Technical Report: Harmonic Subjectivity in Popular Music

Hendrik Vincent Koops

W. Bas de Haas

John Ashley Burgoyne

Jeroen Bransen

Anja Volk

Technical Report UU-CS-2017-018
November 2017

Department of Information and Computing Sciences
Utrecht University, Utrecht, The Netherlands
www.cs.uu.nl

ISSN: 0924-3275

Department of Information and Computing Sciences
Utrecht University
P.O. Box 80.089
3508 TB Utrecht
The Netherlands

TECHNICAL REPORT

Technical Report: Harmonic Subjectivity in Popular Music

Hendrik Vincent Koops[*],W. Bas de Haas[†],John Ashley Burgoyne[‡],Jeroen Bransen[§],Anja Volk[¶]

Abstract

Reference annotation datasets containing harmony annotations are at the core of a wide range of studies in music information retrieval (MIR) and related fields. The majority of these datasets contain single reference annotations describing the harmony of each piece or song. Nevertheless, music theoretical insights on harmonic ambiguity and studies showing differences among annotators in many other MIR tasks make the notion of a single "ground-truth" reference annotation a tenuous one. In order to gain a better understanding of differences between annotators, we introduce and analyze the Harmonic Annotator Subjectivity Dataset (HASD) containing chord labels for fifty songs from four annotators. Our analysis of the chord labels in the dataset reveals a low overlap between the annotators. We show that annotators use distinct chord-label vocabularies, with less than 20 percent chord-label overlap across all annotators. A factor analysis reveals the relative importance of triads, sevenths, inversions, and other musical factors for each annotator on their choice of chord labels and reported difficulty of the songs in the dataset. Between annotators, we find only 73 percent overlap on average for the traditional major–minor vocabulary and 54 percent overlap for the most complex chord labels. Our results suggest the existence of a harmonic *"subjectivity ceiling"*: an upper bound for evaluations in computational harmony research. State-of-the-art chord-estimation systems in MIREX 2017 reported overlap scores that lie beyond this subjectivity ceiling by about 10 percent. This suggests that current ACE algorithms are powerful enough to tune themselves to particular annotators' idiosyncrasies. Overall, our results show that annotator subjectivity is an important factor in harmonic transcriptions that should inform future research on any musical tasks that rely on human annotations.

Keywords: Annotator Subjectivity, Harmony.

1. Introduction

Since the inception of computational harmonic analysis in music information retrieval (MIR) research, several reference annotation datasets for chord labels have been introduced (Mauch et al., 2009; Burgoyne et al., 2011; De Clercq and Temperley, 2011; Ni et al., 2013). These datasets are at the center of a wide range of important computational studies into harmony, including but not limited to: automatic chord estimation (ACE) (McVicar et al., 2014), analysis of harmonic trends over time (Mauch et al., 2015; Burgoyne et al., 2013; Gauvin, 2015), computational hook discovery (Van Balen et al., 2015), chorus analysis of popular music (Van Balen et al., 2013), data fusion of ACE algorithms (Koops et al., 2016), automatic structural segmentation (de Haas et al., 2013), and computational creativity, such as automatic generation of harmony accompaniment (Chuan and Chew, 2007) and harmonic blending (Kaliakatsos-Papakostas et al., 2014).

Virtually all of these studies use datasets that contain *single reference annotations*, i.e., for each corresponding musical moment (e.g., audio frame or section), the reference annotation contains a *single* harmony descriptor (e.g., a chord label) from either a single expert (Mauch et al., 2009) or a unified consensus of multiple experts (Burgoyne et al., 2011). Although

[*] Department of Information and Computing Sciences, Utrecht University, the Netherlands

[†] Chordify, Utrecht, the Netherlands

[‡] Music Cognition Group, Institute for Logic, Language and Computation, University of Amsterdam, Amsterdam, Netherlands

[§] Chordify, Utrecht, the Netherlands

[¶] Department of Information and Computing Sciences, Utrecht University, the Netherlands

most creators of these datasets warn about (harmonic) subjectivity and ambiguity, their annotations are nevertheless used in practice as the *de facto* ground truth for a large number of studies into harmony and related tasks (e.g., MIREX ACE). Moreover, using a single reference annotation is not exclusive to harmony research: a wide range of MIR studies and tasks, such as melody transcription, beat detection and automatic rhythm transcription, also rely primarily or exclusively on single reference annotations.

Theoretical insights on harmonic ambiguity from harmony theory (Schoenberg, 1978; Meyer, 1957; Harte et al., 2005), experimental studies on the large degree of annotator subjectivity (Ni et al., 2013), and the availability of vast amounts of heterogeneous (subjective) harmony annotations in crowd-sourced repositories (e.g., Ultimate-Guitar[6], Chordify[7]) make the notion of a single harmonic "ground-truth" reference annotation a tenuous one.

In an experimental study, Ni et al. found that annotators transcribing the same music recordings disagree on roughly 10 percent of harmonic annotations (Ni et al., 2013). Furthermore, they found that state-of-the-art ACE systems trained on single reference annotations perform worse on a consensus of annotators than on the single reference annotations. They suggest that current ACE systems are starting to overfit single reference annotations, thereby producing models that fail to represent the variability found in human annotations accurately. A similar lack of inter-rater agreement was found in an analysis of human annotations in the MIREX audio similarity task (Flexer, 2014).

The seemingly large differences in chord-label transcriptions among annotators raise questions about the validity of one-size-fits-all automatic chord-label estimation systems and their training and evaluation on single reference annotations. Furthermore, the overfitting problem described by Ni et al. points towards the need for more flexible ACE systems that can adapt themselves to the context (musical proficiency, chord-label vocabulary, etc.) of a user. In a study by Koops et al. (2017), a first approach to such a flexible system is proposed. By taking into account annotator subjectivity in an ACE system, it is shown that a shared harmonic representation can be learned directly from audio which takes into account multiple (heterogeneous) reference annotations. From this representation, chord labels can be personalized for each annotator, yielding more satisfactory chord labels than those generated by the same system trained on a single reference annotation.

Unfortunately, current datasets with harmony annotations contain either single reference annotations (Burgoyne et al., 2011; Mauch et al., 2009), or are restricted in size and sampling (Ni et al., 2013; De Clercq and Temperley, 2011). As a solution to this problem, we introduce a new chord-label dataset containing *multiple reference annotations* for fifty songs from the *Billboard* dataset.[8] Specifically, the new dataset includes four different annotators' transcriptions of each song.

The contribution of this paper is twofold. First, we introduce the Harmonic Annotator Subjectivity Dataset. This open chord-label dataset is linked with other important datasets containing harmonic transcriptions, as well as with major audio music repositories. Secondly, we show that within this dataset, significant differences exist between annotators, in chord labels as well as in perceived difficulty and annotation times. These results show that annotator subjectivity is an important factor in harmonic transcriptions, which should be taken into account in future automatic chord estimation, as well as related computational harmonic research.

The remainder of this paper is structured as follows. Section 2 discusses related work into the analyses of human judgments in music research. In Section 3, we describe the process of selecting songs, annotators and their transcription process. In Section 4, we provide an analysis of the transcriptions obtained from the annotators. The paper closes with a discussion and conclusion in Section 7.

2. Related Work in Analysis of Human Judgments in Music Information Retrieval

Disagreement between human annotators is a well-known problem in a wide variety of tasks in music information retrieval research. The lack of an exact task specification, the differences in the annotators' experiences, musical background, skill level, and instrumental preference, or the usage of different annotation tools are some of the possible causes of disagreement between annotators (Balke et al., 2016; Salamon et al., 2014; Salamon and Urbano, 2012). Annotator disagreement has previously been studied in the contexts of genre classification (Lippens et al., 2004; Seyerlehner et al., 2010), audio music similarity (Flexer, 2014; Flexer and Grill, 2016; Jones et al., 2007), music structure analysis (Nieto et al., 2014; Paulus and Klapuri, 2009; Smith et al., 2011), melody extraction (Balke et al., 2016; Bosch and Gómez, 2014), and human harmony annotations (Ni et al., 2013). Nevertheless, the extent of human disagreement and their impact on these tasks is commonly not taken into account when creating new music information retrieval methods.

The extent to which human judgments coincide is often referred to as *inter-annotator agreement* (or *inter-rater reliability*, *concordance*). The goal of studying inter-annotator agreement is to measure the amount

[6] https://www.ultimate-guitar.com/

[7] https://www.chordify.net/

[8] http://ddmal.music.mcgill.ca/research/billboard

of homogeneity or consensus between different annotators (or *raters*). With high inter-annotator agreement, raters can be used interchangeably without having to worry about the categorization being affected by a significant rater factor. In other words, if interchangeability of raters is guaranteed, then their ratings (or labels) can be used with confidence without asking which rater produced them. Conversely, if the ratings are effected by the raters and interchangeability is not guaranteed, the raters should probably be taken into account when interpreting the ratings (Gwet, 2014).

The joint-probability of agreement is the simplest and least robust measure for studying inter-annotator agreement. Several formal methods have been introduced that improve simple calculations of joint-probability. For example, Kappa (κ) statistics such as Cohen's κ (for two raters) (Cohen, 1960) and Fleiss's κ (for any number of raters) (Fleiss, 1975) correct for the amount of agreement that could be expected through chance. Cohen's κ was for example used in a study into the mood recognition of Chinese pop music (Hu and Yang, 2017). Jones at al. used Fleiss's κ to analyze human similarity judgments of symbolic melodic similarity and audio music similarity (Jones et al., 2007). Balke et al. adapted Fleiss' Kappa for evaluating multiple predominant melody annotations in jazz recordings (Balke et al., 2016). A more versatile statistic, Krippendorff's α (Krippendorff, 1970) assesses the agreement achieved among observers who rate a given set of objects in terms of the values of a variable. Krippendorff's α accepts any number of observers, and can be applied to nominal, ordinal, interval, and ratio levels of measurement. Furthermore, it is able to handle missing data, and corrects for small sample sizes. Schedl et al. (2016) used Krippendorff's α to investigate the agreement of listeners on perceptual music aspects (related to emotion, tempo, complexity, and instrumentation) of classical music.

3. Harmonic Annotator Subjectivity Dataset

We introduce the *Harmonic Annotator Subjectivity Dataset* (HASD), with chord labels for 50 songs from 4 annotators.

3.1 Song Selection

Currently available chord-label annotation datasets containing more than one reference annotation are limited by size, sampling strategy, or lack of a standardized encoding (Ni et al., 2013; De Clercq and Temperley, 2011). To account for these potential problems in our own dataset, we chose to select fifty songs from the *Billboard* dataset (Burgoyne et al., 2011) that have a stable online presence in widely accessible music repositories. This way, listening to the songs is easy, stimulating future research with the dataset. After searching the YouTube website for the title and artist tags of the *Billboard* dataset, we ranked the results of each query by number of views and selected the top fifty songs by this ranking. At the time they were collected, the least-viewed song in the dataset had 67 thousand views and the most-viewed song over 13 million, and an average of 11.9 unique chords according to the *Billboard* dataset annotations.

3.2 Annotator Selection

To study annotator subjectivity and account for a potential instrument bias, we recruited four annotators: two guitarists and two pianists. All annotators had either studied composition or music performance at the undergraduate or graduate level. All annotators were also successful professional music performers, with between 15 and 20 years of experience on their primary instrument. Two of the annotators further identified themselves as composers. We reviewed the first ten transcriptions from each annotator to ensure the annotators had sufficient aptitude to continue; all four annotators completed the initial screening successfully and were hired to continue to annotate the remaining forty songs. The annotators were compensated financially for their annotations at a fixed rate per song.

3.3 Transcription Process

To ensure the annotators were all focused on the same task, we provided them with a guideline for the annotating process. We asked them to listen to the songs as if they wanted to play the song on their instrument in a band, and to transcribe the chords with this purpose in mind. They were instructed to assume that the band would have a rhythm section (drum and bass) and melody instrument (e.g., a singer). Therefore, their goal was to transcribe the complete harmony of the song in a way that, in their view, best matched their instrument.

We used a web interface to provide the annotators with a central, unified transcription method. This interface provided the annotators with a grid of beat-aligned elements, which we manually verified for correctness. Chord labels could be chosen for each beat. The standard YouTube web player was used to provide the reference recording of the song. Through the interface, the annotators were free to select any chord of their choice for each beat. While transcribing, the annotators were able to watch and listen not only to the YouTube video of the song, but also a synthesized version of their chord transcription.

In addition to providing chords and information about their musical background, we asked the annotators to provide for each song a difficulty rating on a scale of 1 (easy) to 5 (hard), the amount of time it took them to annotate the song in minutes, and any remarks they might have on the transcription process.

3.4 Dataset Technical Specifications

To provide the MIR research community with a dataset that is easily accessible, expandable, encourages repro-

Annotator	Primary instrument	Average annotation time	Average reported difficulty	Average number of chord labels per song
$A1$	Guitar	23.10 ($\sigma = 14.91$)	2.40 ($\sigma = 1.16$)	9.46 ($\sigma = 5.13$)
$A2$	Piano	15.66 ($\sigma = 9.91$)	1.60 ($\sigma = 1.18$)	9.42 ($\sigma = 4.20$)
$A3$	Guitar	22.00 ($\sigma = 7.42$)	2.42 ($\sigma = 0.73$)	12.44 ($\sigma = 5.83$)
$A4$	Piano	26.10 ($\sigma = 12.18$)	1.96 ($\sigma = 1.07$)	8.86 ($\sigma = 4.70$)

Table 1: Overview of annotators, their primary instrument and average annotation time and chord labels per song statistics.

ducibility and stimulates future research into annotator subjectivity, we adopted a number of standard encodings that are commonly used in MIR research.

For each of the fifty songs, the dataset contains chord labels provided by four annotators. These chord labels are encoded using the chord-label syntax introduced by Harte et al. (2005). This syntax provides a simple and intuitive encoding that is highly structured and unambiguous to parse with computational means. In addition to chord labels, the dataset contains information about the four annotators, such as musical background, music education and their main instrument. To promote and stimulate future research, we include identifiers for music repositories (e.g., YouTube), allowing researchers to listen to the tracks easily. Furthermore, we provide *Billboard* dataset identifiers which make it possible to cross-reference our dataset with data from the *Billboard* dataset, ACE output from the MIREX task, and other datasets that use these identifiers.

The complete dataset is encoded using the JAMS format: a JSON-annotated music specification for reproducible MIR research, which was introduced by Humphrey et al. (2014). JAMS provides an interface with the standard MIREX evaluation measures used in this paper, making it very easy to evaluate and compare annotations. To provide easy access, we make the dataset publicly available in a Git repository[9]. By way of Git and JAMS, we encourage the MIR community to exchange, update, and expand the dataset.

4. Global View of Annotator Subjectivity

To obtain a general idea of the degree of annotator subjectivity in our dataset, we first analyze the annotations in terms of descriptive statistics. First, we analyze the difficulty scores and remarks (Section 4.1) and the overall chords the annotators provided (Section 4.2). Next, we provide an analysis of the differences in chord labels used by the annotators (Section 6). Building on these findings, we will investigate the cause of annotator subjectivity in more detail with more advanced statistical methods in the sections that follow.

4.1 Reported Annotation Time and Difficulty

Overall, the four annotators ($A1$, $A2$, $A3$, $A4$) took 22 min on average to transcribe a song ($\sigma = 12$), with a minimum of 5 min and a maximum of 60 min. Individually, the averages per annotator were 23 min ($\sigma = 15$), 16 min ($\sigma = 10$), 22 min ($\sigma = 7$), and 26 min ($\sigma = 12$) for $A1$, $A2$, $A3$, and $A4$, respectively.

The annotators also ranked their perceived difficulty of all songs on a scale from 1 (easy) to 5 (difficult). Individually, the annotators reported average difficulties of 2.4 ($\sigma = 1.2$), 1.7 ($\sigma = 1.1$), 2.6 ($\sigma = .8$), and 2.0 ($\sigma = 1.3$), for $A1$, $A2$, $A3$, and $A4$, respectively. Both the average annotation times and reported difficulty for all annotators can be found in Table 1.

Intuitively, the more difficult a song is, the longer it should take to annotate. We can test this relationship using Pearson's correlation coefficient (r). Between the average reported difficulties and average annotation times, we find a very strong positive linear correlation, $r = .93$, $p \ll .05$. The correlations per annotator appear in Figure 1. The figure shows that for $A1$ and $A2$, the correlation is very strong, $r = .92$ and $r = .84$, respectively. $A4$'s measurements are also strongly correlated ($r = .76$); $A3$ shows a strong correlation that is nonetheless perhaps weaker than the rest ($r = .61$). Figure 1 shows that $A3$'s annotations cluster around 20–30 min in length and a reported difficulty of 2–3, while the other annotators exhibit a wider spread across both time and difficulty. The outlier in Figure 1, with a reported difficulty of 1 and a reported annotation time of 60 minutes, can be explained by it being the first song annotated by $A4$, who had to get used to the interface and annotation process. However, in Section 5 we will see that the order of songs does not have a significant effect on annotation time and perceived difficulty for any annotator.

4.2 Chord-Label Statistics

Turning to the harmonic transcriptions, we investigate the extent to which annotator subjectivity in terms of chord labels can be found in our dataset. We analyze the chord-label annotations in several ways. First, we investigate which chord labels are used in our dataset and how much overlap in chord vocabulary there is among annotators. This will provide a general indication of annotator subjectivity in our dataset, as it shows the difference in chord-label vocabularies among annotators. Then we analyze the number of unique chord labels in a song and its reported difficulty.

[9] RepositoryURL removed for double blind reviews

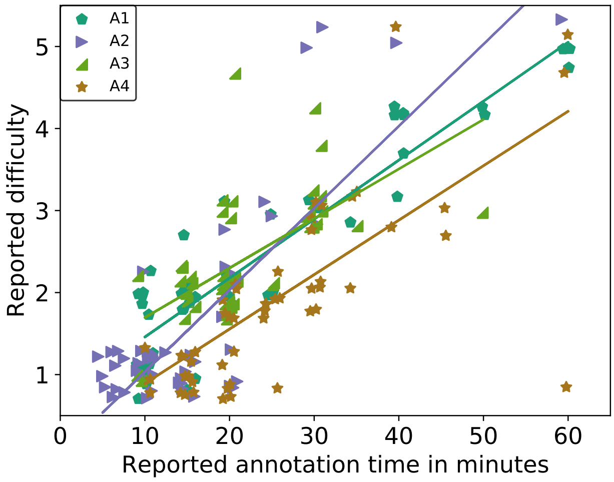

Figure 1: We find strong, but differing, correlations per annotator between reported annotation time and reported difficulty from 1 (easy) to 5 (hard). In general, songs perceived as difficult took longer to annotate than easy songs. Random jitter added to aid visualization.

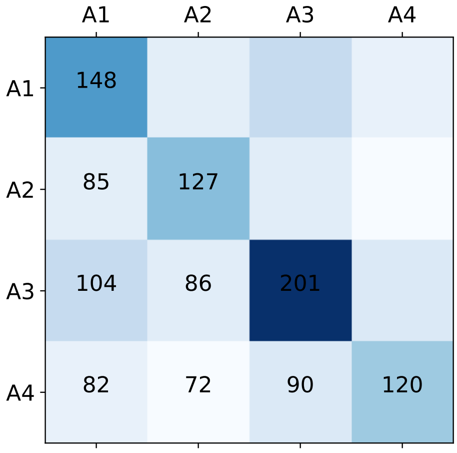

Figure 2: Pairwise intersection sizes of all 290 unique chord-labels in the dataset for all annotators. On average, the annotators share less than half of their chord label vocabulary with the other annotators.

4.2.1 *Chord-label vocabularies*

On average, the four annotators ($A1$, $A2$, $A3$, $A4$) used 10.3 chord labels per song ($\sigma = 5.2$), with a minimum of 3 chord labels and a maximum of 27 chord labels. Individually, the averages per annotator were 9.46 chord labels ($\sigma = 5.13$), 9.42 chord labels ($\sigma = 4.2$), 12.44 chord labels ($\sigma = 5.83$), and 8.86 chord labels ($\sigma = 4.7$) for $A1$, $A2$, $A3$, and $A4$, respectively. These statistics are similar to what was found in the *Billboard* dataset by Burgoyne et al. (2011), in which songs contain on average 11.8 unique chord labels.

Altogether, the annotators used 290 unique chord labels in their transcriptions, of which the most frequently used chords are common chord labels such as `G:maj`, `C:maj`, `D:maj`, and `A:maj`. Annotators $A1$, $A2$, $A3$, and $A4$ used 148, 127, 201, and 120 unique chord labels respectively. The intersection of the unique chords of all annotators contains only 56 chord labels, corresponding to less than 20 percent of all chord labels in the dataset, which already provides some evidence that each annotator uses a distinct set of chord labels. The intersection set contains only two enharmonically equivalent chords, and only three inverted chords: `F:maj/3`, `E:maj/2`, `D:maj/5`. Nevertheless, inversions are generally used by all annotators. Around 11 percent of the chord labels in the dataset contain inversion. Nevertheless, the annotators differ in the amount of chord labels that include inversions. Of all the chord labels that the annotators $A1$, $A2$, $A3$, and $A4$ use, 0.08, 0.04, 0.15 and 0.16 percent include inversions, respectively. Of their unique chord labels, 0.26, 0.27, 0.43, 0.39 percent include inversions for $A1$, $A2$, $A3$, and $A4$ respectively. This seems to suggest that while there is relatively little disagreement about pitch spelling, there is a large amount of disagreement on the level of inversions. If we consider a chord label equivalent to all its possible inversions, we find a total of 139 unique chord labels, and an intersection size of only 38 chord labels, corresponding with 27 percent of all chord labels in the dataset.

The intersection sizes for unique chord labels for all songs for each pair of annotators can be found in Figure 2. This figure shows that $A1$ and $A3$ share the most chord labels (104). Fewer chord labels are shared between $A2$ and $A4$ than with the rest. This is interesting, as $A1$ and $A3$ are both guitar players, and $A2$ and $A4$ are piano players. This seems to suggest that our piano players are on average more diverse in terms of their chord-label vocabulary, while the guitar players seem to be more similar to each other in their chord-label vocabulary – although the usual caveats with respect to small sample size apply.

4.2.2 *Difficulty versus number of chord labels in a song*

It can be expected that songs with a large number of chord labels, and therefore a large number of chord changes should be harder to transcribe than songs with a small number of chord labels. We indeed find a pos-

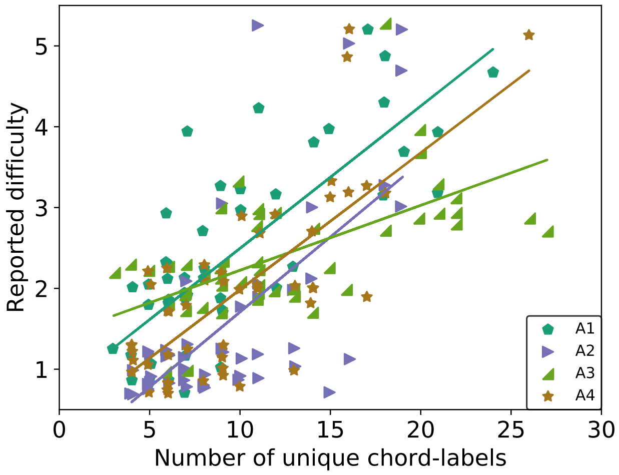

Figure 3: Reported difficulty and number of chord labels per song are strongly correlated for all annotators. The larger the number of chords used, the more difficult to annotate was the song perceived.

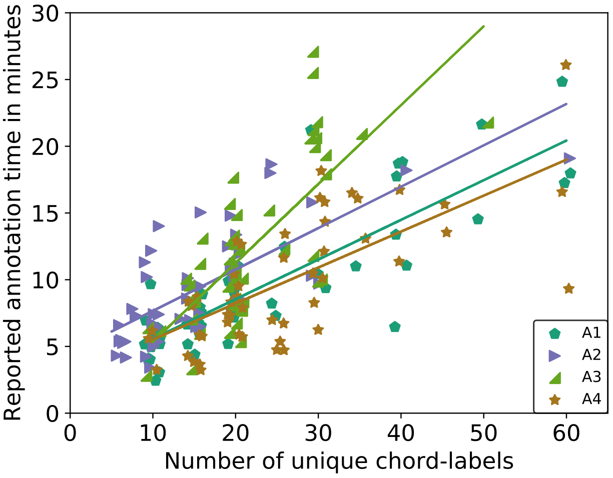

Figure 4: Annotation time and number of chord labels per song are strongly correlated for all annotators. The larger the number of chords used, the more time it took to annotate.

itive correlation between the reported difficulty of a song and the number of unique chord labels for that song. In Figure 3, the number of unique chords used by an annotator for a song is plotted against that annotators' reported difficulty for that song. Furthermore, in Figure 4 the number of unique chords used by an annotator for a song is plotted against that annotators' reported annotation time for that song.

We find a strong positive correlation between the average reported difficulty and average number of unique chords, $r = .80$, $p \ll .01$. Nevertheless, when we turn to individual annotators, we see that not all correlations are similar for all annotators. For $A1$ ($r = .79$) and $A4$ ($r = .75$) the degree of correlation is comparable, but the correlations for $A2$ ($r = .67$) and $A3$ ($r = .65$) are strong but somewhat weaker.

In an inspection of Figure 3, we see that some songs are annotated with a low number of unique chords, but with a relatively high difficulty. When we look at those transcriptions, we find indeed a low number of unique chord labels, but with a high amount of detail. These chord labels are often intricate labels with added sevenths, ninths, or thirteenths, or inversions (e.g., `C#:min7/b7` or `Bb:min9/b3`), which are harder to play and transcribe. These differences among annotators help us understand the subjectivity of perceived difficulty: for some annotators difficulty is about the amount of (change in) chord labels per song, while others report songs to be more difficult if the chord labels themselves are more complex.

5. Individual Differences in Annotation Ability

The previous section highlights several areas of variance among the annotators: annotation time, chord vocabulary, and how difficulty is perceived. In order to formalize the potential causes of this variance, we examined the correlation of these annotator behavior measures – reported annotation time, reported annotation difficulty, and number of unique chords used – with the annotators' agreement with the *Billboard* ground truth. We also considered two potential external causes of difficulty or disagreement, the length of the song (in seconds) and a learning effect after completing several annotations, represented by the tranche in which annotators received each song (first, second, or third). We were particularly interested the following. First, in checking whether there is indeed a general chord complexity factor that goes beyond triads and inversions. Secondly, whether song length or learning affects reported difficulty or annotation disagreement. Thirdly, whether there is a consistent relationship among the behaviour and agreement measures independent of individual annotators. And finally, whether there are differences between annotators with respect to agreement in addition to the differences in the behavioral measures (...). These questions focused on differences among annotators as independent individuals with reference to a global ground truth, without (yet) considering the annotators' agreement with each other.

We measured agreement with the original *Billboard* ground truth using the MIREX weighted chord symbol recall (WCSR) metrics, i.e., the proportion of correct labels weighted by song duration, after both the labels and the ground truth have been simplified to one of seven following vocabularies: ROOT only compares the root of the chords; MAJMIN only compares major, minor, and no-chord labels; MIREX considers a chord label correct if it shares at least three pitch classes with the reference label; THIRDS compares chords at the level of root and major or minor third; TRIADS compares at the level of triads (major, minor, augmented, etc.), i.e., in addition to the root, the quality is considered through a possibly altered 5th; SEVENTHS compares all above plus any notated sevenths; TETRADS compares at the level of the entire quality in *closed voicing*, i.e., wrapped within a single octave. Extended chords (9ths, 11ths and 13ths) are rolled into a single octave with any upper voices included as extensions. For MAJMIN, THIRDS, TRIADS, TETRADS and SEVENTHS, we also test with inversions: MAJMIN_INV, THIRDS_INV, etc. For a detailed explanation of these measures, we refer the reader to the standardized MIR evaluation software package `mir_eval` by Raffel et al. (2014) and the MIREX ACE website[10].

Before computing correlation coefficients, we transformed each of our measures to improve normality. (Using Spearman's correlation coefficients instead of Pearson's to avoid normalization transforms was not possible because some of our research hypotheses involve differences in means.) For annotation time and the number of unique chords per annotator, as well as song length, we used a log transform (base 2). For the MIREX WCSR measures, which range from 0 to 1, we used a probit (standard normal quantile) transform. We also reversed the sign of the transformed WCSR measures so that they would represent difficulty/disagreement rather than easiness/agreement.

We treated reported annotation difficulty as an ordinal variable, using polyserial correlation coefficients instead of Pearson's. Polyserial correlation coefficients assume that an ordinal variable with k levels is a coarse observation of a latent normal variable, with $k-1$ cut points determining which ordinal level is observed. For example, for a binary variable there is one cutpoint, it assumes that all values of the latent variable below the cut point are observed as 0 and all values above the cut point are observed as 1. When using polyserial correlation coefficients in a statistical model, one usually estimates the cut points as extra parameters, sometimes independently for each participant or group. This estimation is not computationally trivial, and it is sensitive to empty rating categories; common estimation procedures can also yield mildly non-positive-definite correlation matrices. We collapsed rating difficulties 4 and 5 into a single category to avoid some of these problems, but Annotator 2 rated such a large majority of songs as having difficulty 1 that violations of positive definiteness were impossible to avoid entirely.

5.1 Exploratory Factor Analysis

We began with an exploratory factor analysis to determine the dimensionality of our set of measures. Both parallel analysis (Humphreys and Jr., 1975) and Velicer's MAP criterion (Velicer, 1976), two common techniques for choosing the dimensionality, suggest that four factors are sufficient. Table 2 presents the four-factor solution, using the principal-factor method (similar to principal-component analysis but allowing for an additional error sources for each measure) with an oblique rotation (oblimin) to maximize interpretability. The pattern in the loadings (correlations between the factors and the original measures) lends itself to a clear and meaningful interpretation of the factors. Factor 1 represents a baseline, triad-level difficulty, Factor 2 represents additional difficulty arising from sevenths, and Factor 4 represents additional difficulty arising from inversions. Factor 3 collects all three of the annotator-dependent difficulty measures, suggesting that there is indeed a distinct complexity aspect to some songs that goes beyond triads, sevenths, and inversions. Because we used an oblique rotation rather than an orthogonal one, correlations among the factors were possible, and all four of the factors are inter-correlated positively, suggesting that a higher-level, general difficulty factor may be present that is partially responsible for all four lower-level types of difficulty. The communalities (h^2, or proportion of variance explained for each measure) are very high for the MIREX vocabularies, showing that the four-factor model does an excellent job explaining these measures. The annotator-dependent indicators have lower communalities, especially the number of unique chords, but still represent a good fit. Overall, the four-factor exploratory model explains 92 percent of the variance in the data we collected.

In summary, the exploratory factor analysis suggested that annotator's performance depends on a baseline triad-level difficulty, additional difficulty arising from sevenths or inversions, and a further chord complexity factor; it also suggests that there may be a general difficulty factor contributing to each of the four difficulty types. As a final check on the four-factor model, we compared three- and five-factor models as alternatives. Neither alternative was compelling. A three-factor model simply eliminates Factor 4 (inversions), which has considerable explanatory value; the extra factor in a five-factor model, in contrast, has no obvious interpretation and no items with loadings of greater magnitude than the four-factor model.

5.2 Confirmatory Factor Analysis

The exploratory factor analysis suggested a basic underlying model for how annotators' perceived difficulty in transcribing a song relates to their agreement with the ground truth for that song. The factors in this model are inter-correlated, suggesting that there may also be a higher-order common cause of difficulty. Exploratory factor analysis is limited, however, in its ability to specify the factor structure further, and it also offers no good way to test for the effect of external factors, such as song length and learning effects. It also makes it difficult to separate which aspects of the model are common to all annotators from those aspects that differ among annotators, i.e., potential as-

[10] `http://www.music-ir.org/mirex/wiki/2017:Audio_Chord_Estimation`

Indicator	Factor 1	Factor 2	Factor 3	Factor 4	h^2
Loadings					
MIREX vocabulary					
THIRDS	**.96**	.02	.01	−.01	.96
MAJMIN	**.95**	.05	−.03	.03	.97
TRIADS	**.92**	.02	.09	.01	.96
ROOT	**.92**	.03	.01	.01	.91
MIREX	**.94**	−.02	.02	.00	.88
THIRDS_INV	.46	.15	.11	**.55**	.97
MAJMIN_INV	.47	.15	.05	**.58**	.99
TRIADS_INV	.48	.12	.14	**.53**	.98
SEVENTHS	.18	**.92**	−.04	.22	.98
TETRADS	.19	**.89**	.05	−.24	.98
SEVENTHS_INV	−.10	**.97**	.00	.23	.99
TETRADS_INV	−.08	**.94**	.08	.20	.98
Difficulty rating	−.04	.00	**.94**	−.06	.83
Annotation time	.07	−.03	**.88**	.00	.83
Number of unique chords	−.07	.02	**.80**	.01	.60
Inter-Correlations (Proportion Variance Explained on Diagonal)					
Factor 1	**.39**				
Factor 2	.67	**.26**			
Factor 3	.49	.36	**.17**		
Factor 4	.39	.29	.24	**.10**	

Note. $N = 200$. The largest factor loading for each indicator appears in boldface. Factor 1 seems to represent a baseline, triad-level difficulty, Factor 2 additional difficulty arising from sevenths, Factor 4 additional difficulty arising from inversions, and Factor 3 a chord-complexity factor beyond these components that also contributes to annotators' perceived difficulty. h^2 = communality, the percent of variance per indicator explained by the factor model.

Output of the R **psych** package, version 1.7.8, using the principal-factor method (Revelle, 2017).

Table 2: Exploratory Factor Analysis of Annotation Difficulty Indicators (Oblimin Rotation)

pects where annotator subjectivity is at work. We thus used the four-factor model as a basis for a confirmatory factor analysis, where we could verify the plausibility of the exploratory model and test for the presence of the general difficulty factor, the effects of song length and learning, and whether annotators differ significantly on each of the factors – or in other words, what exactly causes annotators' transcriptions to vary.

Our first step in the confirmatory analysis was to define the factors more rigorously. Given the loading patterns and high inter-correlations in the exploratory model, we allowed the Triad Difficulty factor to load on all twelve of the MIREX WCSR measures, and thus serving as a baseline for all measures of this type. All other loadings for this factor were constrained to zero. We allowed the Sevenths Difficulty factor to load only the four MIREX vocabularies involving sevenths and the Inversions Difficulty factor to load only on the five vocabularies involving inversions, again constraining all other possible loadings on these factors to zero. We allowed the Annotation Difficulty factor to load only on the three annotator-dependent measures, reported difficulty, reported annotation time, and number of unique chords. To ensure that the model remained identified given the overlapping factors, we enforced independence (zero covariance) between Triad Difficulty and Sevenths Difficulty and also between Triad Difficulty and Inversions Difficulty, but we allowed all other possible pairs of factor to covary.

We fit this first-order model to each annotator individually. Table 3 includes goodness-of-fit statistics for these models. The model fits well for Annotators 3 and 4, adequately for Annotator 1, and less well for Annotator 2. Annotator 2 exhibited so little variance in difficulty ratings that the polyserial correlations lead to a non-positive-definite matrix. So many of the ratings are 1 that it is impossible to estimate an underlying normal variable reliably. Once we combined Annotator 2 back with other annotators in later models, however, the problem subsided somewhat, and despite the overall instability of the fit for Annotator 2, all loadings in this first-order model are large, statistically significant ($p < .05$), and of comparable magnitude for every individual annotator. We accepted the first-order model, and for further analysis, we assumed that all annotators shared a common model form.

In both the exploratory factor analysis and the first-order model, the four factors are highly intercorrelated, which suggested that there may be an underlying General Difficulty factor that is responsible for this correlation, i.e., a second-order model (see Figure 5). The second-order model had one fewer parameter per annotator – in place of the four free correlations between factors in the first-order model there are four loadings from General Difficulty to each of the original four factors, and one of these must be fixed in order to identify the model. As such, second-order model should normally have a poorer fit than the first-order model, but if the difference is not statistically significant and the model still fits acceptably, we should prefer the more parsimonious second-order model. As Table 3 shows, the second-order model does indeed fit acceptably well and the degradation in fit from the first-order model is not statistically significant ($p = .90$). Looking in detail at the model parameters, however, we noticed that the loadings on Sevenths Difficulty was small and not statistically significant for any annotator. As such, we also tested an even more parsimonious model wherein the General Difficulty factor was not allowed to load on Sevenths Difficulty (i.e., we fixed the loading to zero). This second-order model without a connection between General Difficulty and Sevenths Difficulty also fit acceptably well and showed no significant degradation from the model where the loading between General Difficulty and Sevenths Difficulty was free ($p = .44$). We accepted the presence of a General Difficulty factor and used the model without a connection to Sevenths Difficulty as our basis for further testing.

Given the General Difficulty factor, we then examined whether song length or learning affected General Difficulty. Again, we used a backward step-wise selection process for consistency with the other selection procedures. We first tested a model with both of these covariates as exogenous predictors of General Difficulty and found that while song length had a significant effect for all annotators, tranche did not have a significant effect for any annotator. Removing tranche showed no significant degradation in model fit ($p = .38$), but removing song length degraded model fit substantially ($p = .01$). We chose the model with only song length as a predictor of General Difficulty. Figure 5 depicts this model structure.

In order to test whether the latent difficulty factors differed across annotators, we followed the procedure recommended by Brown (2015). We first tested *measurement invariance*: that the relationship between the latent factors in the model and the observed measures is the same for all annotators. In the absence of measurement invariance, comparing the latent factors would be meaningless. Starting with a baseline "equal form" model, namely the model with a General Difficulty factor and song length as an exogenous predictor, we first tested whether the loadings and intercepts in the model were equal for all annotators. As with adding the General Difficulty factor, this restriction should not improve model fit, but because it is more parsimonious, we accept it if the degradation in model fit is not significant. The model with equal loadings and intercepts still fits well, and the degradation with respect to the equal-form model is not significant ($p = .65$). Further restricting the coefficient of the song-length regression on General Difficulty retained a good fit, and the degradation in fit was again not significant ($p = .52$). These restrictions meet the criteria for "strong" measurement invariance, and as such, we

Model	χ^2	df	χ^2_{diff}	Δdf	RMSEA	CFit	SRMR	CFI	TLI
Single-Annotator Models (First-Order)									
Annotator 1 ($n = 50$)	107.68^{**}	77			.090	.07	.028	.92	.89
Annotator 2 ($n = 50$)	112.52^{**}	77			.097	.04	.033	.81	.73
Annotator 3 ($n = 50$)	89.91	77			.059	.38	.059	.94	.92
Annotator 4 ($n = 50$)	87.72	77			.053	.44	.053	.94	.91
Higher-Order Structure									
First-order	392.11^{***}	308			.075	.05	.032	.91	.88
Second-order									
w/ Sevenths Difficulty	395.70^{***}	312	0.12	1.7	.074	.06	.033	.91	.96
w/o Sevenths Difficulty	358.64^{**}	316	0.54	0.9	.052	.43	.035	.96	.94
Exogenous Predictors									
Song length and tranche[a]	446.14	424			.033	.82	.039	.98	.97
Song length	430.54	428	0.75	1.0	.011	.94	.039	1.00	1.00
None	603.30^{***}	436	8.31^{**}	1.3	.088	<.01	.038	.84	.80
Measurement Invariance									
Equal form[a]	372.57	368			.016	.91	.039	1.00	.99
Equal loadings and intercepts	498.18	467	4.32	6.2	.037	.78	.068	.97	.97
Equal predictor coefficients	480.86	470	0.41	1.0	.022	.92	.068	.99	.99
Annotator Heterogeneity									
Equal factor variance	532.66	485	$4.16^{\dagger}$	1.6	.045	.63	.167	.95	.96
Equal first-order factor means[b]	479.78	482	0.36	2.3	<.001	.97	.068	1.00	1.00
Equal second-order factor mean									
w/ free ann. time intercept	467.49	482	0.07	1.0	<.001	.99	.068	1.00	1.01
w/o free ann. time intercept	474.24	485	$2.63^{\dagger}$	0.9	<.001	.98	.068	1.00	1.01

Note. $N = 200$. χ^2_{diff} and Δdf represent nested differences, scaled using Satorra's method. Italics represent the model chosen from each set to be the baseline for the following set. RMSEA = root mean square error of approximation, ideally $\lessapprox .060$; CFit = probability that RMSEA $\leq .050$; SRMR = standardized root mean square residual, ideally $\lessapprox .080$; CFI = comparative fit index, ideally $\gtrapprox .95$; TLI = Tucker–Lewis index, ideally $\gtrapprox .95$. The model selected from each section of the table appears in italics.

[a] Statistics differ from the previous model because of the addition or deletion of potential exogenous indicators in the target correlation matrix.

[b] Factor variances remain free because there is no evidence of homogeneity; the baseline for comparison remains the equal-predictor model.

$^{\dagger}p < .10 \quad ^{*}p < .05 \quad ^{**}p < .01 \quad ^{***}p < .001$

Output of the R **lavaan** package, version 0.5.23.1097 (Rosseel, 2012).

Table 3: Test Statistics for Measurement Invariance and Annotator Heterogeneity on Annotation Difficulty Indicators

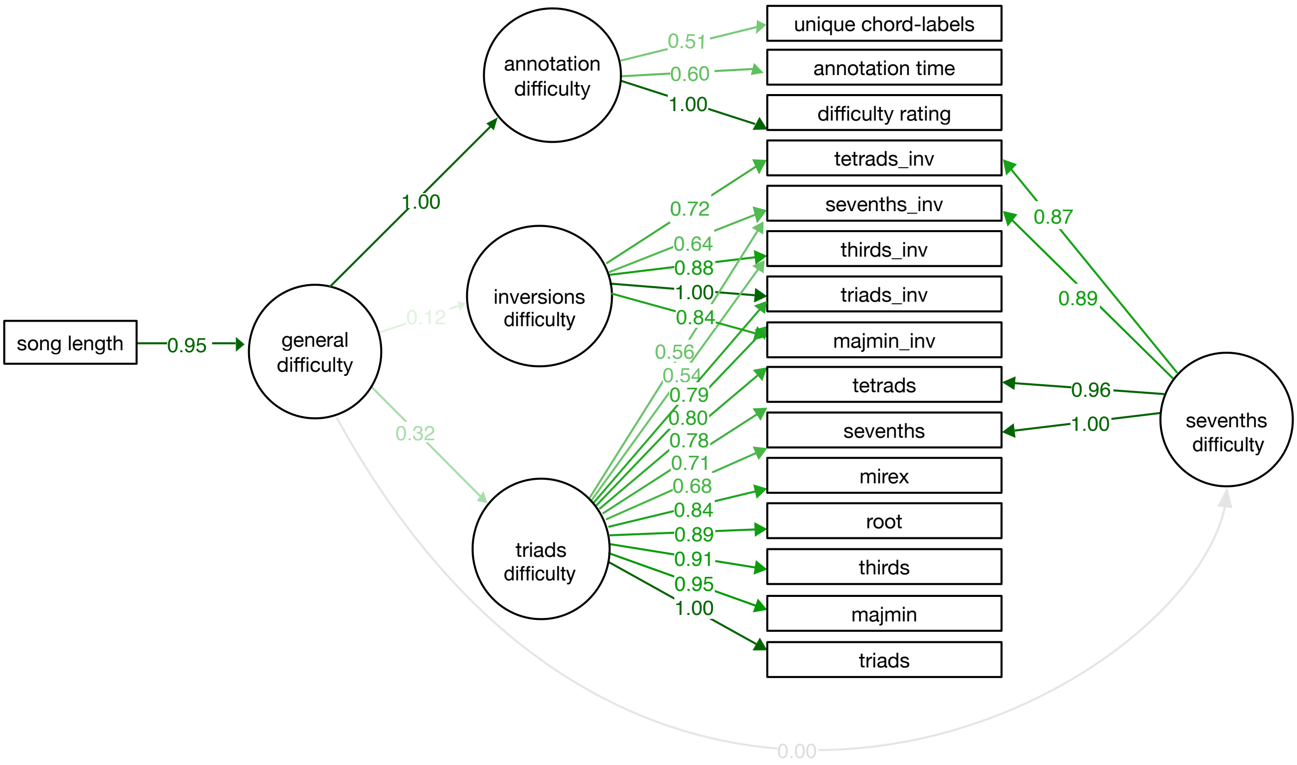

Figure 5: Second-order factor model for indicators of annotation difficulty. Loadings are unstandardized and common to all annotators. Intercepts (which were common across annotators) and residual variances (which were not) are omitted for clarity. A second-order General Difficulty factor predicts three of the four first-order factors. The largest loading on each factor is set to 1.0 in order to fix their scales.

proceeded to testing annotator differences on the latent difficulty factors. Figure 5 includes the common loadings and predictor coefficients for this strong invariance model.

We first tested for differences in factor variances across annotators. When restricting the variances of the factors to be equal across annotators, the degradation in model fit with respect to the strong invariance model is weakly significant ($p = .09$) and the many goodness-of-fit measures drop to borderline levels. The standardized root mean square residual (SRMR) is unacceptably high – .167 – and more than twice as bad as any other model we considered. We rejected the hypothesis of equal factor variance across annotators.

We also tested for difference in factor means across annotators. We began by restricting the factor means to be equal only for the first-order difficulty factors. In contrast to restricting the factor variances, restricting these factors means yields an acceptable model fit and no significant degradation ($p = .88$). Further restricting the second-order mean (General Difficulty) to be the same across annotators still yields an acceptable fit with no significant degradation ($p = .52$). We concluded that although factor variance differs among annotators, the factor means are the same.

At this point, we had a largely acceptable model. As a final step, we examined the *modification indices* for any problematic constraint. Modification indices are an approximation of how much model fit will improve if a single constraint is relaxed. The modification indices suggested that freeing the intercept for annotator time would improve model fit for most annotators, and this was plausible: even given a common level of Annotation Difficulty, it is believable that some annotators will be uniformly faster or slower. We compared a model with a free annotation-time intercept to our model with all intercepts restricted, and the degradation was weakly significant ($p = .09$). We concluded that that intercept for annotation time should remain free.

In summary, we found that a General Difficulty factor can explain both annotators' perceived difficulty and their agreement with the *Billboard* ground truth; more difficult songs exhibit less agreement, and our chosen annotator-dependent measures are consistent with the common external measures of WCSR. While we found no evidence of a learning effect from annotation experience, we found song length had a significant impact on General Difficulty, with longer songs being more difficult on average. Beyond General Difficulty, further differences in perceived difficulty or ground-truth agreement could be explained by four lower-level factors: Triad Difficulty, Sevenths Difficulty, Inversions Difficulty, and other Annotation Difficulty. On average, all annotators found the songs equally difficult with respect to these factors, but the variance differed. Finally, even after taking into account the difficulty factors, some annotators were systematically slower or faster than others.

How should one interpret differences in factor variances when the means are the same? Variance in this case reflects the range of difficulty across the full sample of songs we asked annotators to transcribe, and thus low variance suggests a lack of sensitivity to a particular type of difficulty, whereas high variance suggests that a particular type of difficulty is especially important for a particular annotator. Put differently, the results suggest that the core of annotator subjectivity lies not in differences in raw transription ability *per se*, but in the relative importance of triads, sevenths, inversions, and other musical factors for each annotator. In a context where one must interpret variances, however, one disadvantage of second-order factor models is that it can be difficult to separate how a higher-order factor like General Difficulty is affecting the observed measures as distinct from the first-order factors. The Schmid–Leiman factorization is an equivalent representation of second-order models that can be easier to interpret (Schmid and Leiman, 1957). It separates the loading for each measure into a portion arising exclusively from the higher-order factor and the portions arising from the residual variance of the first-order factors. The factorization is usually standardized so that each loading represents the correlation between a factor – either first- or second-order – and an observed measure. As such, the squared loadings represents the proportions of variance in each measure that are explained by each factor, first-order and second-order.

Table 4 presents the Schmid-Leiman factorization of our chosen confirmatory factor model for each annotator. A number of patterns become clear. Song length has a slightly weaker effect on General Difficulty for Annotator 4 than for the other annotators, but in general, it is responsible for about a quarter of the variance in General Difficulty. For Annotators 1 and 2, the annotator-dependent measures are also influenced by a moderate amount of an independent Annotation Difficulty, whereas Annotators 3 and 4 exhibit no such variation. As mentioned earlier, this independent source of Annotation Difficulty could have something to do with unusual chords or voicings, but a separate study would be necessary to analyze this finding more deeply. At the first-order level, we see that Annotator 2 is highly sensitive to Sevenths Difficulty, and that Annotator 4 is quite sensitive to Inversions Difficulty. The table also includes residual variances, i.e., the proportion of variance due to effects external to the model. Consistent with the earlier tables, the performance of Annotator 2 is more idiosyncratic with respect to the model as compared to the other three annotators. In short, each annotator is indeed unique, exhibiting a distinct pattern of sensitivity to particular types of difficulty in our song sample. Inevitably, these differing sensitivities lead to differing transcriptions.

	General Difficulty				Annotation Difficulty				Residual Variance			
Indicator	A1	A2	A3	A4	A1	A2	A3	A4	A1	A2	A3	A4
Exogenous Predictors												
Song length	.51	.56	.56	.45								
Annotator-dependent												
Difficulty rating	.84	.97	1.08[a]	.91	.49	.37	–	–	–	–	–	.19
Annotation time	.74	.71	.98	.97	.55	.50	–	–	.21	.36	.16	.08
Number of unique chords	.68	.70	.62	.78	.45	.36	–	–	.33	.38	.66	.41
MIREX vocabulary												
TRIADS_INV	.61	.66	.63	.56					–	.21	.04	–
THIRDS_INV	.60	.65	.62	.55					.02	.23	.09	–
MAJMIN_INV	.58	.61	.64	.53					.07	.31	.01	.05
TRIADS	.54	.58	.56	.58					–	.16	.05	–
MAJMIN	.52	.55	.59	.57					.03	.26	–	.01
THIRDS	.52	.57	.54	.58					.02	.19	.11	–
ROOT	.51	.55	.54	.57					.07	.26	.12	–
MIREX	.49	.55	.55	.54					.15	.25	.07	.09
TETRADS_INV	.49	.41	.52	.44					–	.05	–	–
SEVENTHS_INV	.46	.39	.50	.42					.05	.05	–	.02
TETRADS	.41	.33	.42	.40					–	.01	.06	.08
SEVENTHS	.31	.32	.41	.39					.05	–	.04	.03
	Triad Difficulty				Sevenths Difficulty				Inversion Difficulty			
Indicator	A1	A2	A3	A4	A1	A2	A3	A4	A1	A2	A3	A4
MIREX vocabulary												
TRIADS_INV	.67	.55	.62	.55					.42	.22	.42	.71
THIRDS_INV	.68	.56	.62	.56					.39	.20	.37	.65
MAJMIN_INV	.67	.53	.66	.55					.37	.19	.38	.60
TRIADS	.87	.71	.80	.83								
MAJMIN	.84	.67	.84	.82								
THIRDS	.84	.69	.77	.84								
ROOT	.82	.66	.77	.83								
MIREX	.79	.67	.79	.78								
TETRADS_INV	.54	.34	.50	.43	.60	.80	.62	.58	.35	.14	.35	.56
SEVENTHS_INV	.52	.33	.49	.42	.61	.82	.65	.60	.31	.13	.31	.51
TETRADS	.66	.41	.59	.57	.65	.84	.64	.67				
SEVENTHS	.61	.39	.58	.56	.65	.87	.68	.71				

Note. $N = 200$. Although the measurement model is identical for all annotators (see Figure 5), differences in factor and indicator variances across Annotators yield different standardized solutions. Loadings and variances $< .01$ are represented as –.

[a] This Heywood case arises due to the scaling factors in the ordinal regressions.

Output of the R **lavaan** package, version 0.5.23.1097 (Rosseel, 2012).

Table 4: Schmid–Leiman Decomposition of Standardized Factor Loadings and Residual Variance per Annotator

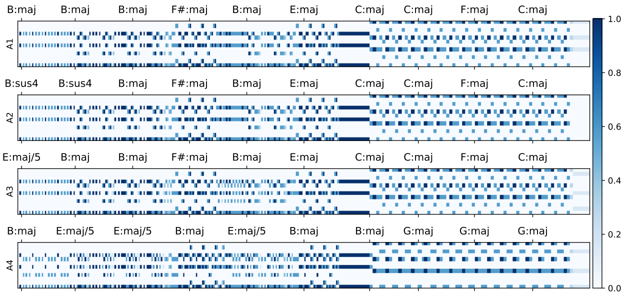

Figure 6: Visualization of annotator subjectivity at the chroma level, for all annotators for *Billboard* dataset song ID 92. The y-axis represents the 12 pitch classes; the x-axis is time. Comparing the chroma reveals large differences in chord detail between annotators. Chroma bins are weighted according to the average MIREX MAJMIN pairwise score, revealing areas of agreement (dark blue) and disagreement (light blue). The figure shows a random sample of chord-labels on beats that have some (nonzero) amount of disagreement.

6. Chord-Label Annotator Subjectivity

The factor analysis in the previous suggest that the relative importance of triads, sevenths, inversions, and other musical factors for each annotator strongly affect annotator subjectivity. Nonetheless, factor analysis must rely on a single set of measures per annotator, and thus it still cannot tell us the extent to which annotators agree among themselves. In this section, we examine a final set of tests on inter-annotator agreement. First, in Section 6.1, we discuss the average pairwise agreement between the annotators using the standard MIREX evaluation measures. After that, in Section 6.2, we discuss the agreement of the annotators with the *Billboard* reference annotations that are commonly used in computational harmony research. These comparisons will give us an intuitive and musically informed idea of the observed proportion of agreement between annotators and of annotators with the *Billboard* annotations. Although the interpretation of these pairwise comparisons is intuitive, we need to adjust for the fact that a certain amount of the agreement could occur due to chance alone. Therefore, in Section 6.2, we discuss the more sophisticated Krippendorff's-α coefficients that measure the inter-annotator agreement of the chord-labels provided by the annotators.

6.1 Pairwise MIREX Chord-Label Agreement

Intuitively, one would expect annotators to agree mostly on fundamental properties of chord labels (e.g. root notes) and would disagree more on intricate parts of chord labels (e.g. inversions and seventh intervals). To investigate how the annotators differ in terms of chord label choice at different chord label granularities, we calculate the average pairwise agreement between all annotators. To this end, we compare the annotations of each annotator with each of the three other annotators, resulting in three agreement scores. The average of these scores shows the average agreement of the four annotators in their transcriptions of each song. By *agreement*, we refer to the commonly used MIREX evaluation of chord-label overlap of the standard MIREX chord-label vocabularies (as explained in Section 5) between two annotations.

The pairwise agreement among all annotators for all fifty songs and all evaluation methods can be found in Figure 7. The rows correspond to the MIREX evaluations; columns correspond to songs. The corresponding *Billboard* dataset IDs can be found below the columns, and the corresponding average reported difficulty scores can be found above the columns. The rows are ordered by average column value, increasing from low average agreement to high. The figure shows that overall, average agreement decreases with an increase in chord-label granularity: annotators agree more on the root notes (ROOT) than on complex chords (e.g., SEVENTHS). Nevertheless, we find that the average agreement of root notes is only .76, with some scores as low as .005. This is surprising, as one would assume that annotators would in general agree on root notes, as well as disagree more on the more intricate chord labels. The root-note disagreement propagates through the disagreement of the other evaluations, which can be seen in the decreasing average agreements plotted at the right x-axis of the figure. This shows that as chord labels become more complex, agreement decreases. The average agreement scores for the remaining chord-label granularities can be found in Table 5.

The amount of detail an annotator can give to a chord label does not end with just the set of pitches. Inversions are an important aspect of harmony, and arguably open to a certain degree of subjectivity. For example, when annotating a song that contains a guitar and a bass guitar, in which the guitarist plays a single chord while the bass guitar plays a descending arpeggio of that chord, an annotator could choose to annotate just the single guitar chord for the entire part but could also choose to include the moving bass line, thereby interpreting it as a new inversion of the same chord for each bass note. Neither of these options is objectively wrong. As a more specific example, Figure 6 shows the differences between annotators for a particular song on the level of *chroma* over time (i.e. a chromagram). Chroma captures the pitch-class content of a chord label in terms of the twelve different pitch classes folded into a single octave. We extracted these chroma using the `mir_eval` software by Raffel et al. (2014). We see that $A1$ annotated rather coarsely, while $A4$ annotated with much more detailed chord labels, inversions, and more frequent chord-label changes.

Figure 7 also shows that for each evaluation measure, the agreement is lower if we take into account inversions. On average the difference is around 5

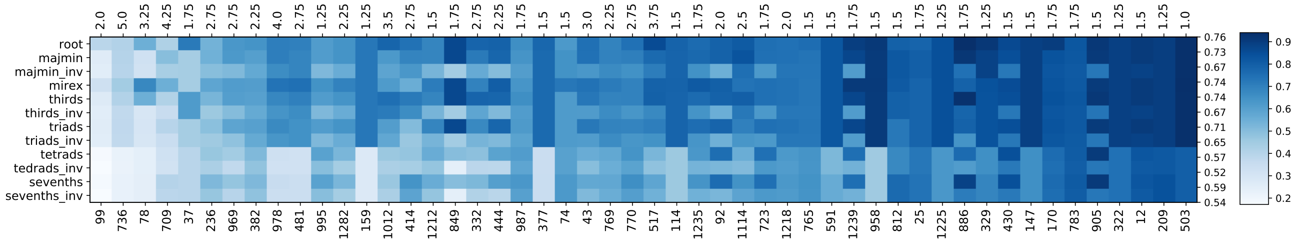

Figure 7: Average pairwise agreement of several MIREX evaluations of all songs in the dataset. Annotator agreement decreases with increased chord-label granularity. The checkerboard-like pattern reveals that for each level of granularity, the level of agreement decreases when inversions are taken into account. *Billboard* dataset IDs can be found below the columns; average reported difficulties can be found above the columns. The numbers on the right show the average agreement for each chord granularity level. Columns are ordered by increasing average pairwise agreement.

	ROOT	MAJMIN	MAJMIN_INV	MIREX	THIRDS	THIRDS_INV	TRIADS	TRIADS_INV	TETRADS	TETRADS_INV	SEVENTHS	SEVENTHS_INV
$\overline{x}$	.76	.73	.67	.74	.74	.67	.71	.65	.57	.52	.6	.54
σ	.19	.2	.24	.18	.19	.24	.21	.24	.24	.24	.24	.25

Table 5: Average ($\overline{x}$) and standard deviation (σ) pairwise agreement results between all annotators. Agreement decreases with increased chord granularity, and is significantly lower when inversions are taken into account.

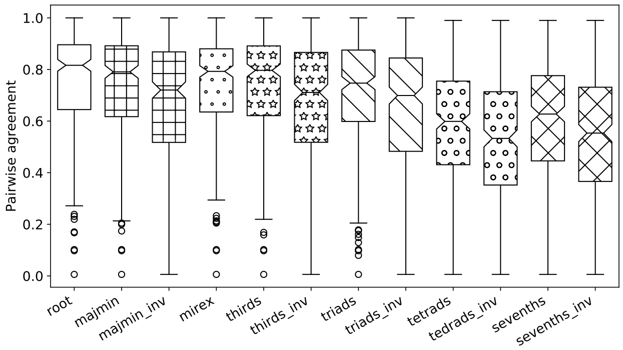

Figure 8: Pairwise agreement among four annotators for all MIREX chord granularity levels. Agreement is significantly lower when inversions are taken into account ($\star$ vs $\star$_inv) with ($p \ll 0.001$).

percentage points, for example, MAJMIN ≈ 0.73 and MAJMIN_INV ≈ 0.67, although the difference in agreement for individual songs can be very large: up to 31 percentage points. All differences are significant in a Wilcoxon signed-rank test to assess whether the results of evaluating a chord granularity level has the same distribution as when taking into account inversions, with $p \ll 0.001$. This shows that for any chord-label type, the amount of annotator subjectivity significantly increases when taking into account inversions. This effect is visualized in Figure 8 which shows the pairwise agreement between all annotators for all MIREX evaluations for all songs.

One could argue that one aspect of a reported difficulty for a song has to do with an annotator's uncertainty about which chord labels to choose for that song: if the annotators find a song to be relatively simple on average, one would expect their chord labels to be relatively more similar. In our dataset, we find indeed that on average, the annotators disagree more when they perceive a song to be more difficult. The average agreement is inversely correlated with the average reported difficulty, $r = -0.6$, $p \ll 0.01$.

6.2 Annotator Agreement with *Billboard* Annotations

The relatively low overall chord-label agreement between expert annotators shown in the previous section raises questions on the creation of one-size-fits all chord-label annotations, which are almost universally used for research relating to computational harmony analysis. One approach to solving the problem of creating chord-label annotations with the broadest appeal is creating a consensus annotation from multiple expert annotations. This was proposed and presented in the *Billboard* dataset. The annotations in this dataset are the result of an expert creating a consensus from two expert annotations (Burgoyne et al., 2011). Assuming that a consensus annotation is on average closer to individual annotations than annotations are to each other, we hypothesize that our annotators would agree on average more with the *Billboard* annotation than with each other. To test in what way our annotators agree with the *Billboard* dataset annotations, we evaluate the annotations from $A1$, $A2$, $A3$ and $A4$ on the corresponding *Billboard* dataset annotation.

Figure 9 shows the pairwise agreement between the annotators and the *Billboard* annotations for all MIREX evaluations. Just like in the results of the Sections 6.1 and 6.2, the figure shows again that overall, agreement decreases with an increase in chord-label granularity: annotators agree more on the root notes (ROOT) than on complex chords (e.g., SEVENTHS) of the *Billboard*

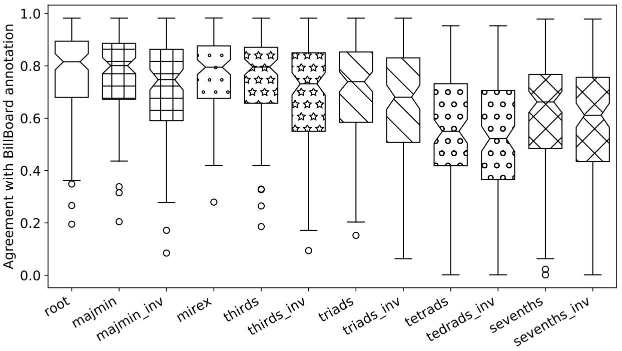

Figure 9: Agreement of the four annotators with the BillBoard annotations for all MIREX chord granularity levels. Agreement is significantly lower when inversions are taken into account ($\star$ vs $\star$_inv) with ($p \ll 0.001$).

annotations. We find that the average agreement of root notes is only 0.77 ($\sigma = 0.16$), with some scores as low as 0.19. The agreement scores for the other chord-label granularities can be found in Table 6.

Figure 9 shows again that for each evaluation measure, the agreement is lower if we take into account inversions. On average the difference is around 5 percentage points, for example, MAJMIN ≈ 0.77 and MAJMIN_INV ≈ 0.72, although the difference in agreement for individual songs can be very large: up to 62 percentage points. All differences in agreement are significant in a Wilcoxon signed-rank test to assess whether the results of evaluating a chord granularity level has the same distribution as when taking into account inversions, $p \ll 0.001$. This shows that for any chord-label type, the amount of annotator subjectivity significantly increases when taking into account inversions.

A first visual comparison of the agreements from Figure 8 and Figure 9 seems to imply that annotators overall agree a little bit more with the *Billboard* annotations than with each other. Nevertheless, none except one of the differences are significant in a Mann-Whitney U test to assess whether the results of annotator agreement has the same distribution as *Billboard* agreement, all $p > 0.05$. The exception is SEVENTHS_INV, $p < 0.05$. While these p-values tell us that there is no significant difference between inter-annotator pairwise agreement and the annotators' agreement with the *Billboard* annotations, we can also measure the magnitude of the difference between groups through the Common-Language Effect Size (CL). CL gives a description of the probability that a score sampled at random from one distribution will be greater than a score sampled from some other distribution. We find CL ranging between 0.48 and 0.56 for the chord granularities, indicating a roughly equal chance of annotators agreeing more with the *Billboard* than with the other annotators. These results show that annotators do not significantly agree more with a *Billboard* annotation than with the annotations from the other three annotators.

These *Billboard* annotations are a staple dataset used in training ACE systems. In 2017, the best performing algorithm in the MIREX ACE task on datasets that intersect with the HASD (Billboard2012 and Billboard2013) reported accuracy scores of .86, .86, .83, .63, and .61 for ROOT, MAJMIN, MAJMIN_INV, SEVENTHS, and SEVENTHS_INV, respectively.[11] Table 7 presents the results for all datasets in the MIREX ACE task. Although our dataset only overlaps with the Billboard2012 and Billboard2013 datasets, they all contain comparable music in terms of genre and popularity. Comparing these to the average pairwise agreement scores found in our dataset shows that the state-of-the-art ACE algorithms perform beyond the *"subjectivity ceiling"* found in our dataset.

6.3 Krippendorff's α Inter-Annotator Agreement

While the pairwise tests in the previous sections provide a musically informed view on the average pairwise agreement between the annotators, it does not account for agreement by random chance. Therefore, we also evaluate the four annotators' chord-labels using Krippendorff's α measure of inter-annotator agreement (Krippendorff, 1970).

Krippendorff's α measures the agreement between annotators on the labeling of units (in our case beats) on a scale from 0 (no agreement), to 1 (full agreement). α becomes negative when disagreement is beyond that what can be expected from chance. Values between .4 and .75 represent a fair agreement beyond chance. To be able to evaluate the chord-labels at the different MIREX granularity levels, we re-label the chord-labels. We follow the standardized MIREX chord vocabulary mappings that were introduced by Pauwels and Peeters (2013). Calculating α for each chord label granularity provides a detailed view into the chance-corrected agreement of the annotators' annotations in our dataset.

Figure 10 shows Krippendorff's α coefficients of all annotators for all songs for all chord-label granularities. Similar patterns as in the average pairwise agreement in Figure 7 can be observed. A higher inter-annotator agreement can be found in root notes (ROOT), with decreasing agreement for more complex chord-label granularities. As a general baseline, $\alpha \geq 0.8$ is often brought forward as good agreement, and $\alpha \geq 0.667$ for where "tentative conclusions are still acceptable" (Krippendorff, 2004). With the exception of ROOT, we find that the average $\alpha \leq 0.667$ indicating a fair inter-annotator agreement. Nevertheless, overall α is quite low for the other chord-label granularities, with arithmetic means ranging from 0.63 (THIRDS, $\sigma = 0.18$) to 0.42 (TETRADS_INV, $\sigma = 0.17$). The figure exhibits the same checkerboard-like pattern as in

[11] http://www.music-ir.org/mirex/wiki/2017:Audio_Chord_Estimation_Results

	ROOT	MAJMIN	MAJMIN_INV	MIREX	THIRDS	THIRDS_INV	TRIADS	TRIADS_INV	TETRADS	TETRADS_INV	SEVENTHS	SEVENTHS_INV
$\overline{x}$	.77	.77	.72	.77	.75	.70	.71	.66	.57	.54	.63	.59
σ	.16	.16	.19	.13	.16	.19	.18	.20	.22	.23	.21	.23

Table 6: Average ($\overline{x}$) and standard deviation (σ) agreement results between the annotators and the *Billboard* annotations. Agreement decreases with increased chord granularity, and is significantly lower when inversions are taken into account.

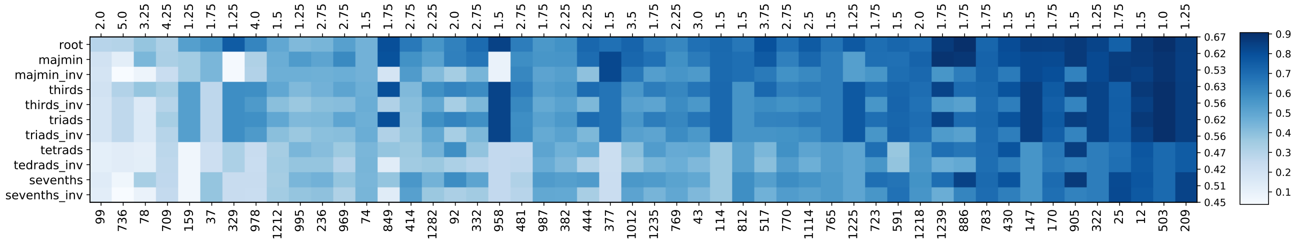

Figure 10: Krippendorff's α inter-rater agreement of all songs in the dataset. The checkerboard-like pattern reveals that for each level of granularity, the level of agreement decreases when inversions are taken into account. *Billboard* dataset IDs can be found below the columns; average reported difficulties can be found above the columns. The numbers on the right show the average agreement for each chord granularity level. Columns are ordered by increasing average pairwise agreement.

Dataset	ROOT	MAJMIN	MAJMIN_INV	SEVENTHS	SEVENTHS_INV
HASD	.76	.73	.67	.6	.54
Isophonics2009	.87 (KBK)	.87 (KBK)	.83 (KBK)	.76 (KBK)	.73 (KBK)
Billboard2012	.86 (KBK)	.86 (KBK)	.83 (KBK)	.63 (WL)	.61 (JLW)
Billboard2013	.81 (KBK)	.78 (KBK)	.76 (KBK)	.58 (WL)	.56 (JLW)
JayChou29	.83 (WL)	.82 (WL)	.79 (WL)	.62 (WL)	.59 (WL)
RobbieWilliams	.89 (KBK)	.88 (KBK)	.85 (KBK)	.83 (KBK)	.81 (KBK)
RWC-Popular	.87 (KBK)	.87 (KBK)	.81 (KBK)	.70 (WL)	.68 (JLW)
USPOP2002Chords	.82 (KBK)	.81 (WL)	.78 (JLW)	.69 (WL)	.66 (JLW)

Note. KBK = Korzeniowski et al. (2017), WL = Wu et al. (2017), JLW = Jiang et al. (2017)

Table 7: MIREX 2017 ACE evaluation results. Evaluation results consistently surpass the subjectivity ceiling found in the HASD.

Figure 7, indicating that the inter-annotator agreement for chord-label granularities is lower when inversions are taken into account.

7. Conclusions and Discussion

In this paper, we presented a new harmonic annotator subjectivity dataset of expert-annotated chord labels of popular songs, and an analysis of the extent of annotator subjectivity found in this dataset. We have shown that the annotators in this dataset each use a particular chord-label vocabulary, with overlap among all annotators of less than 20 percent.

Furthermore, in a pairwise analysis of the annotations using the commonly used MIREX evaluation measures, we find that annotators agree on average on only 73 percent of root notes. This disagreement increases with the complexity of chord labels, with only 59 percent agreement for the most complex vocabulary. Agreement is even lower when we take into account inversions, with an average of 5 percentage points less agreement for chords with inversions. In a comparable experiment using annotations from formally trained amateur musicians, Ni et al. (2013) reported annotator subjectivity of around 10% among the annotators when compared to a consensus. Although the research of Ni et al. concerned amateurs annotators in contrast to our expert annotators, comparable but slight higher amounts of average pairwise agreement can be found in their dataset.

In an inter-annotator agreement analysis using Krippendorff's α, we find disagreements that underline the findings from the pairwise comparisons. Comparing the annotators and the commonly used standard *Billboard* reference annotation, we find that annotators on avererage agree just as much with each other as with the *Billboard* annotations. This suggests that the *Billboard* annotations can be seen as another expert annotation that is equally valid as an expert annotation from our dataset.

The large differences among annotators show that annotator subjectivity is an important factor in harmonic transcriptions, which should figure into serious computational research on harmony. ACE in particular should take annotator subjectivity into account by providing personalized chord labels, tuned to the idiosyncrasies of each user. Ni et al. (2013) similarly found that state-of-the-art ACE systems perform closely to that of the annotators found in their dataset when evaluated on the MAJMIN chord-label granularity. Chord-label estimation performances beyond a subjectivity ceiling suggest that state-of-the-art ACE systems are starting to tune themselves to a particular subjective annotation, and could also be powerful enough for chord-label personalization. In fact, a first approach to such a system has already been introduced by Koops et al. (2017), showing that chord labels can be tuned to an annotator's specific vocabulary from a representation shared by multiple annotators.

We conclude by suggesting that the root causes of annotator subjectivity should be addressed in future research. The first instrument of annotators (i.e., a bias towards listening to the instrument they are accustomed to listening to), their preferred level of transcription detail, their musical sophistication (e.g., instrument and music theory proficiency), and even their harmonic taste (i.e., simply preferring the sound of a chord over another) could all be reasons why annotators differ in their transcriptions. Furthermore, a harmonic similarity analysis of the chord-label annotations provided by annotators could provide insight into the relative distances between the annotators' annotations, if clusters of annotators exist and if these clusters correlate with the possible root causes of annotator subjectivity. As mentioned in the introduction, a vast amount of heterogeneous (subjective) harmony annotations can be found in crowd-sourced repositories. It is currently an unsolved problem how to computationally find useful annotations within these repositories, and how these can be used for computational harmony research. A better understanding of annotator subjectivity would help reveal which crowd-sourced chord-label annotations are within the bounds of subjectivity, therefore appropriate for research. In the long-term, results from the growing body of work that reveals the extent and cause of annotator subjectivity calls for the development of more flexible computational harmony MIR (e.g. ACE) systems that can take into account annotator subjectivity and the reasons why annotators may differ. Moreover, it is not unlikely that annotator subjectivity plays a role in other MIR tasks, as ambiguity plays a large part in music in general.

References

Balke, S., Driedger, J., Abeßer, J., Dittmar, C., and Müller, M. (2016). Towards evaluating multiple predominant melody annotations in jazz recordings. In *Proceedings of the 17th International Society for Music Information Retrieval Conference (ISMIR 2016)*, pages 246–252.

Bosch, J. and Gómez, E. (2014). Melody extraction in symphonic classical music: a comparative study of mutual agreement between humans and algorithms. In *Proc. 9th Conference on Interdisciplinary Musicology–CIM14, Berlin, Germany*.

Brown, T. A. (2015). *Confirmatory Factory Analysis for Applied Research*. Guilford, New York, 2nd edition.

Burgoyne, J., Wild, J., and Fujinaga, I. (2011). An expert ground truth set for audio chord recognition and music analysis. In *Proc. of the 12th International Society for Music Information Retrieval Conference, ISMIR*, volume 11, pages 633–638.

Burgoyne, J., Wild, J., and Fujinaga, I. (2013). Compositional data analysis of harmonic structures in popular music. In *International Conference on Mathematics and Computation in Music*, pages 52–63. Springer.

Chuan, C.-H. and Chew, E. (2007). A hybrid system for automatic generation of style-specific accompaniment. In *Proceedings of the 4th International Joint Workshop on Computational Creativity*, pages 57–64.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

De Clercq, T. and Temperley, D. (2011). A corpus analysis of rock harmony. *Popular Music*, 30(01):47–70.

de Haas, W., Volk, A., and Wiering, F. (2013). Structural segmentation of music based on repeated harmonies. In *IEEE International Symposium on Multimedia (ISM)*, pages 255–258. IEEE.

Fleiss, J. L. (1975). Measuring agreement between two judges on the presence or absence of a trait. *Biometrics*, pages 651–659.

Flexer, A. (2014). On inter-rater agreement in audio music similarity. In *Proc. of the 15th International Society for Music Information*, pages 245–250.

Flexer, A. and Grill, T. (2016). The problem of limited inter-rater agreement in modelling music similarity. *Journal of New Music Research*, 45(3):239–251.

Gauvin, H. (2015). “The Times They Were A-Changin”': A database-driven approach to the evolution of musical syntax in popular music from the 1960s. *Empirical Musicology Review*, 10(3):215–238.

Gwet, K. L. (2014). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC.

Harte, C., Sandler, M., Abdallah, S., and Gómez, E. (2005). Symbolic representation of musical chords: A proposed syntax for text annotations. In *Proc. of the 6th International Society for Music Information Retrieval Conference, ISMIR*, volume 5, pages 66–71.

Hu, X. and Yang, Y.-H. (2017). The mood of chinese pop music: Representation and recognition. *Journal of the Association for Information Science and Technology*, 68(8):1899–1910.

Humphrey, E., Salamon, J., Nieto, O., Forsyth, J., Bittner, R., and Bello, J. (2014). JAMS: A JSON annotated music specification for reproducible MIR research. In *Proc. of the 15th International Society for Music Information Retrieval Conference, ISMIR*, pages 591–596.

Humphreys, L. G. and Jr., R. G. M. (1975). An investigation of the parallel analysis criterion for determining the number of common factors. *Multivariate Behavioral Research*, 10(2):193–205.

Jiang, J., Li, W., and Wu, Y. (2017). Extended abstract for mirex 2017 submission: Chord recognition using random forest model. In *MIREX evaluation results*.

Jones, M. C., Downie, J. S., and Ehmann, A. F. (2007). Human similarity judgments: Implications for the design of formal evaluations. In *ISMIR*, pages 539–542.

Kaliakatsos-Papakostas, M., Cambouropoulos, E., Kühnberger, K.-U., Kutz, O., and Smaill, A. (2014). Concept invention and music: creating novel harmonies via conceptual blending. In *In Proceedings of the 9th Conference on Interdisciplinary Musicology (CIM2014), CIM2014*. Citeseer.

Koops, H., de Haas, W., Bountouridis, D., and Volk, A. (2016). Integration and quality assessment of heterogeneous chord sequences using data fusion. In *Proc. of the 17th International Society for Music Information Retrieval Conference, ISMIR*, pages 178–184.

Koops, H. V., de Haas, W. B., Bransen, J., and Volk, A. (2017). Chord label personalization through deep learning of integrated harmonic interval-based representations. In *Proceedings of the 1st Workshop on Deep Learning for Music*, pages 19–25.

Korzeniowski, F., Böck, S., and Krebs, F. (2017). Mirex ssubmissions for chord recognition and key estimation 2017. In *MIREX evaluation results*.

Krippendorff, K. (1970). Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, 30(1):61–70.

Krippendorff, K. (2004). Reliability in content analysis. *Human Communication Research*, 30(3):411–433.

Lippens, S., Martens, J.-P., and De Mulder, T. (2004). A comparison of human and automatic musical genre classification. In *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, volume 4, pages iv–iv. IEEE.

Mauch, M., Cannam, C., Davies, M., Dixon, S., Harte, C., Kolozali, S., Tidhar, D., and Sandler, M. (2009). OMRAS2 metadata project 2009. In *Late-breaking demo session at 10th International Society for Music Information Retrieval Conference, ISMIR*.

Mauch, M., MacCallum, R., Levy, M., and Leroi, A. (2015). The evolution of popular music: Usa 1960–2010. *Royal Society open science*, 2(5):150081.

McVicar, M., Santos-Rodríguez, R., Ni, Y., and De Bie, T. (2014). Automatic chord estimation from audio: A review of the state of the art. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 22(2):556–575.

Meyer, L. (1957). Meaning in music and information theory. *The Journal of Aesthetics and Art Criticism*, 15(4):412–424.

Ni, Y., McVicar, M., Santos-Rodriguez, R., and De Bie, T. (2013). Understanding effects of subjectivity in measuring chord estimation accuracy. *IEEE Trans-*

actions on Audio, Speech, and Language Processing, 21(12):2607–2615.

Nieto, O., Farbood, M. M., Jehan, T., and Bello, J. P. (2014). Perceptual analysis of the f-measure for evaluating section boundaries in music. In *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR 2014)*, pages 265–270.

Paulus, J. and Klapuri, A. (2009). Music structure analysis using a probabilistic fitness measure and a greedy search algorithm. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(6):1159–1170.

Pauwels, J. and Peeters, G. (2013). Evaluating automatically estimated chord sequences. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 749–753. IEEE.

Raffel, C., McFee, B., Humphrey, E., Salamon, J., Nieto, O., Liang, D., Ellis, D., and Raffel, C. (2014). mir_eval: A transparent implementation of common MIR metrics. In *Proc. of the 15th International Society for Music Information Retrieval Conference, ISMIR*, pages 367–372.

Revelle, W. (2017). *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University, Evanston, Illinois. R package version 1.7.8.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2):1–36.

Salamon, J., Gómez, E., Ellis, D. P., and Richard, G. (2014). Melody extraction from polyphonic music signals: Approaches, applications, and challenges. *IEEE Signal Processing Magazine*, 31(2):118–134.

Salamon, J. and Urbano, J. (2012). Current challenges in the evaluation of predominant melody extraction algorithms. In *ISMIR*, volume 12, pages 289–294.

Schedl, M., Eghbal-Zadeh, H., Gómez, E., and Tkalcic, M. (2016). An analysis of agreement in classical music perception and its relationship to listener characteristics. ISMIR.

Schmid, J. and Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika*, 22(1):53–61.

Schoenberg, A. (1978). *Theory of harmony*. University of California Press.

Seyerlehner, K., Widmer, G., and Knees, P. (2010). A comparison of human, automatic and collaborative music genre classification and user centric evaluation of genre classification systems. In *International Workshop on Adaptive Multimedia Retrieval*, pages 118–131. Springer.

Smith, J. B. L., Burgoyne, J. A., Fujinaga, I., De Roure, D., and Downie, J. S. (2011). Design and creation of a large-scale database of structural annotations. In *ISMIR*, volume 11, pages 555–560.

Van Balen, J., Burgoyne, J., Bountouridis, D., Müllensiefen, D., and Veltkamp, R. (2015). Corpus analysis tools for computational hook discovery. In *Proc. of the 16th International Society for Music Information Retrieval Conference, ISMIR*, pages 227–233.

Van Balen, J., Burgoyne, J., Wiering, F., and Veltkamp, R. (2013). An analysis of chorus features in popular song. In *Proc. of the 14th International Society for Music Information Retrieval Conference, ISMIR*, pages 107–112.

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika*, 41(3):321–327.

Wu, Y., Feng, X., and Li, W. (2017). Mirex 2017 submission: Automatic audio chord recognition with miditrained deep feature and blstm-crf sequence decoding model. In *MIREX evaluation results*.