AN EXPERT GROUND-TRUTH SET FOR AUDIO CHORD RECOGNITION AND MUSIC ANALYSIS

John Ashley Burgoyne **Jonathan Wild** **Ichiro Fujinaga**
Centre for Interdisciplinary Research in Music Media and Technology
McGill University, Montréal, Québec, Canada
{ashley,jon,ich}@music.mcgill.ca

ABSTRACT

Audio chord recognition has attracted much interest in recent years, but a severe lack of reliable training data—both in terms of quantity and range of sampling—has hindered progress. Working with a team of trained jazz musicians, we have collected time-aligned transcriptions of the harmony in more than a thousand songs selected randomly from the *Billboard* "Hot 100" chart in the United States between 1958 and 1991. These transcriptions contain complete information about upper extensions and alterations as well as information about meter, phrase, and larger musical structure. We expect that these transcriptions will enable significant advances in the quality of training for audio-chord-recognition algorithms, and furthermore, because of an innovative sampling methodology, the data are usable as they stand for computational musicology. The paper includes some summary figures and statistics to help readers understand the scope of the data as well as information for obtaining the transcriptions for their own research.

1. WHY CHORDS?

Ever since Alexander Sheh and Dan Ellis's first foray into recognizing musical chords directly from audio [11], this challenging problem has fascinated researchers at ISMIR. From the beginning, however, the challenges have been more than just engineering: there has not been nearly enough labelled, time-aligned data to train reliable recognizers. Sheh and Ellis worked with just twenty songs. Gradually, more data has become available, most famously Christopher Harte's transcriptions of the entire output of the Beatles [8], but even the most recent Music Information Retrieval Evaluation Exchange (MIREX) contests [1] have had only 210 songs available [10]. Some researchers have tried to circumvent the problem by synthesizing audio from MIDI [9], but there has remained a significant interest in developing a larger, human-annotated data set of chords from commercial recordings.

Audio chord recognition is not the only use for a larger data set. The analysis of harmony in popular music has been drawing more and more attention from music theorists [2, 6]. Due to the limitations on the amount of available data, these analyses and theories are usually based on a very limited number of examples and cannot be generalized with statistical guarantees of accuracy. A large-scale empirical analysis of harmony in popular music would be an enormous contribution to musicology, but such analysis would require not only more data, just as audio chord recognition does, but also a wider range of data. Of the 210 songs in the MIREX data set, 174 (83 percent) are by the Beatles. While that may be admirable in terms of musical quality, it makes it impossible to draw more general conclusions about how harmony operated in the music of other artists and other periods. We believe that a single, well-conceived data set can address the needs of both communities.

We are pleased to announce the release of a new data set that comprises detailed transcriptions of the chords in more than one thousand songs selected at random from *Billboard* magazine's "Hot 100" charts. Each transcription represents the combined opinion of three or more experts in jazz and popular music, and the chord symbols have been time-aligned with the musical meter and with commercially available audio recordings. This paper describes the methodology for selecting songs (section 2), explains the process used to transcribe them (section 3), and presents some basic descriptive statistics to help readers understand how they might use these data (section 4). In addition to the contribution of the data set, we hope that information about how we produced them—a process that was considerably more involved than we had originally expected—will benefit other research groups who are interested in transcribing still more chords themselves.

[1] http://www.music-ir.org/mirex/

2. THE *BILLBOARD* "HOT 100"

The *Billboard* "Hot 100" is a weekly compilation of the most popular music singles in the United States, all genres included, based on a combination of radio airplay and retail sales (and more recently, digital downloads). [2] The "Hot 100" has been published continuously in *Billboard* magazine since 4 August 1958, replacing earlier charts like "Best Sellers in Stores," "Most Played by Jockeys," and "Most Played in Jukeboxes." Although it is far from a perfect representation of popularity, like any ranking, it is generally regarded to be the gold standard among charts of popular music in North America [4]. Because it includes all genres, it seemed particularly well-suited to the goals of training broadly-applicable chord recognizers and drawing broadly-applicable musicological conclusions. It has also been the basis for several previous attempts to draw statistical conclusions about the behavior of popular singles over time [1, 4, 7].

2.1 Sampling Methodology

The date of the first chart, 4 August 1958, is a natural starting date for selecting songs, but choosing an end date is less straightforward. Hip-hop music does not lend itself readily to harmonic analysis as traditionally understood, and because hip-hop became more popular in the 1990s and 2000s, a larger portion of the music on the "Hot 100" chart from these periods falls out of the scope of the data set. Furthermore, there have been several changes to the formula for computing the "Hot 100" over time, including a particularly significant shift in December 1991, when the data for generating the charts shifted from being self-reported to being generated automatically through Nielsen's BDS and SoundScan system. [3] After this date, songs tended to stay on the charts for so much longer than before that *Billboard* established limits on how many weeks any given single would be allowed to remain on the "Hot 100" chart, added a "Recurrent Singles" chart to capture singles knocked off the chart due to the new rule, and has averaged songs pre-1991 differently from those post-1991 when generating historical summaries like the "50th-Anniversary" charts [3]. We chose to restrict our sample to charts prior to December 1991 in order to avoid these problems.

As stated earlier, our goal in constructing this data set was not only to provide a higher-quality set for audio chord recognition but also to provide a data set that would be useful for computational musicology and the analysis of popular music. As such, it was important to choose a sample of songs that would allow for general questions about how popular music and the factors that made it popular evolved throughout the latter half of the twentieth century. Like most projects,

1. Divide the set of all chart slots into three eras:
   (a) 4 August 1958 to 31 December 1969,
   (b) 1 January 1970 to 31 December 1979, and
   (c) 1 January 1980 to 30 November 1991.
2. Subdivide the chart slots in each era into five subgroups corresponding to quintiles on the chart:
   (a) ranks 1 to 20,
   (b) ranks 21 to 40,
   (c) ranks 41 to 60,
   (d) ranks 61 to 80, and
   (e) ranks 81 to 100.
3. Select a fixed percentage p of possible chart slots at random from each era-quintile pair.
4. For each selected chart slot:
   (a) attempt to acquire the single at the target slot;
   (b) if that fails, toss a virtual coin to choose between either the single directly above or directly below the target slot on the chart from the same week;
   (c) if that fails, choose the single that was not selected by the coin toss in 4b;
   (d) if that fails, toss a virtual coin to choose between either the single two ranks above or two ranks below the target single on the chart from the same week;
   (e) if that fails, choose the single that was not selected by the coin flip in 4d; and
   (f) if that fails, consider the chart position to be a missing data point.

Figure 1. Sampling algorithm for the *Billboard* "Hot 100." The algorithm is designed to minimize the distortion from "convenience sampling" while reducing the expense of collecting an audio collection. We believe that this algorithm yields a data set that, as cost-effectively as possible, is valid for drawing conclusions about relative positioning and changes in the behavior of music on the charts over time.

[2] http://www.billboard.com/charts/hot-100

[3] http://nielsen.com/us/en/industries/media-entertainment.html

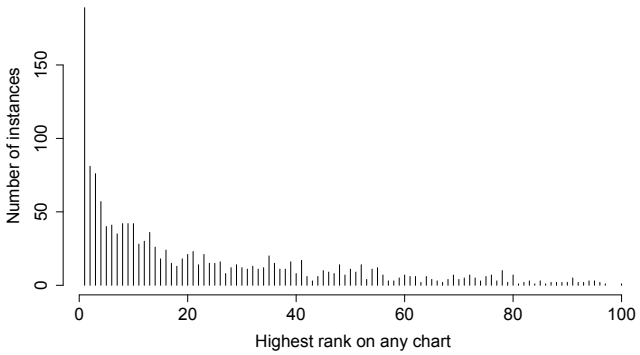

Figure 2. Histogram of the highest rank achieved on any chart among singles in the random sample. Because of the behavior of popular songs—namely that they tend to stay on the chart for a long time and rise and fall through different ranks—our sampling method still weighs the most popular songs more heavily. We consider this behavior desirable.

however, the budget was limited, and we wanted to make the best use possible of the recordings we already had available without unduly biasing the final data set. In consultation with a professional statistician, we devised the sampling methodology detailed in figure 1. The first two steps guarantee that even the most unfavorable random draw would still provide some information about time and chart position. The final step balances the desire to maximize use of recordings on hand with the need to achieve a sample that is representative of the underlying charts; it works on the assumption that singles within two chart positions of each other in any given week should behave similarly. In limit of an infinite number of samples drawn in this way, one would expect to retrieve all recordings on hand weighted proportionally to their behavior on the charts. The more recordings of "missing" chart positions that one acquires later, the more accurately the final sample will represent the underlying charts.

2.2 Properties of the Sample

Overall, from a sample of 2000 slots, we were able to acquire audio for 1365 slots (68 percent): 424 of 683 from before the 1970s, 505 of 664 from the 1970s, and 436 of 653 from after the 1970s. Because the sample was taken over slots and not individual singles, some singles, especially popular singles, appear more than once (and would need to be weighted accordingly for the most accurate statistics). Of the 1100 unique singles in our sample, performed by 533 unique artists, the great majority of singles (869) do appear only once, but 202 appear twice, 24 three times, and 5 four times. A more interesting artifact of sampling over slots instead of singles is that even though the original sample was drawn evenly across all chart ranks, there is still more weight in the sample toward the most popular songs. Songs tend to remain

```
# Love Will Keep Us Together
# Captain and Tenille
# 4/4
# key: B

| B | B | B | B |
| B | B | D#:hdim7/b5 | D#:hdim7/b5 | G#:7 | G#:7 |
| E | E | E:min | E:min |
| B | B:aug | B:maj6 | B:7 |
| E E/7 | C#:min7 F#:9(*3,11) . . |
| B | B | B | B |
| B | B | D#:hdim7/b5 | D#:hdim7/b5 | G#:7 | G#:7 |
```

Figure 3. Prototypical transcription illustrating features of the transcription format. The format encodes a number of high-level musicological features such as key, meter, beat, and phrase. Chord symbols follow the format proposed in [8] and include as much detail as possible about inversions and upper extensions.

on the charts for many weeks (10 on average, although this figure is much greater for the most popular songs and much less for the least popular), rising and falling through different ranks. Figure 2 illustrates the distribution of peak ranks in our sample, which corresponds well to that of the full set of chart slots during the time period spanned in the sample.

3. THE TRANSCRIPTION PROCESS

Annotating such a large data set was a considerably greater undertaking than we had expected, ultimately involving a team of more than two dozen people. We began by developing a file format for transcriptions that would capture as much musicologically-relevant information as possible, designed a web site to manage transcriptions, and organized a series of auditions to identify musicians with sufficient skill to transcribe reliably and efficiently at a high level of detail.

3.1 The Transcription Format

The transcription format was a plain-text format in order to facilitate transfer across platforms. The full specification is available for download with the transcriptions themselves, but the basic premises are illustrated in figure 3. All non-musical material is preceded by a comment character (#), and comments are allowed at the end of any line. The annotators used them freely. Each transcription begins with a four-line header containing the title of the song, the name of the artist, the meter, and the key, and new meter and key lines are added as necessary to reflect changes throughout the song. Each transcription is broken with line breaks into phrases, which are defined loosely as any point where a group might choose to start playing during a rehearsal. Pipes (|) denote barlines, and although transcribers were allowed to mark chords using whatever notation came most naturally to them, all have since been converted to the format proposed in [8].

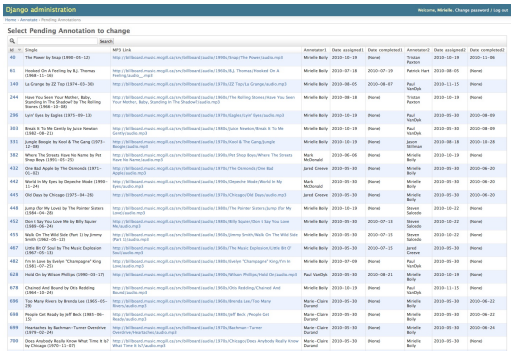

Figure 4. Screenshot of the web site that annotators used to manage their work. The page contains a list of all assignments as well as information about to whom each single was assigned and when.

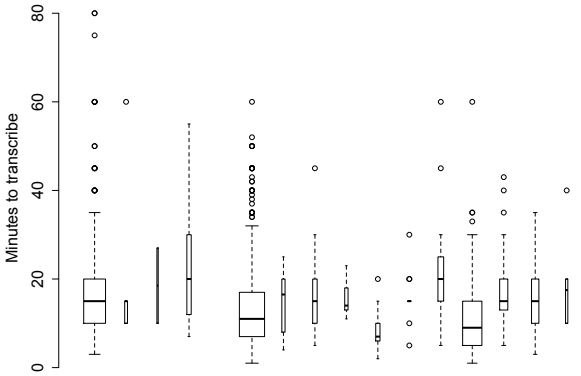

Figure 5. Transcribing times for each annotator. Box widths are scaled proportional to the square root of the number of transcriptions completed. Points more than 1½ times the inter-quartile range are plotted as outliers. The majority of songs took between 8 and 18 minutes to transcribe, although a few extremely difficult songs took more than an hour.

Chords are marked for every beat, with some shorthand to improve readability. For quadruple meters, which are the most common, a bar with a single chord symbol is assumed to have the same chord for all four beats. Bars with two symbols are assumed to have the chord change on beat 3. For bars with less than four chords that follow other patterns, periods are used to denote chords that have not changed. For example, in the first bar of the fifth line of the transcription in figure 3 contains E on the first two beat and E/D♯ on the second two beats, whereas the second bar contains C♯min7 on the first beat only followed by what might be noted as F♯11 in a fake book on the last three beats. Chord changes that are faster than the beat level are simplified. Notable silences in the music are marked with the special tag `&pause`.

3.2 Auditions and the Transcription Process

Over several recruitment periods between April and December 2010, 30 musicians were invited to audition for the project. With one exception (an undergraduate), these musicians were either graduate students in music performance or professional jazz performers (often both). Of those invited to audition, 23 completed the audition and 17 were ultimately hired. We prepared a detailed description of the file format for those auditioning, as well as a set of six sample songs with full transcriptions, in order to help the potential transcribers understand the format and the level of detail expected. After studying these materials, all those auditioning transcribed a set of five test songs that were chosen to be representative of the more difficult songs one would encounter. We reviewed these test transcriptions, decided whether the annotator had sufficient potential to continue, and provided detailed feedback on the audition to each transcriber we hired in order to ensure as much consistency as possible across transcriptions.

After hiring, following the principle of double-keying to minimize mistakes, two annotators were assigned to each song. Working with a custom-designed web interface (see figure 4), the annotators were able to access the audio for their assignments and, although they were asked to work independently, to see who their partner annotator was in case of any difficult questions. Annotators worked at different speeds, and in order to reward more efficient annotators, we paid per song with a bonus system to compensate for songs that were unusually difficult to transcribe. The majority of songs were transcribed in 8 to 18 minutes (median 12 minutes), but the most difficult songs could take an hour or more (see figure 5). Most annotators also reported that regardless of the amount of time spent, it was difficult to do more than a dozen songs in a single day: due to the intense concentration necessary, it was simply too exhausting for them to work more.

After the two assigned annotators for any given song had completed their transcriptions, a third meta-annotator compared the two versions—inevitably, there were usually differences in notation or musical opinion in addition to actual errors—and combined them into a master transcription. This combined version was then time-aligned and annotated with structural information based on musical similarity, functional information (verse, chorus, etc.), and instrumentation [12]. Factoring in the salaries of all involved, it cost more than $20 per song to arrive at this final file, but we believe that the richness and accuracy of the data justify the cost.

4. THE DATA SET

There are 414 059 labeled beats in our corpus, spread over 638 distinct chords and 99 chord classes. Each song contains 11.8 unique chords on average, ranging from a minimum of 1 to a maximum of 84; songs from the late 1970s exhibit the most harmonic variety. Figures 6 and 7 present the relative

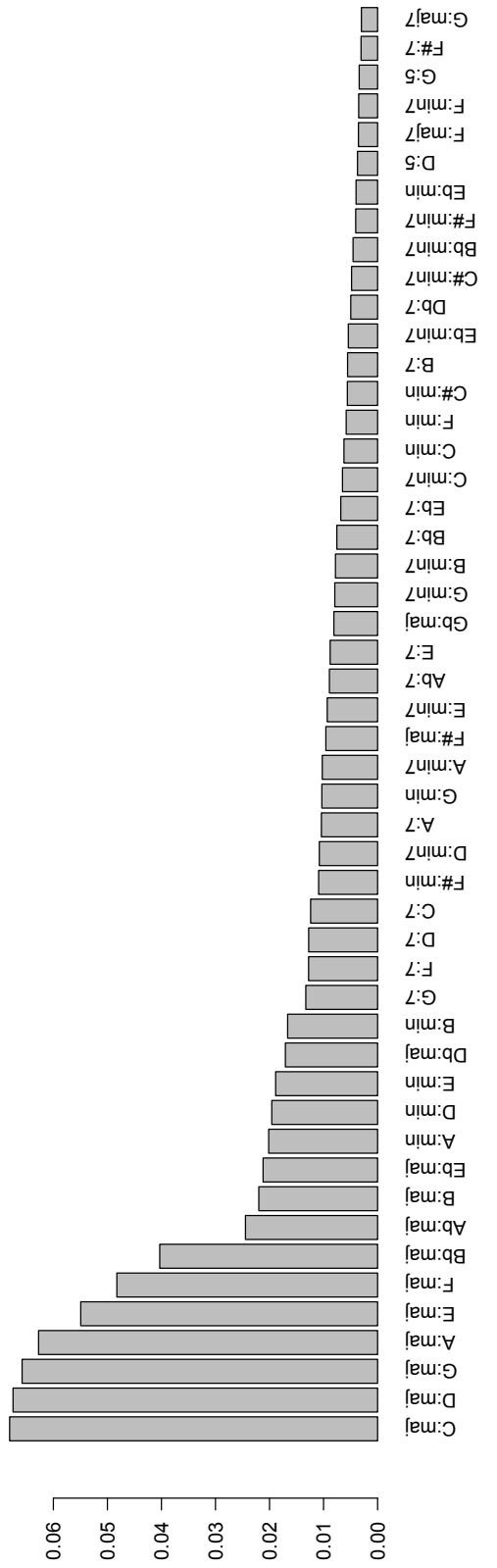

Figure 6. Frequency distribution of the 50 most common chords in the data set. There is a sharp drop after the most common major triads and a long tail afterward.

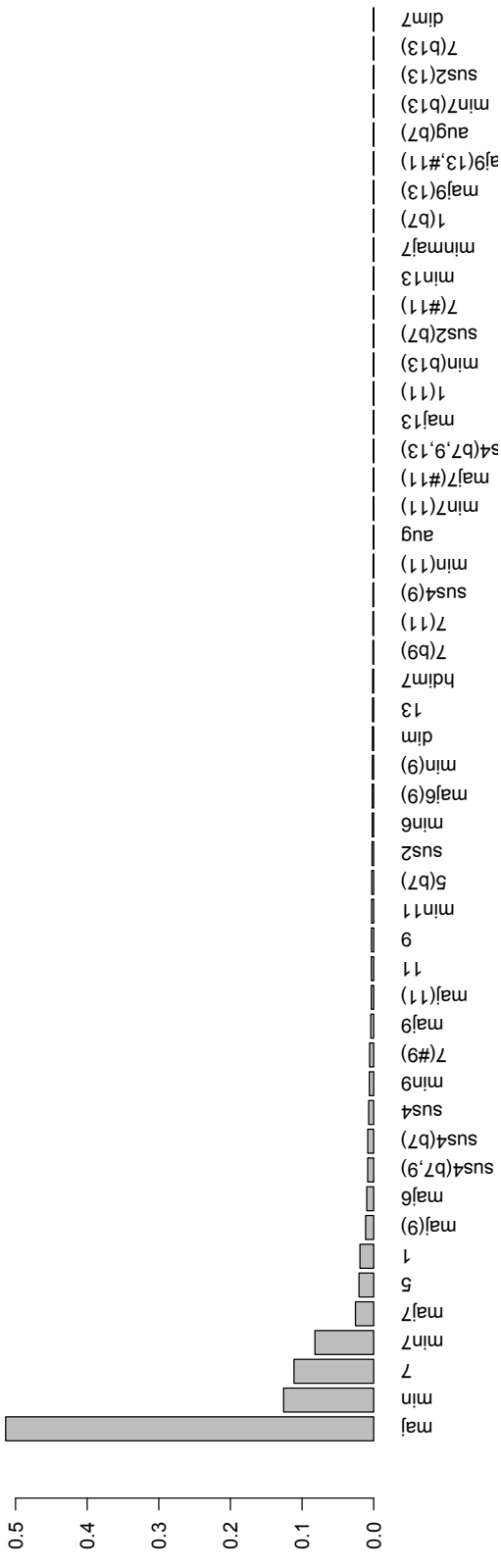

Figure 7. Frequency distribution of the 50 most common chord classes in the data set. Major chords alone account for more than half of the data set, followed by minor chords and the basic 7th chords.

frequencies of the top 50 chords and chord classes from the new data set. The most noticeable pattern is a sharp falloff after the seven most popular chords (all major): C, D, G, A, E, F, and B♭. Indeed, a milder falloff begins even after the four most popular chords. Certainly these chords are a useful set—they are sufficient to play in the five most common major keys—but such a sharp decline even for minor chords was unexpected. For chord classes, the falloff is even more extreme, although this is to be expected. The dominance of major and minor chords and simple seventh chords is consistent with most approaches to simplifying chords symbols (see [10], among others). The ordering suggests that with a data set of this size, it might be reasonable to start training systems that can also recognize simple 9th and 11th chords.

To our knowledge, there is no other curated corpus of popular harmony that equals this new data set in terms of size or scope. It is roughly five times the size of the existing MIREX set and contains a considerably broader range of artists, genres, and time periods. Trevor de Clercq and David Temperley have annotated another impressive data set of 200 songs from *Rolling Stone*'s "500 Greatest Songs of All Time," but their set is not time-aligned with audio [5]. We are currently working on a corpus analysis to compare our set to theirs and to explore deeper structures that may be discoverable with a larger data set.

5. SUMMARY AND CONCLUSION

Seeking to benefit both researchers interested in audio chord recognition and researchers interested in computational approaches to studying harmony in popular music, we have created a database more than four times the size of any existing database with detailed, curated musicological information and time-alignment with commercial audio recordings. The data set benefits from a special sampling methodology that was designed to maximize its utility both for musicological and for engineering purposes. Other researchers who wish to extend this data set or build a similar one of their own should be warned that the process is labor-intensive, but the statistics in this paper should provide guidelines for planning and budgeting. We are very excited to start working on the many questions this database will allow researchers to answer, and we are proud to make it available to the community at no cost and with minimally restrictive licensing. [4]

6. ACKNOWLEDGEMENTS

We would like to thank the Social Sciences and Humanities Research Council of Canada for funding this research, Rhonda Amsel for her advice on sampling, and all of the annotators who worked on the project, especially Reiko Yamada and Tristan Paxton for their tirelessness as meta-annotators.

[4] `http://billboard.music.mcgill.ca/`

7. REFERENCES

[1] S. Bhattacharjee, R. D. Gopal, J. R. Marsden, and R. Telang. A survival analysis of albums on ranking charts. In E. M. Noam and L. M. Pupillo, editors, *Peer-to-Peer Video: The Economics, Policy, and Culture of Today's New Mass Medium*, pages 181–204. Springer, New York, NY, 2008.

[2] N. Biamonte. Triadic modal and pentatonic patterns in rock music. *Music Theory Spectrum*, 32(2):95–110, 2010.

[3] Billboard Magazine. Hot 100 50th anniversary charts FAQ, 2008. Available `http://www.billboard.com/specials/hot100/charts/hot100faq.shtml`.

[4] E. T. Bradlow and P. S. Fader. A Bayesian lifetime model for the "Hot 100" Billboard songs. *Journal of the American Statistical Association*, 96(454):368–81, 2001.

[5] T. de Clercq and D. Temperley. A corpus analysis of rock harmony. *Popular Music*, 30(1):47–70, 2011.

[6] W. Everett. *The Foundations of Rock: From "Blue Suede Shoes" to "Suite: Judy Blue Eyes."* Oxford University Press, New York, NY, 2008.

[7] D. E. Giles. Survival of the hippest: Life at the top of the Hot 100. *Applied Economics*, 39(15):1877–87, 2007.

[8] C. Harte, M. Sandler, S. A. Abdallah, and E. Gómez. Symbolic representation of musical chords: A proposed syntax for text annotations. In *Proc. 6th ISMIR*, pages 66–71, London, England, 2005.

[9] K. Lee and M. Slaney. Acoustic chord transcription and key extraction from audio using key-dependent HMMs trained on synthesized audio. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):291–301, 2008.

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

[11] A. Sheh and D. P. W. Ellis. Chord segmentation and recognition using EM-trained hidden Markov models. In *Proc. 4th ISMIR*, pages 185–91, Baltimore, MD, 2003.

[12] J. B. L. Smith, J. A. Burgoyne, I. Fujinaga, D. De Roure, and J. S. Downie. Design and creation of a large-scale database of structural annotations. In *Proc. 12th ISMIR*, Miami, FL, 2011.