# CREATING GROUND TRUTH FOR AUDIO KEY FINDING: WHEN THE TITLE KEY MAY NOT BE THE KEY

**Ching-Hua Chuan**
University of North Florida
School of Computing
c.chuan@unf.edu

**Elaine Chew**
Queen Mary, University of London
Centre for Digital Music
elaine.chew@eecs.qmul.ac.uk

## ABSTRACT

In this paper, we present an effective and efficient way to create an accurately labeled dataset to advance audio key finding research. The MIREX audio key finding contest has been held twice using classical compositions for which the key is designated in the title. The problem with this accepted practice is that the title key may not be the perceived key in the audio excerpt. To reduce manual annotation, which is costly, we use a confusion index generated by existing audio key finding algorithms to determine if an audio excerpt requires manual annotation. We collected 3224 excerpts and identified 727 excerpts requiring manual annotation. We evaluate the algorithms' performance on these challenging cases using the title keys, and the re-labeled keys. The musicians who aurally identify the key also provide comments on the reasons for their choice. The relabeling process reveals the mismatch between title and perceived keys to be caused by tuning practices (in 471 of the 727 excerpts, 64.79%), and other factors (188 excerpts, 25.86%) including key modulation and intonation choices. The remaining 68 challenging cases provide useful information for algorithm design.

## 1. INTRODUCTION

The typical trend in technology development is for systems proposed later to outperform earlier ones, but this does not seem to have been the case for audio key finding, judging by the results of the MIREX audio key finding contest. The first MIREX audio key finding contest was held in 2005, and the second contest took place six years later. The same dataset was used in the two contests, and based on the numbers, the systems in MIREX 2011 seem to have performed worse than the ones in 2005 on average. This points to the fact that the statistics of the contest results alone have not provided sufficient information for future researchers to develop better systems. If the goal of the contest is to move the research community forward, a detailed examination of the results is required.

An effective way to improve audio key finding is to examine the cases in the dataset for which most systems have difficulties. For this paper, we constructed a dataset with 3324 music audio recording excerpts, 2.6 times the number in the MIREX dataset. It is worth noting that this dataset created from actual music recordings is distinct from the MIREX dataset synthesized from MIDI files. We implemented five existing audio key finding systems and tested them on the dataset. Using the title key as ground truth, let the confusion index, *I*, be the number of systems that disagree with this ground truth. We then extracted a subset of the data consisting of excerpts for which no more than two systems reported correct answers, i.e. $I \geq 3$. This subset, called the challenging set, was re-examined by three professional musicians and their keys manually labeled. By comparing the relabeled keys with the title keys, we observe reasons why the excerpt's perceived key might be different from the title key. We also present the musicians' comments about their annotations to show the factors that impact audio key finding. Finally, we describe some controversial audio key finding cases for which we received three conflicting answers.

The paper is organized as follows. Section 2 provides the background of MIREX audio key finding contests. Section 3 presents the five audio key finding systems implemented for the study. In Section 4, we describe the experiment design, followed by a detailed examination of the experiment results. We state our conclusions and suggestions for future work in Section 5.

## 2. BACKGROUND

In MIREX 2005, Chew, Mardirossian and Chuan proposed contests for symbolic and audio key finding [7]. For audio key finding, a wave file is given as input and one answer including a key name (for example, "C") and a mode (such as "major") is expected as output. The ground truth used in the contest is the key defined by the composer in the title of the piece, as is the practice in key finding research. Each output key is compared to the ground truth and assigned a score as follows: 1 point 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. In the contest, 1252 audio files synthesized from MIDI were used as the test dataset, consisting of symphonies from various time periods. The best system achieved an 87% correct rate, with a composite score of 89.55% [8]. The second audio key finding/detection contest was held in 2011 [9]. The same dataset was used and same evaluation method was employed. The best system achieved a 75% correct rate and weighted score of 82%.

## 3. AUDIO KEY FINDING SYSTEMS

In this section we first describe the general structure of an audio key finding system, then the five systems implemented in this study. Systems that rely on training data are not implemented because the title key (ground truth) may not be the key of the excerpt.

### 3.1 A General Structure

Figure 1 illustrates a general structure of an audio key finding system. The system can be divided to two major parts as shown in the dashed boxes. The components in the left dashed box are designed to transform low-level acoustic features such as frequency magnitude into high-level music information such as pitches or pitch classes. Some audio key finding systems start with some preprocessing of the data, such as the removal of noise and silence. The next two steps, windowing and Fast Fourier Transform (FFT), indicate a basic approach for spectral analysis. After spectral analysis, the step labeled Pitch Class Profile (PCP) generation converts the frequency spectrum information into a pitch class distribution called a pitch class profile. This is often the step where most audio key finding systems differ.

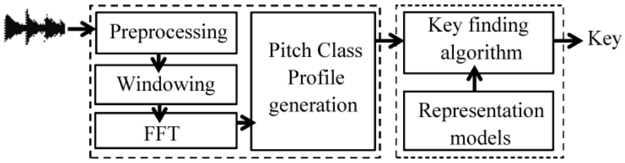


**Figure 1**. General structure of audio key finding systems.

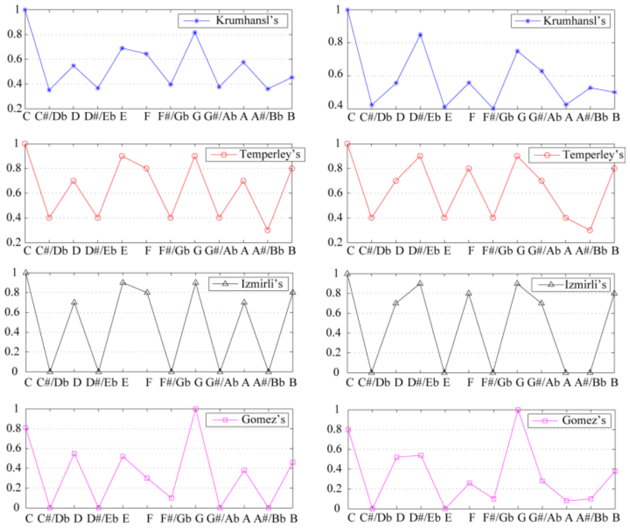
After a pitch class profile is generated, it is then compared with 24 predefined key templates or profiles, 12 major and 12 minor, to determine the key in the audio excerpt. This step is shown in the right dashed box in Figure 1, consisting of two components: a key finding algorithm and a representation model of keys. A representation model provides a typical pitch class distribution for a specific key. The key profiles produced by the representation model are then used to determine the key in a key finding algorithm. For example, a simple key finding algorithm calculates the correlation between the pitch class profile of the audio excerpt and the 24 key profiles, and the key profile that reports the highest correlation value is selected as the key.

In the following sub sections, we describe the audio key finding systems implemented in this study in detail, emphasizing the uniqueness of each system.

### 3.2 Krumhansl's and Temperley's Key Profiles

Tonality has been studied extensively not only in music theory, but also from many other perspectives. In 1986, Krumhansl and Schmuckler [6] developed a widely accepted model called the probe tone profile method (referred to as the K-S model for convenience), which constructs pitch class profiles for major and minor keys by using user ratings from probe tone experiments. In 1999, Temperley [10] improved upon the K-S model by modifying the key profiles to emphasize the differences between diatonic and chromatic scales. Temperley also adjusted the weights of the forth and seventh pitches so as to differentiate between keys with highly similar pitch class signatures. The key profiles generated by the K-S model and Temperley are shown in Figure 2.

These key profiles can be directly used to build a symbolic key finding system. In this study, we added an audio signal processing module to these key profiles, and created two base audio key finding systems for comparison. The audio signal processing module consists of a silence removal preprocessing step, rectangular windowing with non-overlapped frames, FFT, and generation of a PCP using a uniform weighting function across the frequency range of a pitch and also for all pitches folded into 12 pitch classes.



(a) C major key profiles

(b) C minor key profiles

**Figure 2**. C major and C minor key profiles proposed by Krumhansl, Temperley, Izmirli and Gómez.

### 3.3 Izmirli's System

Izmirli proposed a template-based correlation model for audio key finding in [5]. During the PCP generation step shown in Figure 1, called chroma template calculation in Izmirli's system, pitches are weighted using a decreasing function that gives low-frequency pitches more weight. In Izmirli's system, he constructed key templates from monophonic instruments samples, weighted by a combination of the K-S and Temperley's modified pitch class profiles, as shown in Figure 2. Izmirli's system also tracks the confidence value for each key answer, and the global key is then selected as the one having the highest sum of confidence values over the length of the piece.

### 3.4 Gómez's HPCP

In [5], Gómez detected pitches using three times the standard resolution of the pitch frequency spectrum of the FFT method, and distributed the frequency values among the adjacent frequency bins using a triangular weighting function to reduce boundary errors. A Harmonic Pitch Class Profile (HPCP) is generated as input to the key finding algorithm, using a modified version of Krumhansl's key templates as shown in Figure 2.

### 3.5 Chuan and Chew's FACEG

In [3], Chuan and Chew proposed an audio key finding system called Fuzzy Analysis Center of Effect Generator (FACEG). 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 refined periodically using the current key information. The representation model used in the system is Chew's Spiral Array model [1, 2], 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 in real-time: an instantaneous key answer is generated in each window based on past information. It is the only model amongst the ones considered with pitch spelling.

## 4. EXPERIMENT DESIGN

In this section we describe the manner in which the dataset is prepared, and the experiment design for exploring the reasons for the challenging cases.

### 4.1 Data Collection and Selection

The dataset used in this study is provided by Classical KUSC, a classical public radio station. The entire dataset consists of over 40,000 audio recordings of classical performances. For this study, we selected compositions by Bach, Mozart and Schubert. We chose these three composers' work because (1) tonality is generally more clearly defined in these pieces than in more recent compositions; and (2) the composers represent three different styles with distinguishable levels of tonal complexity. We further refined the dataset by filtering out the recordings that do not have key information in the title. For multi-movement works, we used only the first and last movement, because they are generally in the title key.

As a result, the dataset we used in this study consists of 1662 recordings of varying lengths. Similar to the evaluation procedure in the MIREX 2005 and 2011 audio key finding contests, 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 reserved the beginning and end sections of a recording because these two sections are more likely to be in the key shown in the title. As a result, we ended up with a total of 3324 different excerpts in the experiment.

### 4.2 Evaluation Method

To improve the performance of existing audio key finding systems, some more detailed examination of the challenging cases is necessary. An excerpt is considered challenging if no more than two systems out of the five implemented reported the key identical to the one in the title. A challenging set was thus built from such cases.

The excerpts from the challenging cases were given to two professional musicians for key annotations. During the process, the two musicians were provided with the 15-second long excerpts instead of the entire pieces, to ensure that they have the same acoustic information as the systems. No key labels or titles were revealed to the musicians. The only information other than the audio excerpt provided is the name of the composer.

The musicians were asked to write down one answer as the global key for each excerpt, based on the 15 seconds they heard. They were also asked to comment on the reasons behind their answers, particularly for excerpts that they felt were difficult to annotate. When the key annotations by the two musicians differed, the excerpt was given to a third professional musician for relabeling. The final relabeled key was determined by majority vote from the three annotations.

## 5. EXPERIMENT RESULTS

### 5.1 Results Using Title Keys as Ground Truth

Out of the 3324 excerpts, there were 727 excerpts (21.87%) for which no more than two systems reported answers identical to the title keys, i.e. $I \geq 3$. We focused on these 727 excerpts, the challenging cases in this study, to examine the difficulties most key finding systems encounter. Table 1 shows the distribution of the challenging set, in absolute numbers and as a percentage of the number of excerpts we considered by each composer.

| 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%) |

**Table 1.** Details of the entire data set and the challenging set by Bach, Mozart and Schubert.

We divided the reported key into 9 categories based on its relation to the title key: correct, dominant (Dom), subdominant (SubD), parallel major/minor (Par), relative major/minor (Rel), same mode with the root one half-step higher (Mode +1), same mode with the root one half-step lower (Mode – 1), same mode but not in the previous categories (Mode Others), and the rest of relations not included in any of the previous categories (Others).

Figure 3 shows the results for the Bach challenging set in (a) first 15 seconds and (b) last 15 seconds respectively. It can be observed that most of the incorrect answers reported by the systems fall into the last three categories, especially in the category (Mode – 1), indicating that tuning may be an issue in key finding for Bach's pieces.

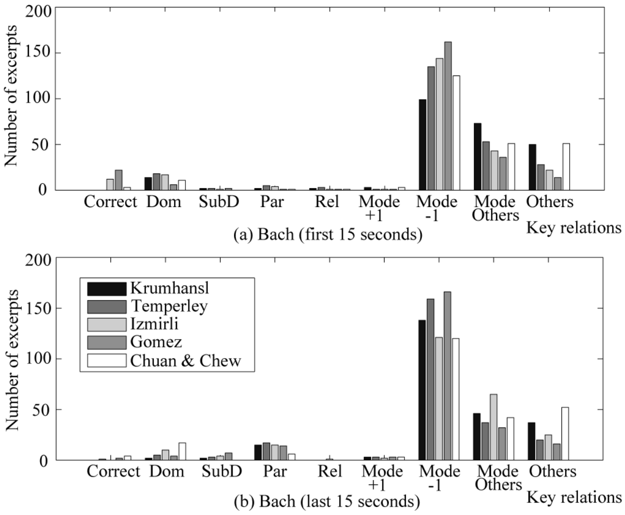


**Figure 3**. Key finding results for the challenging Bach dataset using the title key as ground truth.

Figure 4 shows the results for the Mozart challenging set in the (a) first and (b) last 15 seconds, respectively. In Figure 4 (a), similar to the results in Figure 3 (a), the last three categories account for the majority of the results in the first 15 seconds. However, unlike Figure 3 (b), the parallel major/minor (Par) category accounts for a significant proportion of the results in Figure 4 (b). The reported keys are also more evenly distributed than in Figure 3.

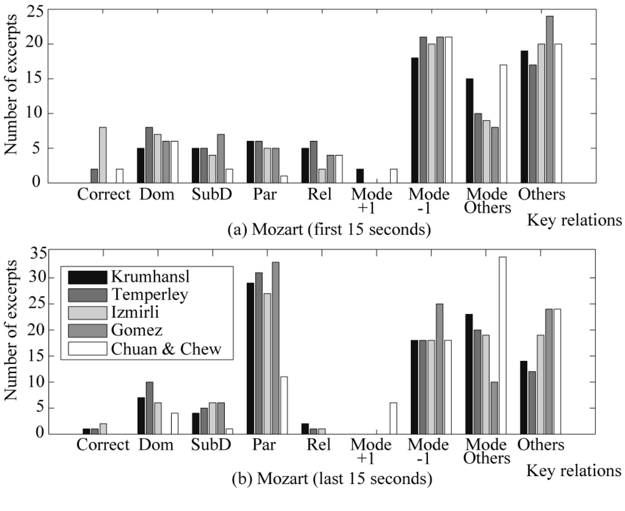


**Figure 4**. Key finding results for the challenging Mozart dataset using the title key as ground truth.

Figure 5 shows the results for the Schubert challenging set in the (a) first and (b) last 15 seconds, respectively. The results of Schubert challenging set are more similar to Mozart's than Bach's, as the results are more evenly distributed in the first 15 seconds and the parallel major/minor dominates in the last 15 seconds. One distinct feature observed in the Schubert results is that the (Mode – 1) category is much less significant.

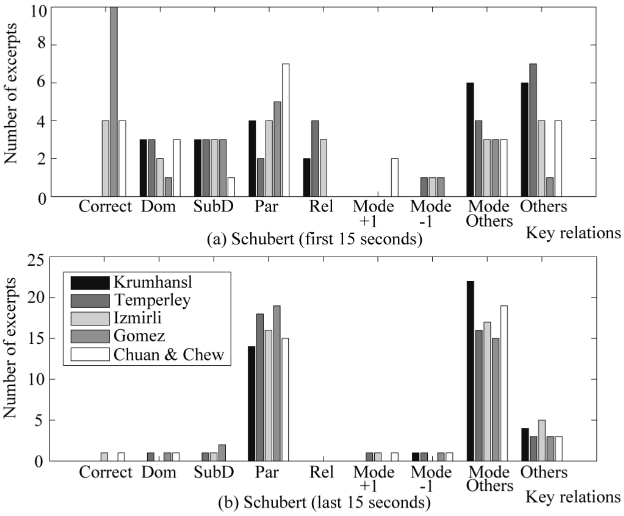


**Figure 5**. Key finding results for the challenging Schubert dataset using the title key as ground truth.

### 5.2 Results Using Re-labeled Keys as Ground Truth

Table 2 shows the statistics of re-labeled keys in relation to title keys. The tuning category consists of re-labeled keys one half step away from title keys, while the other category includes relabeled keys that are neither identical to title keys nor in the tuning category.

| Composer | First 15 seconds | | Last 15 seconds | |
|---|---|---|---|---|
| | tuning | other | tuning | other |
| Bach | 183 (74.7%) | 36 (14.7%) | 182 (74.6%) | 54 (22.1%) |
| Mozart | 48 (64%) | 16 (21. 3%) | 55 (56.1%) | 38 (38. 8%) |
| Schubert | 1 (4.2%) | 8 (33. 3%) | 2 (4.9%) | 36 (87.8%) |

**Table 2.** Relations between title keys and re-labeled keys.

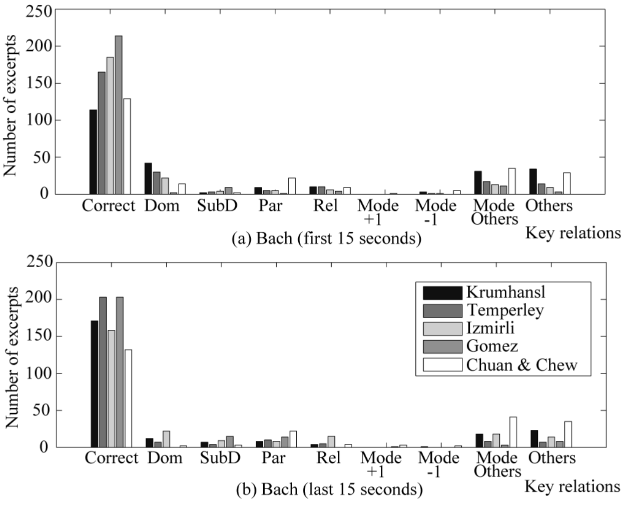


**Figure 6**. Key finding results for the challenging Bach dataset using the re-labeled key as ground truth.

Figure 6 shows the results for the Bach challenging set in the (a) first 15 seconds and (b) last 15 seconds using the relabeled keys as ground truth. Comparing the results in Figure 3 with those in Figure 6, it is clear that many of the recordings of Bach's compositions are not tuned to modern definitions of the title key. Pitches ranged from one quarter to one half-step lower than what one might expect in modern tuning. Therefore, the cases labeled as (Mode – 1) in Figure 3 could be considered correct. This also points to the importance of verifying the title keys manually for the audio key finding. However, it is debatable whether the key should be relabeled based on modern tuning. For example, a recording may be recognized as being in the key of B major according to modern tuning, but B major is a very uncommon key in Baroque music and some musicians still prefer to call it C major despite the flattened tuning.

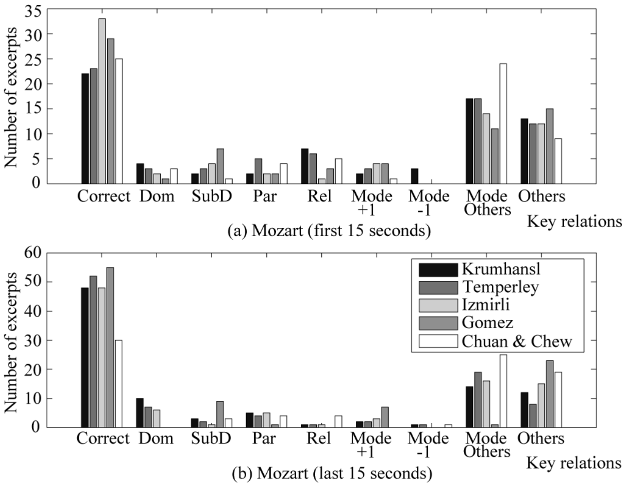


**Figure 7**. Key finding results for the challenging Mozart dataset using the re-labeled key as ground truth.

Figure 7 shows the results for the Mozart challenging set using the relabeled keys as ground truth in the (a) first 15 and (b) last 15 seconds. Observe that the number of correct answers is increased, indicating that the first and last 15 seconds of these pieces are actually in a key other than the title key. By comparing Figure 7 (a) and Figure 4 (a), we observe that the increase in correct answers in Figure 7 (a) mainly results from decreasing numbers in the four categories: Mode – 1, Others, Dominant (Dom) and Parallel (Par). This shows that for Mozart, a piece may start in a related parallel major/minor or even a foreign key. For the last 15 second excerpts, a piece may end in a parallel major/minor key. The tuning problem, indicated by the (Mode – 1) category, can still be observed in Mozart's recordings in both the first and last 15 seconds.

Figure 8 shows the results on the Schubert challenging set in the (a) first 15 and (b) last 15 seconds. The number of correct answers does not increase much in Figure 8 (a) comparing to Figure 5 (a), and the increment in Figure 8 (a) is the result of decrement in the Parallel major/minor (Par) category. When the relabeled keys are used as the ground truth, almost all the systems recognize the keys correctly in the last 15-second excerpts as shown in Figure 8 (b). This result shows that Schubert's pieces may end in the parallel major/minor key, or even some more distant keys in the same mode.

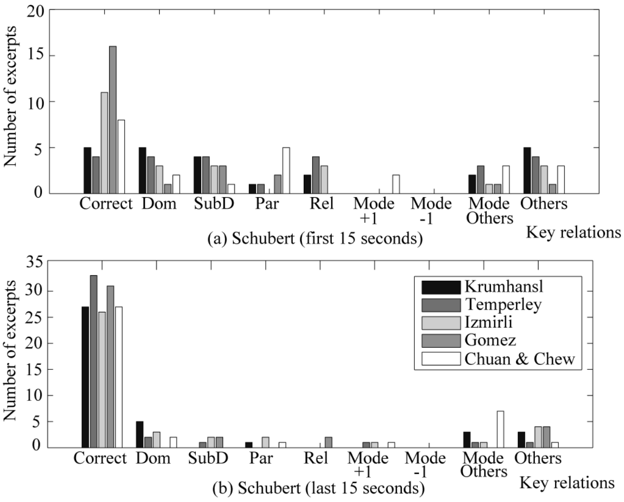


**Figure 8**. Key finding results for the challenging Schubert dataset using the re-labeled key as ground truth.

### 5.3 Musicians' Comments and Case Studies

In this section we present the musicians' comments alongside their answers, and excerpts where they disagreed with each other.

The musicians were encouraged to write down comments with their answers but were not restricted in terms of the words they can use. Table 3 shows the most frequently used keywords and their meanings.

| Keywords | Meanings | Num. of occurrence |
|---|---|---|
| sharp/flat | Notes sound sharp/flat compared to modern tunings | 79 |
| easy | Clear V-I chord progression | 66 |
| picardy 3rd | A piece in a minor key that ends in the parallel major | 28 |
| modulation | A piece changes from one key to another | 16 |
| tough/tricky | Difficult to determine the key, mostly due to missing cadence | 15 |
| cadence | Cadence cut off; cadence spotted in the middle of the piece; misleading cadence | 4 |

**Table 3.** Keywords with meanings and number of occurrences in musicians' comments.

Among 727 excerpts, there are only 8 excerpts in which all three musicians disagreed on the key. Table 4 gives the details of the 8 excerpts, the musicians' annotated keys, and our notes on why these excerpts might have confused the annotators.

## 6. CONCLUSIONS AND FUTURE WORK

In this paper we presented an approach to effectively and efficiently develop a well-annotated dataset for audio key finding. Having a well-annotated dataset is essential for any kind of algorithm testing and development, but it is very time-consuming to create one with numerous examples. In this paper we implemented five audio key finding systems, and used them to select the examples that require manual examination. Three professional musicians re-labeled the keys for these difficult cases.

| Composer | Recording (excerpt)/ Performer | Title key | Relabeled keys |
|---|---|---|---|
| Bach | Cello Suite #3 BWV 1009 (last 15 secs)/Yo-Yo Ma | C major | C major,<br>B major,<br>C minor |
| | Cello. Briefly in minor mode at the beginning, but ends unequivocally in C major; pitches flat. | | |
| Schubert | Moments Musicaux: #6 Op 94 (last 15 secs)/David Fray | Ab major | Eb major,<br>G# major,<br>A major |
| | Piano. In Ab major; tuning sharp. Annotator 1 misled by Bb's in beginning. | | |
| Mozart | String Quartet #16 K428 (first 15 secs)/Quartetto Italiano | Eb major | F minor,<br>C minor,<br>Eb major |
| | String Quartet. Chromatic start and notes following led to ambiguity in mode; ends clearly in Eb major. | | |
| Mozart | String Quartet #16 K428 (first 15 secs)/Quatuor Mosaiques | Eb major | B minor,<br>D major,<br>Eb minor |
| | Same piece as above; annotations completely different. | | |
| Mozart | String Quartet #19 K465 (first 15 secs)/ Quatuor Mosaiques | C major | C minor,<br>B minor,<br>unsure |
| | String Quartet. In C but Eb in vln 2 and flat A in vln 1 (an intonation choice) led to perceived minor mode. | | |
| Mozart | Gran Partita Serenade K361 (last 15 secs)/Octophorus | B major | Bb major,<br>D major,<br>A major |
| | Strings. Tuning flat, which explains the Bb and A. | | |
| Mozart | Symphony #22 K162 (last 15 secs)/Amsterdam Baroque Orchestra | C major | B major,<br>E major,<br>D major |
| | Orchestra. Tuning flat, which explains the B. | | |
| Mozart | Requiem K626 (last 15 secs)/Vienna Philharmonic | D minor | C major,<br>unsure,<br>F major |
| | Voices/Str/Winds/Perc. Flat; insufficent information. | | |

**Table 4.** Information of the excerpts where three musicians disagree on the key.

By examining the relabeled keys, we discovered potential causes for the difficulties and make the following observations and recommendations:

(a) **tuning:** In recorded performances, different tuning or intonation choices can cause confusion. Evaluations could either account for all possible categorical key name interpretations (e.g. flat C might be interpreted as B), or allow for tuning (and letter name) independent key finding, for example by requiring systems to locate the most stable tone.

(b) **modulations:** some excerpts may not be entirely in one key, modulating midstream or ending with a Picardy 3rd. These excerpts could either be removed, or more nuanced ground truth created with key change annotations.

(c) **missing cadences:** a key is theoretically established when a complete cadence confirms its identity. Many excerpts from the first 15 seconds of pieces may not have these cadences. Care could either be taken to make sure these cadences exist in the evaluation samples, or the scoring system could account for levels of difficulty of assessing key based on annotators' notes.

The established dataset and annotations, and the process of collecting them, will benefit the audio key finding community and MIREX contests. While we cannot publicly share the music files, we will post the results of the annotations online. Future plans include augmenting the dataset with automatically generated key labels.

## 7. ACKNOWLEDGEMENTS

We like to thank the annotators, Fred Bertino, Victor Hormilla, and Jaime Kaufman, for their time and effort.

## 8. REFERENCES

[1] E. Chew: *Towards a Mathematical Model of Tonality*, Doctoral dissertation, Department of Operations Research, Massachusetts Institute of Technology, Cambridge, Mass, USA, 2000.

[2] E. Chew: "Modeling Tonality: Applications to Music Cognition," in *Proc. of the 23rd Annual Meeting of the Cognitive Science Society*, pp. 206–211, Edinburgh, Scotland, UK, 2001.

[3] C.-H. Chuan and E. Chew: "Fuzzy Analysis in Pitch-Class Determination for Polyphonic Audio Key Finding," in *Proc. of the 6th International Conference on Music Information Retrieval*, pp. 296–303, London, UK, 2005.

[4] E. Gómez, "Tonal Description of Polyphonic Audio for Music Content Processing," INFORMS *Journal on Computing*, summer 2006, Vol. 18, No. 3, pp. 294–304, 2006.

[5] Ö. İzmirli, "Template Based Key Finding from Audio," in *Proc. of the International Computer Music Conference*, Barcelona, Spain, 2005.

[6] C. L. Krumhansl, "Quantifying Tonal Hierarchies and Key Distances," in *Cognitive Foundations of Musical Pitch*, chapter 2, pp. 16–49, Oxford University Press, New York, USA, 1990.

[7] MIREX 2005 Audio Key Finding Contest, www.music-ir.org/mirex/wiki/2005:Audio_and_Symbolic_Key

[8] MIREX 2005 Audio Key Finding Contest Results, www.music-ir.org/mirex/wiki/2005:Audio_Key_Finding_Results

[9] MIREX 2011 Audio Key Detection Contest, nema.lis.illinois.edu/nema_out/mirex2011/results/akd/index.html

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