, many algorithms also produce a numerical value indicating the strength or confidence of the output key. We did not observe clear relations between this key strength/confidence value and the algorithm's accuracy. We also analyzed the relations between the four algorithms' key finding accuracy and musicians' key recognition difficulty ratings. Although accuracy was not strongly correlated with difficulty rating, the result showed that the mean key difficulty rating of excerpts receiving weighted scores above 0.7 is significantly lower than that for excerpts with weighted accuracy scores lower than 0.7. Finally, we examined the relations between the algorithms' accuracy and instruments tags. Some patterns were observed, but these were not sufficient for drawing strong conclusions.

For future work, we plan to expand the dataset by including more excerpts and generating features and annotations for other music information retrieval tasks.

REFERENCES

[1] Papadopoulos, H. & Peeters, G. (2012) "Local key estimation from an audio signal relying on harmonic and metrical structures," IEEE Transaction on Audio, Speech, and Language Processing, Vol. 20, No. 4, pp. 1297–1312.

[2] Krumhansl, C. L. (1990) "Quantifying tonal hierarchies and key distances," Cognitive Foundations of Musical Pitch, chapter 2, pp. 16–49, Oxford University Press, New York.

[3] Temperley, D. (1999) "What's Key for Key? The Krumhansl-Schmuckler Key-Finding Algorithm Reconsidered," Music Perception, Vol. 17, No. 1, pp. 65–100.

[4] Peeters, G. (2006) "Musical key estimation of audio signal based on HMM modeling of chroma vectors," Proceedings of the International Conference on Digital Audio Effects, 127-131.

[5] Chuan, C.-H. and Chew, E. (2005) "Fuzzy analysis in pitch class determination for polyphonic audio key finding," Proceedings of the 6th International Conference on Music Information Retrieval.

[6] Gómez, E. and Herrera, P. (2004) "Estimating the tonality of polyphonic audio files: Cognitive versus machine learning modeling strategies," Proceedings of the 5th International Conference on Music Information Retrieval, 92-95.

[7] Chai, W. and Verco, B. (2005) "Detecting of key change in classical piano music," Proceedings of the 6th International Conference on Music Information Retrieval, 468-473.

[8] Chuan, C.-H. and Chew, E. (2006) "Audio key finding: considerations in system design, and case studies on 24 Chopin's preludes," Ichiro Fujinaga, Masataka Goto, George Tzanetakis (eds), EURASIP Journal on Applied Signal Processing, Special Issue on Music Information Retrieval.

[9] İzmirli, Ö. (2005) "Template based key finding from audio," Proceedings of the International Computer Music Conference.

[10] İzmirli, Ö. (2007) "Localized key finding from audio using non-negative matrix factorization for segmentation," Proceedings of the 8th International Conference on Music Information Retrieval, 195-200.

[11] Schuller, B. & Gollan, B. (2012) "Music theoretic and perception-based features for audio key determination," Journal of New Music Research, Vol. 41, No. 2, pp. 175–193.

[12] Sun, J., Li, H., & Ma, L. (2011) "A music key detection method based on pitch class distribution theory," International Journal of Knowledge-based and Intelligent Engineering Systems, Vol. 15, pp. 165–175.

[13] Weiss, C. (2013) "Global key extraction from classical music audio recordings based on the final chord," Proceedings of the Sound and Music Computing Conference, pp. 742–747.

[14] Müller, M., Konz, V., Bogler, W. and Arifi-Müller, V. (2011) "Saarland music data," Proceedings of the 12th International Conference on Music Information Retrieval.

[15] Pauws, S. (2004) "Musical key extraction from audio," Proceedings of the 5th International Conference on Music Information Retrieval.

[16] Lartillot, O. & Toiviainen, P. (2007) "Matlab toolbox for musical features extraction from audio," Proceedings of the 10th International Conference on Digital Audio Effects, pp. 237–244.

[17] Gómez, E. (2006) Tonal Description of Music Audio Signals, PhD thesis.

[18] Noland, K. & Sandler, M. (2007) "Signal processing parameters for tonality estimation," Proceedings of Audio Engineering Society 122nd Convention.

[19] Müller, M. & Ewert, S. (2011) "Chroma toolbox: matlab implementations for extracting variants of chroma-based audio features," Proceedings of the 12th International Society for Music Information Retrieval Conference, pp. 215–220.

[20] Mauch, M. & Dixon, S. (2010) "Approximate note transcription for the improved identification of difficult chords," Proceedings of the 11th International Society for Music Information Retrieval Conference, pp. 135–140.

[21] Mauch, M. (2010) Automatic Chord Transcription from Audio Using Computational Models of Musical Context, PhD thesis, Queen Mary University of London.

[22] www.vamp-plugins.org/plugin-doc/qm-vamp-plugins.html#qm-keydetector, accessed July 2014.

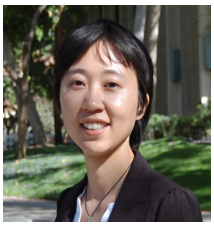
[23] Brown, J. (1991) "Calculation of a constant Q spectral transform," Journal of the Acoustical Society of America, Vol. 89, No. 1, pp. 425–434.

[24] Chuan, C.-H. and Chew, E. (2012) "Generating ground truth for audio key finding: When the title key may not be the key," Proceedings of the 13th International Conference on Music Information Retrieval.

[25] Chew, E. (2000) "Towards a mathematical model of tonality," doctoral dissertation, Department of Operations Research, Massachusetts Institute of Technology, Cambridge, Mass, USA.

[26] Chew, E. (2001) "Modeling tonality: applications to music cognition," Proceedings of the 23rd Annual Meeting of the Cognitive Science Society, pp. 206–211.
[27] Chuan, C.-H. and Charapko, A. (2013) "Predicting key recognition difficulty in polyphonic audio," Proceedings of the 9th IEEE International Workshop on Multimedia Information Processing and Retrieval.

AUTHORS



Ching-Hua Chuan is an assistant professor of computing at Universityof North Florida. She received her Ph.D. in computer science from University of Southern California Viterbi School of Engineering in 2008. She received her B.S. and M.S. degrees in electrical engineering from National Taiwan University. Dr. Chuan's research interests include audio signal processing, music information retrieval, artificial intelligence and machine learning. She was the recipient of the best new investigator paper award at the Grace Hopper Celebration of Women in Computing in 2010.

Elaine Chew received the B.A.S. degree in mathematical and computational sciences with honors, and in music with distinction, from Stanford University, Stanford, California, in 1992, and the S.M. and Ph.D. degrees in operations research from the Massachusetts Institute of Technology, Cambridge, Massachusetts, in 1998 and 2000, respectively, in the United States. She joined Queen Mary, University of London, in the United Kingdom, as Professor of Digital Media in Fall 2011, where she serves as Director of Music Initiatives in the Centre for Digital Music. Her research interests center on mathematical and computational modelling of music prosody and structure, on which she has authored over 80 refereed journal and conference articles. Prof. Chew was the recipient of a National Science Foundation Faculty Early Career Award, Presidential Early Career Award in Science and Engineering, and Edward, Frances, and Shirley B. Daniels Fellowship at the Radcliffe Institute for Advanced Study.

.