Annotator subjectivity in harmony annotations of popular music

Koops, H.V.; de Haas, W.B.; Burgoyne, J.A.; Bransen, J.; Kent-Muller, A.; Volk, A.

DOI
10.1080/09298215.2019.1613436

Publication date
2019

Document Version
Final published version

Published in
Journal of New Music Research

JOURNAL OF NEW MUSIC RESEARCH
2019, VOL. 48, NO. 3, 232–252
https://doi.org/10.1080/09298215.2019.1613436

Routledge
Taylor & Francis Group

Annotator subjectivity in harmony annotations of popular music

Hendrik Vincent Koops [a], W. Bas de Haas[b], John Ashley Burgoyne[c], Jeroen Bransen[b], Anna Kent-Muller[d] and Anja Volk[a]

[a]Department of Information and Computing Sciences, Utrecht University, Utrecht, The Netherlands; [b]Chordify, Utrecht, The Netherlands; [c]Music Cognition Group, Institute for Logic, Language and Computation, University of Amsterdam, Amsterdam, The Netherlands; [d]Department of Music, University of Southampton, Southampton, UK

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. Nevertheless, studies showing differences among annotators in many other MIR tasks make the notion of a single 'ground-truth' reference annotation a tenuous one. In this paper, we introduce and analyse the Chordify Annotator Subjectivity Dataset (CASD) containing chord labels for 50 songs from 4 expert annotators in order to gain a better understanding of the differences between annotators in their chord label choice. Our analysis reveals that annotators use distinct chord-label vocabularies, with low chord-label overlap across all annotators. 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. 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. Our results further substantiate the existence of a harmonic 'subjectivity ceiling': an upper bound for evaluations in computational harmony research. Current state-of-the-art chord-estimation systems perform 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, which should inform future studies into harmony perception and computational models of harmony.

ARTICLE HISTORY
Received 25 May 2018
Accepted 4 April 2019

KEYWORDS
Annotator subjectivity; harmony; inter-rater agreement

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 (Burgoyne, Wild, & Fujinaga, 2011; De Clercq & Temperley, 2011; Mauch et al., 2009; Ni, McVicar, Santos-Rodriguez, & De Bie, 2013). These datasets are at the centre of a wide range of important computational studies into harmony, including but not limited to: automatic chord estimation (ACE) (McVicar, Santos-Rodríguez, Ni, & De Bie, 2014), analysis of harmonic trends over time (Burgoyne, Wild, & Fujinaga, 2013; Gauvin, 2015; Mauch, MacCallum, Levy, & Leroi, 2015), computational hook discovery (Van Balen, Burgoyne, Bountouridis, Müllensiefen, & Veltkamp, 2015), chorus analysis of popular music (Van Balen, Burgoyne, Wiering, & Veltkamp, 2013), data fusion of ACE algorithms (Koops, de Haas, Bountouridis, & Volk, 2016), automatic structural segmentation (Haas, Volk, & Wiering, 2013) and computational creativity, such as automatic generation of harmony accompaniment (Chuan & Chew, 2007) and harmonic blending (Kaliakatsos-Papakostas, Cambouropoulos, Kühnberger, Kutz, & Smaill, 2014).

Virtually all of these studies use datasets that contain *single reference annotations*,that is, 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 annotator (Mauch et al., 2009) or a unified consensus of multiple annotators (Burgoyne et al., 2011). 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. Although most creators of these datasets mention the problem of (harmonic) subjectivity and ambiguity, these single reference annotations are nevertheless used in practice as the *de facto* 'ground truth' for a large number of computational studies into harmony

CONTACT Hendrik Vincent Koops h.v.koops@gmail.nl

and related tasks (e.g. Music Information Retrieval Evaluation eXchange[1] [MIREX] ACE).

However, recent research has revealed that among different annotators of the same musical task, often a low agreement is found, which is problematic for creating single reference annotations. A study by Flexer and Grill (2016) demonstrates a low inter-rater agreement between annotators of the MIREX music similarity task. In a study that evaluates automatic approaches for the task of predominant melody estimation, Balke, Driedger, Abeßer, Dittmar, Müller (2016) found that the evaluated performances of these automated methods vary substantially depending on which human annotation is taken as the reference. Balke et al. (2016) argued that the existence of a single 'ground-truth' reference annotation is questionable. Ni et al. (2013) similarly argued that a low inter-rated agreement can be found between harmony transcriptions of different annotators. In a related study, De Clercq and Temperley (2011, p. 95) report an agreement of 94.4% on root judgements and 92.4% on chromatic relative root (root relative to the key). However, these results are based on two annotations made by the authors of the study, which were discussed, compared and corrected for what the authors considered to be errors. In a study using a larger corpus of 200 songs, Temperley Clercq (2013) report 93.3% agreement on chromatic relative root.

Real-world evidence for disagreement between harmonic annotators can be found in the vast amounts of heterogeneous (subjective) harmony transcriptions in crowd-sourced repositories for popular music (e.g. Ultimate-Guitar,[2] Chordify[3]). Often multiple, conflicting harmonic transcriptions of the same song can be found, of which it is unclear which one is 'correct'. Moreover, harmonic disagreement is not only found in (amateur) crowd-sourced annotations, but also between expert harmonic analysts. To provide a more specific example of harmonic disagreement, the next sections provide an overview of disagreement between songwriters, experts and amateur analysts of popular music.

1.1. Harmonic annotator disagreement

A particularly notorious example of harmonic disagreement relates to the opening of the song *A Hard Day's Night* by The Beatles, of which countless music theorists, experts and amateurs have tried to find and explain its particular pitch-class content. Beatles experts, such as Pedler (2010), refer to it as having a 'holy grail' status of 'one of popular music's great unsolved mysteries'. In interviews, the Beatles have given partial and sometimes conflicting definitions of the sonority in terms of a specific chord label. In an interview, George Harrison referred to the chord as an 'F with a G on top (on the 12-string), but you'll have to ask Paul [McCartney] about the bass note to get the proper story' (AKA, 2001). On a different occasion, guitarists Gary Moore and George Harrison were discussing the chord, which Moore thought to be a `G7sus4`. Harrison denied this and showed him a different fingering. Moore responded: 'Are you sure? It doesn't sound like that!', to which Harrison replied: 'Yes, I'm sure, actually, Gary' (Spitz, 2005).

A wide range of expert analysts have tried to decipher the collection of pitches, which has resulted in a myriad of chord possibilities, of which we will provide a couple of examples. We refer to Koops (2019) for a more detailed overview. Spitz (2005) analyses the chord as a 'a `G7`, with an added ninth and a suspended fourth, so unique that it is neither major nor minor'. Hickey (2010) describes it as a `G11sus4`. Womack (2017) writes that Lennon and Harrison played `F` major chords with an added `G` on their 12-string guitars, and McCartney plucked a `D` on his bass. Fujita, Hagino, Kubo, and Sato (1993) defines the main guitar chord as a `Gsus4/D`. Pedler (2010) provides several solutions: `G7sus4` as 'the buskers' choice', but suggests a more accurate label as `G11/D`. Gould (2014) labels the chord a `D` minor 11th. Many more interpretations to the sonority have been proposed, e.g. `Gsus4` (Roessner, 2009) and `G7sus4/A` (Bennett, 2001). Despite his detailed analysis, Pedler (2010) argues that there is no single perfect guitar chord that covers all the bases of the actual sonority. Hickey (2010) refers to the sonority as 'probably the most famous opening chord since Wagner's *Tristan* chord', which similarly has been exposed to numerous harmonic analyses and heated debates (Nattiez, 1990).

Amateur transcriptions differ wildly as well. To facilitate playing popular songs, several online archives exist, containing (mostly crowd-sourced) chord-label annotations for popular music songs. A large amount of variation can be found in these repositories. Table 1 provides an overview of several of these repositories and the chord labels they present for the opening chord of the *A Hard Day's Night*. Most of the versions from Table 1 seem to suggest the label `G7sus4` – 'the buskers' choice' according to Pedler (2010) – which was disputed by composer Harrison himself, according to Gary Moore. Overall, the table shows a wide range of possible chord labels whose pitches in some cases overlap, and in some cases clash.

One could argue that the opening sonority to *A Hard Day's Night* is a particularly difficult one to match with

[1] MIREX is a community-based formal evaluation framework coordinated and managed by the International Music Information Retrieval Systems Evaluation Laboratory.
[2] https://www.ultimate-guitar.com/
[3] https://www.chordify.net/

Table 1. Different online repositories provide different, conflicting chord labels for the opening chord for *A Hard Day's Night* by the Beatles.

Source	Number of transcriptions	Chord label(s)
ultimate-guitar.com	8	`G7sus4/D`, `D7sus4`, `D7sus4/A`, `Fadd9`, `Fadd9/A`, `Fadd9/C`, `Fadd9/D`, `C7sus4`, `G7sus4`, `Gsus4`, `Fadd9`, `Dadd9*`, `Dadd9/F#` (capo at fret 3), `D7sus4`, `Gadd9sus4` (no 3rd), `F6/9` (no tonic)
e-chords.com	1	`G4(7)`
guitaretab.com	1	`G`
tabs4ukulele.com	1	`G7sus4`
ukulele-tabs.com	1	`G7sus`
chords-and-tabs.net	3	`Fadd9`, `G`, `C9`
guitarfloor.com	2	`A#add9`, `G7/4`
guitartabsexplorer.com	4	`G`, `C9`, `Fadd9`, `G7sus4`

Note: Transcriptions retrieved April 2019.

a chord label, one that is not representative of chords in general. Although this particular example is indeed notoriously complex, variation is ubiquitous in harmony analyses of popular music. Doll (2017) provides a large number of examples of songs which are built from the same basic 'chordal loops', yet convey centric and functional ambiguity because of different musical dimensions (meter, texture, parallelism). The wide use of production effects that distort the original content of the acoustic signal in some way is an important factor in perceptual harmonic ambiguity.

George Martin, the long-time producer of the Beatles, referred to (harmonic) ambiguity stemming from production effects (such as distortion) when he said: 'it shouldn't be expected that people are necessarily doing what they appear to be doing on records' (Martin & Hornsby, 1994). Martin argues that recording a song is like film making, where all sorts of effects are used in order to create aural illusions. However aesthetically pleasing these illusions might be, they are problematic for determining what chords are played in a musical recording.

The right amount of signal distortion can aurally *imply* some harmonies that are not physically played by the musician. Rock guitarists for example often use *power chords*, which contain only a root note and a fifth interval. When these power chords are distorted, the complex overtones of the fifth interval are amplified and can imply chord notes that are not part of the original power chord. A distorted perfect fifth has an overtone at the major tenth (one octave plus a major third), which can be perceived as a chordal major third. It is unclear whether this major third should be regarded as part of the sonority, and thus part of a label in a transcription. Doll (2017) notes for example that the power chords in the opening of Bikini Kill's 'Rebel Girl' imply a major third, which is not actually played by the performers.

Another example of implied harmony is by sketching out a harmony and relying on the listener's culturally or stylistically informed expectations to perceptually extend the sonorities to 'hear' the complete harmony. Moore (2012) provides the example of the 'boogie pattern' in Chuck Berry's 'Johnny B. Goode'. Here, the guitar plays alternating fifth and sixth degrees of the scale, above an open string root where possible. Since the roots of this technique lie in the blues tradition, it is commonly interpreted as a sequence of major chords from a blues chord sequence. In this case, the choice of chord labels is less grounded on the songs' actual harmonic content, but more on its relationship to a particular cultural or historical context. In summary, disagreement in the perception of harmony is a common phenomenon. Reasons for diverging analyses are differences in application of music theoretical concepts to ambiguous musical passages, and fundamental and inherent human perceptual and cultural differences that stem from enculturation differences.

1.2. Studying harmonic disagreement

In the previous section, we discussed well-known examples of disagreement. While it is generally known in musicology that different analysts for the same piece may have different opinions, its investigation in computational studies is rare. Therefore, we introduce here a new chord-label dataset containing *multiple reference annotations* for 50 songs from the *Billboard* dataset.[4] Specifically, the new dataset includes four different annotators' transcriptions of each song. The study of annotator subjectivity is important for perceptual and computational harmony research. The former relates to gaining insight into the perceptual variation of human harmony hearing, while the latter relates to questions of how to model harmony for computational approaches, such as automatic chord estimation. For example, the large differences in

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

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 (Humphrey & Bello, 2015; Ni et al., 2013).

As a first step into studying disagreement in this dataset, we investigate the phenomenon described above: having different, perhaps conflicting harmonic descriptions for the same piece of music. More specifically, we investigate *harmonic annotator subjectivity*: the agreement (or variation) found between chord labels provided by expert annotators of popular music. In a previous experimental study concerning 5 annotators and 20 songs by The Beatles and Queen, Ni et al. (2013) already showed that annotators transcribing the same music recordings disagree on roughly 10 percent of harmonic annotations. This study expands on work such as Ni et al. (2013) and De Clercq and Temperley (2011) by studying harmonic annotator subjectivity on a larger scale in a wider range of popular music, using a standardised chord-label encoding. The study of annotator subjectivity in expert annotations, in this paper, aims to reveal the variability found in human annotations, which arguably should be taken into account in modelling human harmonic perception.

Contributions. The contribution of this paper is twofold. First, we introduce the *Chordify Annotator Subjectivity Dataset*. This publicly available chord-label dataset is linked with other important datasets containing harmonic transcriptions, as well as with major audio music repositories. Second, we investigate annotator subjectivity within this dataset and show that when using the standard ACE evaluations measures, significant differences exist between transcriptions, 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.

Synopsis. The remainder of this paper is structured as follows. Section 2 discusses related work in the analyses of human judgements in music information retrieval research. In Section 3, we introduce our dataset and 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. In Section 5, we explore the agreement between annotators using pairwise analyses, a comparison with the *Billboard* reference annotation and inter-rater agreement statistics. Section 6 describes the individual differences between annotators using a factor analysis. The paper closes with a discussion and conclusion in Section 7.

2. Related work in analysis of human judgements in Music Information Retrieval

Disagreement between human annotators is a well-known problem in a wide variety of tasks in MIR research. The lack of an exact task specification, the differences in the annotators' experiences, musical background, skill level, and instrumental preference, or the use of different annotation tools are some of the possible causes of disagreement between annotators (Balke et al., 2016; Salamon, Gómez, Ellis, & Richard, 2014; Salamon & Urbano, 2012). Annotator disagreement has previously been studied in the contexts of genre classification (Lippens, Martens, & De Mulder, 2004; McVicar et al., 2016; Seyerlehner, Widmer, & Knees, 2011), audio music similarity (Flexer, 2014; Flexer & Grill, 2016; Jones, Downie, & Ehmann, 2007), music structure analysis (Nieto, Farbood, Jehan, & Bello, 2014; Paulus & Klapuri, 2009; Smith, Burgoyne, Fujinaga, De Roure, & Downie, 2011), melody extraction (Balke et al., 2016; Bosch & Gómez, 2014), musical tempo extraction and beat tracking (McKinney, Moelants, Davies, & Klapuri, 2007), ratings of guitar tabs (Macrae & Dixon, 2011) 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 judgements 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 of homogeneity or consensus between different annotators (or *raters*). With high inter-annotator agreement, annotators can be used interchangeably without having to worry about the categorisation being affected by a significant annotator factor. In other words, if interchangeability of annotators is guaranteed, then their labels (or ratings) can be used with confidence without asking which annotator 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, 2010).

To study inter-annotator agreement, several measures have been introduced, of which joint probability of agreement is the simplest and least robust measure. 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 & Yang, 2017). Jones et al. used Fleiss's κ to analyse human similarity judgments of symbolic melodic similarity and audio music similarity (Jones et al., 2007). Balke et al. (2016) adapted Fleiss' Kappa for evaluating multiple predominant melody annotations in jazz recordings (Balke et al., 2016).

In this study, in addition to pair-wise analyses and descriptive statistics of our dataset, we use Krippendorff's α to assess the inter-annotator agreement between the chord labels of popular music from four annotators. 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, Eghbal-Zadeh, Gomez, and Tkalčič (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. Chordify Annotator Subjectivity Dataset

In this section, we introduce a new dataset containing chord labels for 50 songs from 4 different annotators, time-aligned with commercially available audio recordings. Each transcription represents the subjective opinion of one out of four harmony experts in popular music. After a brief introduction to the *Billboard* dataset, the following sections detail the annotation procedure of the *Chordify Annotator Subjectivity Dataset* (CASD).

Burgoyne et al. (2011) introduced the *Billboard* dataset containing chord label annotations for songs sampled from the *Billboard 'Hot 100'* music charts, the definitive weekly ranking of the most popular songs in North America (Bradlow & Fader, 2001). Each *Billboard* annotation represents the consensus of two experts in jazz and popular music for a popular song. This data set has quickly become a standard reference set for several MIR tasks relating to harmony such as ACE. Therefore, we model our data set after the *Billboard*. Specifically, we (a) sample songs from the *Billboard* to have a widely used and accepted reference annotation to compare the annotation from our annotators to, and (b) employ a similar annotator selection process. In this way, our dataset can be seen as an incremental step for the *Billboard* dataset to include multiple annotators for the study of annotator subjectivity.

3.1. Song selection

Currently available chord-label annotation datasets containing more than one reference annotation are limited by size and song sampling strategy (e.g. Ni et al., 2013) or lack a standardised chord-label encoding (e.g. De Clercq & Temperley, 2011). Therefore, we selected 50 songs from the *Billboard* dataset that have a stable online presence in widely accessible music repositories (e.g. official artist YouTube channel uploads), with chord-label annotations in an encoding that is standardised in MIR research.

In 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 plays and selected the top 50 songs by this ranking. At the time they were collected, the least-viewed song in the dataset had 76,000 plays and the most-viewed song over 13 million plays, and an average of 11.9 unique chords according to the *Billboard* reference annotations.

3.2. Annotator selection

To ensure we obtained high-quality transcriptions, we set out to find annotators who studied music and harmony at the undergraduate or graduate level. Furthermore, to ensure familiarity with the music used in this study, we searched for annotators who were experienced in playing (e.g. in cover bands) as well as transcribing popular music. We found four annotators that were successful professional musicians with a broad knowledge of harmony and chords, who have an academic degree in music and have between 15 and 25 years of experience on their primary instrument.

Annotator 1 (A1) is a professional music transcriber and composer of popular and classical music. A1's transcriptions have been published by several publishing houses and websites. A1 studied music theory and composition and has been an active guitar player for 15 years.

Annotator 2 (A2) studied guitar at a conservatoire and has been playing the guitar and drums for 19 years. A2 plays in several bands semi-professionally, runs a music production company and teaches guitar and drum.

Annotator 3 (A3) studied piano and composition at the undergraduate level. A3 is a prize winning composer for film, television and commercials, and has been playing piano for 25 years. In addition to being a composer, A3 plays piano in several cover bands and works as a professional pop music transcriber.

Annotator 4 (A4) is a classically trained pianist who has been playing popular and classical music for over 20 years. After graduating from a conservatoire, A4

studied new music, sound design and contemporary music composition at the graduate level. A4 furthermore teaches piano lessons, works as a producer and plays in several bands.

With this musical expertise, we assume that these annotators are capable of producing harmonic annotations for popular music, like those created for the *Billboard* dataset. After inviting the annotators to join the experiment, we reviewed their first 10 transcriptions to ensure they had sufficient aptitude to continue; all four annotators completed the initial screening successfully and were hired to continue to annotate the remaining 40 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. Their task was therefore a *practical* one: to listen to the music and transcribe the chord labels of the songs as they perceive them, so they could reproduce what they have heard. This task hence differs from a music theoretic analysis of a musical score according to a particular harmony theory and is very similar to the set-up in the annotation tasks in Burgoyne et al. (2011) and Ni et al. (2013).

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. We assumed that studying chord-label subjectivity beyond the beat level would have a marginal effect on our findings, because (a) chord labels are rarely notated beyond the beat level and (b) sub-beat notation would introduce a positive bias towards subjectivity. The standard YouTube web player was used to provide the reference recording of the song. Through the interface, the annotators were free to select a chord label for each beat from a drop down menu with all chord labels that are available in the *Billboard* dataset. If the chord label of their choice was not available, the annotators notified us and we added the chord label to the system. In this way, the annotators were completely free to choose any chord label for each beat. While transcribing, the annotators were able to watch and listen to the YouTube video of the song, and if they wanted, a synthesised version of their chord transcription. The chord labels are encoded using the commonly used chord-label syntax introduced by Harte, Sandler, Abdallah, and Gómez (2005). This syntax provides a simple and intuitive encoding that is highly structured and unambiguous to parse with computational means.

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 reproducibility and stimulates future research into annotator subjectivity, we adopted a number of standard chord-label and annotation practices that are commonly used in MIR research.

For each of the 50 songs, the dataset contains the chord labels provided by our four annotators. 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 have made the dataset publicly available in a Git repository.[5] Through utilising 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 analyse the annotations in terms of descriptive statistics. First, we analyse the difficulty scores and remarks (see Section 4.1) and the overall chords the annotators provided (see Section 4.2). Next,

[5] https://www.github.com/chordify/CASD

we provide an analysis of the differences in chord labels used by the annotators (see Section 5). 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 on average 20.25 minutes to transcribe a song ($\sigma = 11.55$), with a minimum of 5 minutes and a maximum of 60 minutes. The annotators also ranked their perceived difficulty of all songs on a scale from 1 (easy) to 5 (difficult). On average, the annotators gave the songs a difficulty rating of 2.1 ($\sigma = 1.09$). The average annotation times and reported difficulty for each annotator can be found in Table 2.

Naturally, 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 reported difficulties and annotation times, we find a very strong positive linear correlation, $r = 0.93$, $p \ll 0.05$. The correlations per annotator appear in Figure 1. The figure shows that for $A1$ and $A2$, the correlation is very strong, $r = 0.92$ and $r = 0.84$, respectively. $A4$'s measurements are also strongly correlated ($r = 0.76$); $A3$ shows a strong correlation that is nonetheless perhaps weaker than the rest ($r = 0.61$). Figure 1 shows that $A3$'s annotations cluster around 20–30 minutes 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 6 we will see that the order of songs does not have a significant effect on annotation time and perceived difficulty for any annotator.

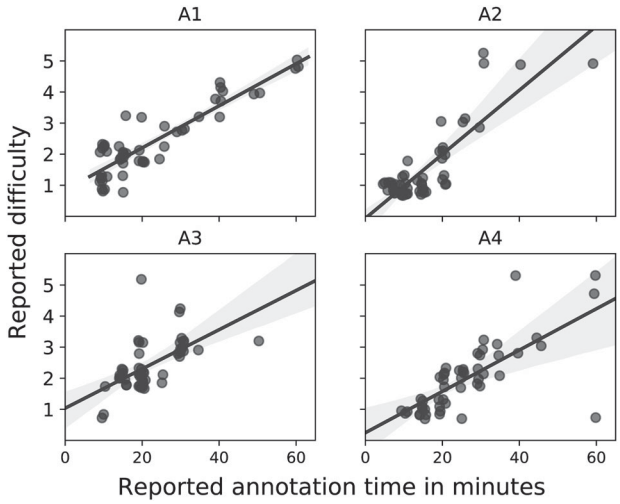

Figure 1. We find strong ($r = 0.93, 0.84, 0.61, 0.76$ for A1, A2, A3, A4), but differing, correlations per annotator between reported annotation time and reported difficulty from 1 (easy) to 5 (hard). In general, songs perceived as difficult took longer to annotate than easy songs. Translucent bands around the regression line indicate the confidence interval for the regression estimate.

4.2. Chord-label statistics

Turning to the harmonic transcriptions, we investigate the disagreement in terms of chord labels in our dataset. To better explain the chord-label content of our dataset, we first provide an example of chord labels provided by the annotators for couple of bars of a given song. Although the score was not provided to the annotators during the annotation task, we provide it here to visualise the chords provided by the annotators. Figure 3 provides such an example, in which we have aligned the chord labels of each annotator to the score of Rick James' 'Super Freak'.

This example shows a variety of different types of annotator disagreement and is taken from bars 24 to 28 of *Super Freak*. Here we can see basic disagreements, such as in the second half of bar 28 where the annotators disagree on whether the chord is `A:min` or `A:min7` – having only a single pitch class disagreement (agreeing on `A`, `C`, `E` but disagreeing on the importance of the pitch `G` in the harmony). More complex levels of disagreement are present for example between `E:sus4` and `A:min` in bar 25. However, when we observe the pitch class content of these chords we can see the pitches `E` and `A` feature in both chords (`E`, `G#`, `A`, `B` for `E:sus4`, and `A`, `C`, `E` for `A:min`).

Table 2. Overview of annotators, their primary instrument, musical background and average (and standard deviation) annotation time (in minutes), number of chord labels per song, and reported difficulty statistics. Difficulty is reported on a scale from 1 (easy) to 5 (hard).

Annotator	Primary instrument	Background/occupation	Annotation time (min)	Reported difficulty	Unique Chord labels per song
A1	Guitar	Transcriber, composer	23.10 (14.91)	2.40 (1.16)	9.46 (5.13)
A2	Guitar	Musician, teacher	15.66 (9.91)	1.60 (1.18)	9.42 (4.20)
A3	Piano	Transcriber, composer	22.00 (7.42)	2.42 (0.73)	12.44 (5.83)
A4	Piano	Composer, producer	26.10 (12.18)	1.96 (1.07)	8.86 (4.70)

A similar disagreement can be found when exploring the difference between `F:maj` and `A:min` from bar 24. `F:maj` and `A:min` are similar in pitch-class make up (`F`, `A`, `C` for `F:maj`, and `A`, `C`, `E` for `A:min`), sharing the pitch classes `A` and `C`. Observing the beginning of bar 28, we can see the annotators disagree on whether the first two beats of each bar are in `D` major (with *A*1 specifying the power chord `D:5`) or `A` minor as perceived by *A*3. This disagreement features prominently throughout the piece and could be attributed to the harmonic disagreement present between the falling bass guitar riffs that relate more to `D` and the piano part that relates more to `A`. Interestingly, the guitar players in our dataset more often choose the chords that related to the guitar part of the piece, while the piano players in our dataset more often choose the chords that related to the piano part. This example is also interesting harmonically, due to the *N.C.* notated above the stave over the guitar riff, indicating 'no chord', which suggests no chord should be inferred because the arranger of this score does not consider this to be a harmony. However, it is important to note here that the annotators did not have access to the score. Their task was to transcribe the harmony by ear, and as such, this kind of information was not available to the annotators. Consequently, these parts are in fact annotated with chord labels by each annotator.

For all songs in our dataset, we analyse the chord-label annotations in several ways. First, we investigate which chord labels are used in our dataset and how much overlap in chord-label vocabularies there is among annotators. We investigate this overlap on the level of string representation of the chord labels, and as such at this point do not take mitigating factors such as enharmonic equivalence into account. This analysis will provide a general indication of annotator subjectivity in our dataset, as it shows the difference in the use of absolute chord label strings among annotators. Then we analyse the number of unique chord labels in a song and its reported difficulty. Further on in this paper, in Section 5, we will analyse the agreement between annotators using standard chord-label overlap measures.

4.2.1. Chord-label vocabularies

On average, the four annotators (*A*1, *A*2, *A*3, *A*4) used 10.3 unique chord labels per song ($\sigma = 5.2$), with a minimum of 3 and a maximum of 27 unique chord labels. Individually, the averages per annotator were 9.46 unique chord labels ($\sigma = 5.13$), 9.42 unique chord labels ($\sigma = 4.2$), 12.44 unique chord labels ($\sigma = 5.83$) and 8.86 unique chord labels ($\sigma = 4.7$) for *A*1, *A*2, *A*3 and *A*4, respectively. These statistics are similar to what was found by Burgoyne et al. (2011) in the *Billboard* dataset, in which songs contain on average 11.8 unique chord labels.

Altogether, the annotators used 290 unique chord labels in their transcriptions. The most frequently used chords are common chord labels such as `G:maj`, `C:maj`, `D:maj`, and `A:maj`. Annotators *A*1, *A*2, *A*3 and *A*4 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 are inversions. Nevertheless, the annotators differ in the amount of chord labels that are inversions. Of all the chord label tokens that the annotators *A*1, *A*2, *A*3 and *A*4 use, 8, 4, 15 and 16 percent are inversions, respectively. Of their unique chord label types, 26, 27, 43, 39 percent are inversions for *A*1, *A*2, *A*3, and *A*4 respectively. This seems to suggest that while there is relatively little disagreement on pitch spelling, there is a large amount of disagreement on the level of inversions. If we consider all possible inversions of a chord label, 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 *A*1 and *A*3 share the most chord labels (104). Fewer chord labels are shared between *A*2 and *A*4 than with the rest. This is interesting, as *A*1 and *A*3 are both guitar players, and *A*2 and *A*4 are piano players. This seems to suggest that our piano players are on average more diverse in terms of their chord-label

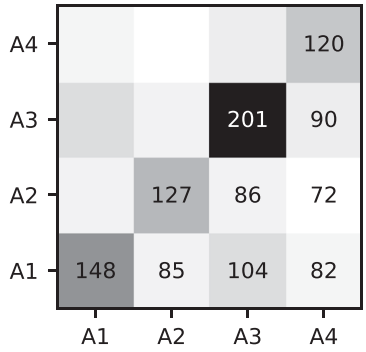

Figure 2. Pairwise intersection sizes of all 290 unique chord labels in the dataset for all annotators. On average, the annotators share around half of their chord label vocabulary with the other annotators.

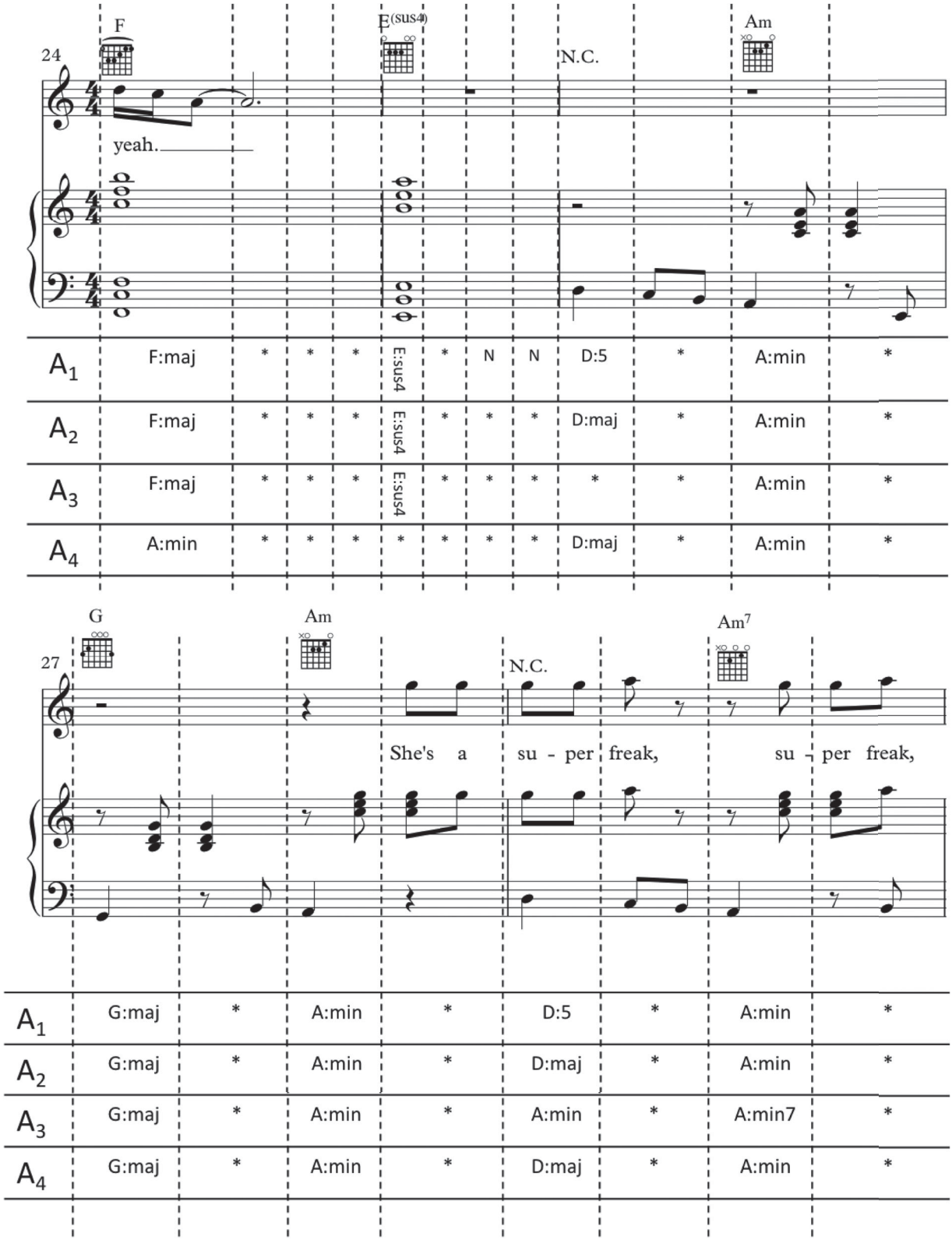

Figure 3. Bars 24–28 of Rick James' 'Super Freak'. The figure shows the musical score aligned with the chord labels provided by the annotators. A star (⋆) indicates a repetition of the previous chord label. Note that the annotators transcribed the harmony of the song solely by ear from the original audio recording. The score was not provided to the annotators.

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 unique chord labels, and therefore a large number of chord changes should be harder to transcribe than songs with a small number of unique chord labels. We indeed find a positive correlation between the reported difficulty of a song and the number of unique chord labels for that song. In Figure 4, 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 5 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, $r = 0.80$, $p \ll 0.01$. Nevertheless, when we turn to individual annotators, we see that not all correlations are similar for all annotators. For $A1$ ($r = 0.79$) and $A4$ ($r = 0.75$) the degree of correlation is comparable, but the correlations for $A2$ ($r = 0.67$) and $A3$ ($r = 0.65$) are strong but somewhat weaker.

In an inspection of Figure 4, 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 to 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.

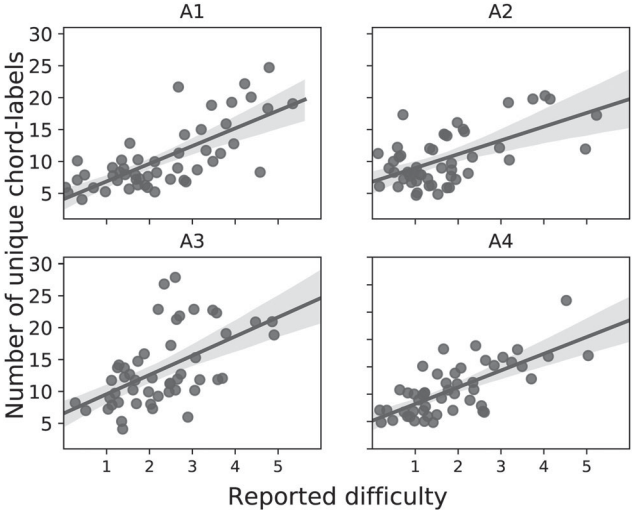

Figure 4. Reported difficulty and number of chord labels per song are strongly correlated, with $r = 0.79, 0.75, 0.67, 0.65$ and $p < 0.01$ for A1, A2, A3, A4. The larger the number of unique chords used, the more difficult to annotate was the song perceived.

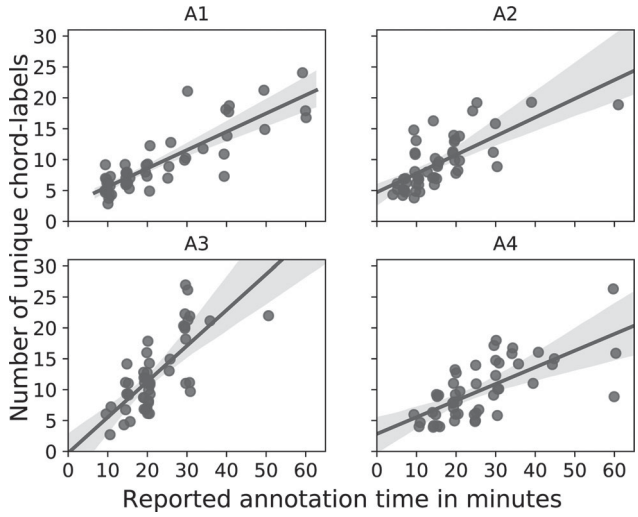

Figure 5. Annotation time and number of chord labels per song are strongly correlated, with $r = 0.86, 0.73, 0.74, 0.69$ and $p < 0.01$ for A1, A2, A3, A4. The larger the number of chords used, the more time it took to annotate.

5. Chord-label annotator subjectivity

In this section, we examine a set of formal measures of inter-annotator agreement. First, in Section 5.1, we discuss the average pairwise agreement among the annotators of CASD using the standard MIREX evaluation measures. After that, in Section 5.2, we discuss the agreement of the annotators with the original *Billboard* reference annotations that are commonly used as a ground truth in computational harmony research. Finally, although pairwise comparisons are intuitively easy to understand, we need to take into account that a certain amount of agreement could be attributed to chance. Therefore, in Section 5.3, we discuss the more sophisticated Krippendorff's α coefficients that measure the inter-annotator agreement of the chord labels provided by the annotators.

5.1. Average pairwise chord-label agreement

In general, 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

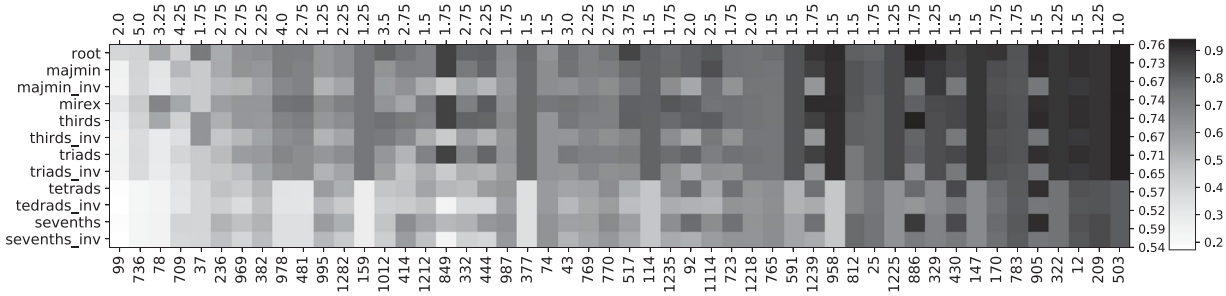

Figure 6. Average pairwise agreement of several MIREX evaluations of all songs in the dataset. Annotator agreement decreases with increased chord-label granularity. The chequerboard-like pattern reveals that for each level of granularity, the level of agreement decreases when inversions are taken into 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.

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 standard MIREX weighted chord symbol recall (WCSR) metrics, i.e. the proportion of correct labels weighted by song duration, after the chord labels from the annotators 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 fifth; 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 example, `C:7` and `C:9` are equivalent but `C:7` and `C:maj7` are not. For MAJMIN, THIRDS, TRIADS, TETRADS and SEVENTHS, we also test with inversions: MAJMIN_INV, THIRDS_INV, etc.

Before computing the agreement, `mir_eval` first separates each chord label into three parts: the root, the set of root-invariant intervals and the bass interval. Depending on the vocabulary, `mir_eval` then compares the corresponding parts. ROOT for example compares only the root note, while MAJMIN requires equal roots as well as the intervals subject to the reference chord quality being major or minor. For a detailed explanation of these measures, we refer the reader to the standardised MIR evaluation software package `mir_eval` by Raffel et al. (2014) and the MIREX ACE website.[6]

The pairwise agreement among all annotators for all 50 songs and all evaluation methods can be found in Figure 6. 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 to high average agreement. 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 0.76, with some scores as low as 0.005. This is surprising, as one would assume that annotators would in general agree on root notes and 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 3.

Tuning. We investigated whether the low root note agreement could be caused by tuning issues. After all, the right amount of tuning deviation from 440 Hz concert tuning in an audio recording could cause annotators to disagree on root notes by exactly one semitone, resulting in zero root note agreement. Using the Pearson correlation coefficient to measure linear correlation, we found no significant correlation between absolute tuning (as measured using the method proposed by Mauch

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

Table 3. Average ($\bar{x}$) 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.

Chord label vocabulary	$\bar{x}$	σ
ROOT	0.76	0.19
MAJMIN	0.73	0.20
MAJMIN_INV	0.67	0.24
MIREX	0.74	0.18
THIRDS	0.74	0.19
THIRDS_INV	0.67	0.24
TRIADS	0.71	0.21
THIRDS_INV	0.65	0.24
TETRADS	0.57	0.24
TETRADS_INV	0.52	0.24
SEVENTHS	0.60	0.24
SEVENTHS_INV	0.54	0.25

& Dixon, 2010) and agreement, nor could we find a significant correlation between absolute tuning deviation from 440 Hz and agreement. To test the impact of individual detuned songs on the overall agreement, we iteratively removed the top most detuned songs from 1 to 15, and tested whether agreement significantly changes. Still, no significant change in agreement was found. This suggests that the tuning difference found in the audio of our dataset is not a major contribution to annotator disagreement.

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 visual example of disagreement, Figure 7 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 12 pitch classes folded into a single octave. We extracted these chroma using the `mir_eval` software by Raffel et al. (2014). The figure reveals that the annotators all agreed on the pitch classes of the opening chord of the song. Nevertheless, we see that *A*1 used a low number of chord-label changes, while the others were more meticulous in their chord-label choice.

Figure 6 also shows 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 ≈ 0.73 and MAJMIN_INV ≈ 0.67, although the difference in agreement for individual songs can be very large: up to 31 percentage points. All differences (i.e. the five pairs of X vs. X_inv) are significant in a Wilcoxon signed-rank test to assess whether the results of evaluating a chord granularity level have the same distribution as when taking into account inversions (e.g. MAJMIN vs. MAJMIN_INV), 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 visualised 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,

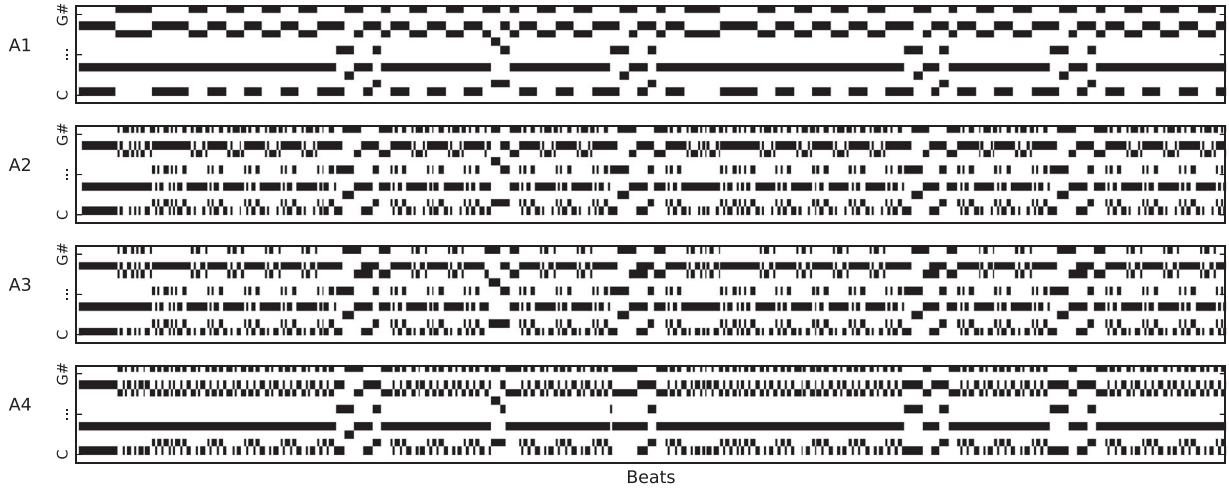

Figure 7. Visualisation of annotator subjectivity at the pitch class level, for all four annotators for the *Billboard* dataset song with ID 995. The *y*-axes represent the 12 pitch classes for each annotator; the *x*-axes represent time in terms of beats. Comparing the chromagrams reveals large differences in chord detail between annotators.

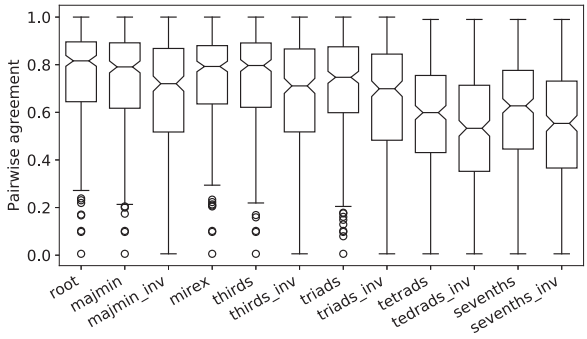

Figure 8. Pairwise agreement among four annotators for all MIREX chord granularity levels. Agreement is significantly lower when inversions are taken into account (⋆ vs ⋆_inv) with ($p \ll 0.001$).

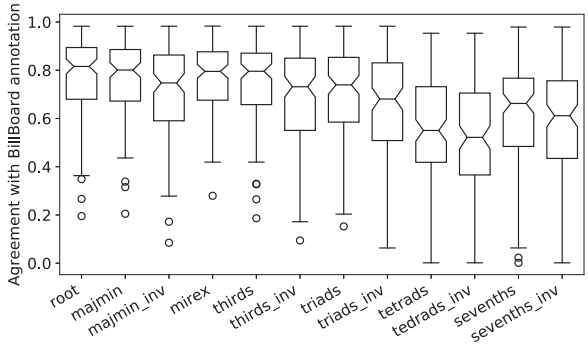

Figure 9. Agreement of our four individual annotators with the consensus from the Billboard annotation for all MIREX chord granularity levels. Agreement is significantly lower when inversions are taken into account (⋆ vs. ⋆_inv) with ($p \ll 0.001$).

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, $r = -0.6$, $p \ll 0.01$.

5.2. Annotator agreement with Billboard annotations

The relatively low overall chord-label agreement between 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 how to create chord-label annotations with the broadest appeal is creating a consensus annotation from multiple annotations. This was proposed and presented in the *Billboard* dataset. The annotations in this dataset are the result of a meta-annotator creating a consensus from two annotations (Burgoyne et al., 2011). Assuming that a consensus annotation is on average closer to individual annotations than annotations of individuals are to each other, we hypothesise that our annotators would agree on average more with the *Billboard* annotation than with each other. To test this, 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 Sections 5.1 and 5.2, 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* 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 4.

Table 4. Average ($\bar{x}$) 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.

Chord label vocabulary	$\bar{x}$	σ
ROOT	0.77	0.16
MAJMIN	0.77	0.16
MAJMIN_INV	0.72	0.19
MIREX	0.77	0.13
THIRDS	0.75	0.16
THIRDS_INV	0.70	0.19
TRIADS	0.71	0.18
TRIADS_INV	0.66	0.20
TETRADS	0.57	0.22
TETRADS_INV	0.54	0.23
SEVENTHS	0.63	0.21
SEVENTHS_INV	0.59	0.23

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 have 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 comparison of the agreements from Figures 8 and 9 seems to imply that annotators overall agree a little bit more with the *Billboard* annotations than with each other. Nevertheless, only one of the differences is significant in a Mann–Whitney U-test, which assesses whether the results of annotator agreement have the same distribution as *Billboard* agreement, with $p > 0.05$. Thus being the SEVENTHS_INV, $p < 0.05$. While these results show us that there is no significant difference between

Table 5. MIREX 2017 ACE evaluation results. Evaluation results consistently surpass the subjectivity ceiling found in the HASD.

Dataset	ROOT	MAJMIN	MAJMIN_INV	SEVENTHS	SEVENTHS_INV
HASD	0.76	0.73	0.67	0.6	0.54
Isophonics2009	0.87 (KBK)	0.87 (KBK)	0.83 (KBK)	0.76 (KBK)	0.73 (KBK)
Billboard2012	0.86 (KBK)	0.86 (KBK)	0.83 (KBK)	0.63 (WL)	0.61 (JLW)
Billboard2013	0.81 (KBK)	0.78 (KBK)	0.76 (KBK)	0.58 (WL)	0.56 (JLW)
JayChou29	0.83 (WL)	0.82 (WL)	0.79 (WL)	0.62 (WL)	0.59 (WL)
RobbieWilliams	0.89 (KBK)	0.88 (KBK)	0.85 (KBK)	0.83 (KBK)	0.81 (KBK)
RWC-Popular	0.87 (KBK)	0.87 (KBK)	0.81 (KBK)	0.70 (WL)	0.68 (JLW)
USPOP2002Chords	0.82 (KBK)	0.81 (WL)	0.78 (JLW)	0.69 (WL)	0.66 (JLW)

Note. kbk, Korzeniowski, Böck, Krebs, and Widmer (2017); wl, Wu, Feng, and Li (2017); jlw, Jiang, Li, and Wu (2017).

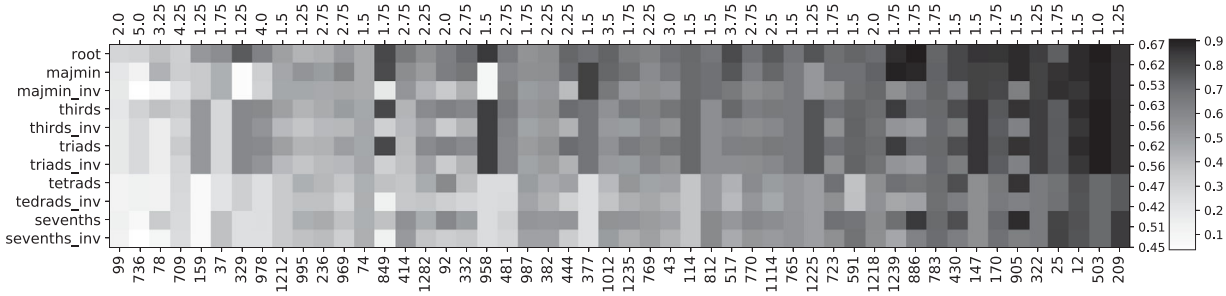

Figure 10. Krippendorff's α inter-rater agreement of all songs in the dataset. The chequerboard-like pattern reveals that for each level of granularity, the level of agreement decreases when inversions are taken into account. *Billboard* dataset IDs can be found below the columns; average reported difficulties 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.

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* annotations 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 0.86, 0.86, 0.83, 0.63 and 0.61 for ROOT, MAJMIN, MAJMIN_INV, SEVENTHS and SEVENTHS_INV, respectively.[7] Table 5 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, which suggest that they are overfitting to a single subjective annotation.

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

5.3. 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 labelling 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 0.4 and 0.75 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 standardised 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 6 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 chequerboard-like pattern as in Figure 6, indicating that the inter-annotator agreement for chord-label granularities is lower when inversions are taken into account. In summary, accounting for agreement by random chance using Krippendorff's α, we find a quite low inter-rater agreement across all annotators.

6. Individual differences in annotation ability

The previous sections highlight several areas of variance among the annotators: annotation time, chord vocabulary, how difficulty is perceived and their agreement with the *Billboard* reference annotation. In order to formalise the potential causes of this variance, we examine the correlation of these annotator behaviour measures (reported annotation time, reported annotation difficulty and number of unique chords used) with the annotators' agreement with the *Billboard* reference annotation. Here, we use the *Billboard* reference annotation as a ground truth in explaining how the annotators' perceived difficulty in transcribing a song relates to their chord-label agreement for that song. In particular we are interested to investigate if each annotator is indeed unique, and whether they have a particular pattern of sensitivity to chord labels that influence how difficult they perceive a song to be.

Put more formally, we are interested in describing the variability among the annotators in terms of a potentially lower number of unobserved variables called factors. To do so, we first perform an exploratory factor analysis to uncover the underlying structure of the measurements from our annotators. This analysis suggests a model that describes an annotator's performance as depending on four difficulty factors that have high explanatory power. Using the factor model as a basis for a confirmatory factor analysis, we verify the plausibility of the exploratory model and test for the presence of higher level factors, the effects of song length and learning, and whether annotators differ significantly from each of the factors – or in other words, what exactly causes annotators' transcriptions to vary.

In the following sections, we report on the findings of an exploratory and confirmatory factor analysis, of which the detailed statistical intricacies can be found in Koops (2019).

6.1. Exploratory factor analysis

To investigate the number of factors (or dimensionality) that could explain our set of measures, we begin with an exploratory factor analysis. Both parallel analysis (Humphreys & Montanelli, 1975) and Velicer's MAP criterion (Velicer, 1976), two common techniques for choosing the dimensionality suggest that four factors are sufficient. Table 6 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 maximise 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 (*Triad Difficulty*), Factor 2 represents additional difficulty arising from sevenths (*Sevenths Difficulty*) and Factor 4 represents additional difficulty arising from inversions (*Inversions Difficulty*). Factor 3 (*Annotation Difficulty*) collects all three of the annotator-dependent difficulty measures, suggesting that there is indeed a distinct difficulty 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 intercorrelated 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 suggests that annotator's performance depends on a baseline triad-level difficulty, additional difficulty arising from sevenths or inversions, and a further chord difficulty 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 Difficulty*), which has considerable explanatory value; the extra factor in a

Table 6. Exploratory factor analysis of annotation difficulty indicators (Oblimin Rotation).

Indicator	Factor 1 *Triad Difficulty*	Factor 2 *Sevenths Difficulty*	Factor 3 *Annotation Difficulty*	Factor 4 *Inversions Difficulty*	h^2
			Loadings		
MIREX vocabulary					
THIRDS	0.96	0.02	0.01	−0.01	0.96
MAJMIN	0.95	0.05	−0.03	0.03	0.97
TRIADS	0.92	0.02	0.09	0.01	0.96
ROOT	0.92	0.03	0.01	0.01	0.91
MIREX	0.94	−0.02	0.02	0.00	0.88
THIRDS_INV	0.46	0.15	0.11	0.55	0.97
MAJMIN_INV	0.47	0.15	0.05	0.58	0.99
TRIADSlowbar;INV	0.48	0.12	0.14	0.53	0.98
SEVENTHS	0.18	0.92	−0.04	0.22	0.98
TETRADS	0.19	0.89	0.05	−0.24	0.98
SEVENTHS_INV	−0.10	0.97	0.00	0.23	0.99
TETRADS_INV	−0.08	0.94	0.08	0.20	0.98
Difficulty rating	−0.04	0.00	0.94	−0.06	0.83
Annotation time	0.07	−0.03	0.88	0.00	0.83
Number of unique chords	−0.07	0.02	0.80	0.01	0.60
			Intercorrelations (proportion variance explained on diagonal)		
Factor 1	0.39				
Factor 2	0.67	0.26			
Factor 3	0.49	0.36	0.17		
Factor 4	0.39	0.29	0.24	0.10	

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.
Estimates from the R **psych** package, version 1.7.8, using the principal-factor method (Revelle, 2018).

five-factor model, in contrast, has no obvious interpretation and no items with loadings of greater magnitude than the four-factor model.

6.2. Confirmatory factor analysis

The exploratory factor analysis suggests a basic underlying model for how annotators' perceived difficulty in transcribing a song relates to their agreement with the *Billboard* reference annotation 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 aspects where annotator subjectivity is at work. We thus use the four-factor model as a basis for a *confirmatory factor analysis*, where we can verify the plausibility of the exploratory model and test for (1) the presence of the *General Difficulty* factor, (2) the effects of song length and learning, and (3) individual differences among annotators.

Our first step in the confirmatory analysis is to define the factors more rigorously. Given the loading patterns and high intercorrelations in the exploratory model, we allow Factor 1 (*Triad Difficulty*) to load on all 12 of the MIREX WCSR measures. All other loadings for this factor are constrained to zero. We allow Factor 2 (*Sevenths Difficulty*) to load only the four MIREX vocabularies involving sevenths and Factor 4 (*Inversions Difficulty*) to load only on the five vocabularies involving inversions, again constraining all other possible loadings on these factors to zero. Because Factor 1 (*Triad Difficulty*) loads on every MIREX measure, we know that Factors 2 and 4 are specifically measuring the extra difficulty caused by sevenths or inversions, independent of the difficulty of identifying any other harmonic aspects. We allow Factor 3 (*Annotation Difficulty*) to load only on the three annotator-dependent measures, reported difficulty, reported annotation time and number of unique chords. To ensure that the model remains identified given the overlapping factors, we enforce independence (zero covariance) between Factor 1 (*Triad Difficulty*) and Factor 2 (*Sevenths Difficulty*), and also between Factor 1 (*Triad Difficulty*) and Factor 4 (*Inversions Difficulty*); we allow all other pairs of factors to covary.

We fit this first-order model to each annotator individually and find that 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 (so many of Annotator 2's ratings are 1) that it is impossible to estimate an underlying normal variable reliably. Despite the overall instability of the fit for Annotator 2, all loadings in this first-order model are large, statistically significant at the 5% level ($p < 0.05$) and of comparable magnitude for every individual annotator, meaning that all four factors are important for all four

annotators. We accepted the first-order model, and for further analysis, we assumed that all annotators shared this four-factor model structure.

6.3. Underlying General Difficulty factor

The four factors (*Triad*, *Sevenths*, *Annotation* and *Inversions Difficulty*) are highly inter-correlated, which suggests that there may be an underlying *General Difficulty* factor that is responsible for this correlation, i.e. a second-order model (see Figure 11). A second-order model does indeed fit acceptably well and the degradation in fit from the first-order model is not statistically significant ($p = 0.90$). Looking in detail at the model parameters, however, we notice that the loadings on *Sevenths Difficulty* are small and not statistically significant for any annotator. As such, we also test an even more parsimonious model wherein the *General Difficulty* factor is not allowed to load on *Sevenths Difficulty* (i.e. we fix the loading to zero). This second-order model without a connection between *General Difficulty* and *Sevenths Difficulty* also fits acceptably well and shows no significant degradation from the model where the loading between *General Difficulty* and *Sevenths Difficulty* is free ($p = 0.44$). We accept the presence of a *General Difficulty* factor and use the model without a connection to *Sevenths Difficulty* as our basis for further testing.

Given the *General Difficulty* factor, we then examine whether song length or learning affects *General Difficulty*. As a proxy for learning, we simply use the tranche in which annotators received each song (first, second or third). We first test a model with both of these covariates as exogenous predictors of *General Difficulty* and find that while song length has a significant effect for all annotators, the tranche does not have a significant effect for any annotator. Removing tranche shows no significant degradation in model fit ($p = 0.38$), but removing song length degrades model fit substantially ($p = 0.01$). We choose the model with only song length as a predictor of *General Difficulty*. Figure 11 depicts this model structure.

6.4. Difference in difficulty factors across annotators

In order to test whether the latent difficulty factors (*Triad*, *Sevenths*, *Annotation*, *Inversions* and *General Difficulty*) differ across annotators, we follow the procedure recommended by Brown (2015). After testing for measurement invariance, differences in factor variances and differences

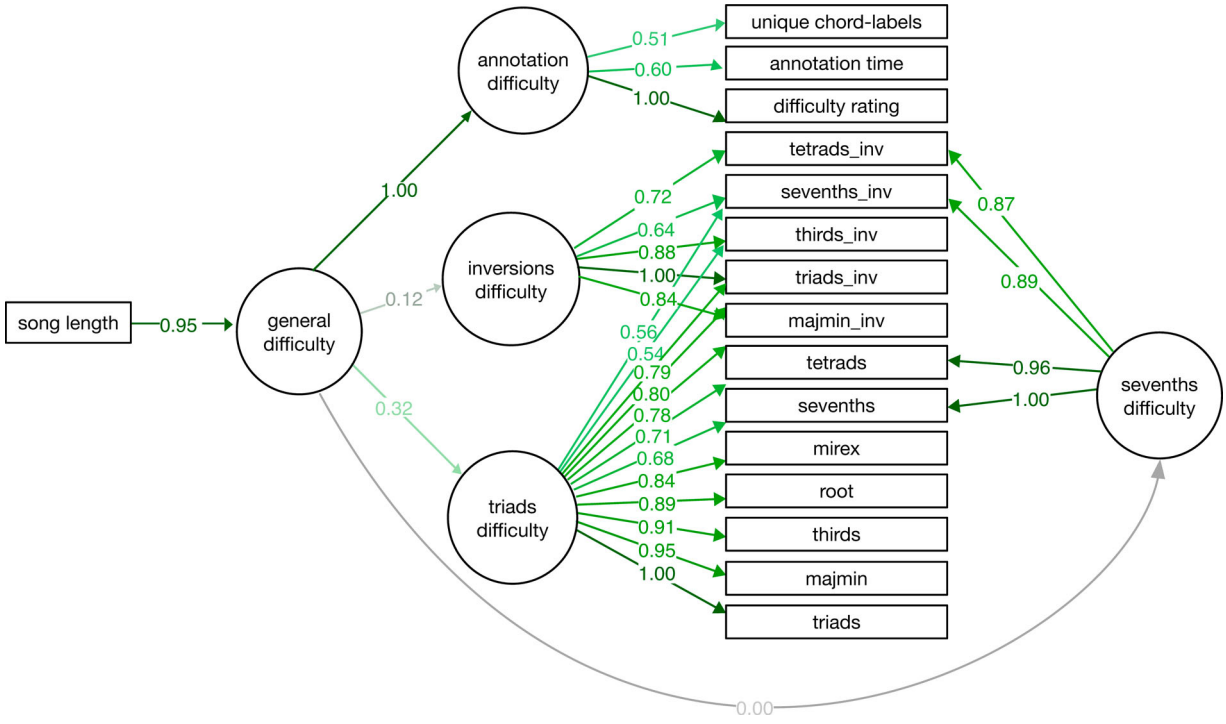

Figure 11. Second-order factor model for indicators of annotation difficulty. The exploratory factor analysis suggests that annotator's performance depends on a baseline triad-level difficulty (*Triad Difficulty*), additional difficulty arising from sevenths (*Sevenths Difficulty*) or inversions (*Inversions Difficulty*), and a further chord difficulty factor (*Annotation Difficulty*). A second-order *General Difficulty* factor predicts three of the four first-order factors. Loadings are unstandardised and common to all annotators. Intercepts (which were common across annotators) and residual variances (which were not) are omitted for clarity. The largest loading on each factor is set to 1.0 in order to fix their scales.

in factor means, we reject the hypothesis of equal factor variance across annotators ($p = .09$, but with substantial degradation in goodness-of-fit statistics), but retain the hypothesis that the factor means are the same ($p = .88$).

In summary, the *General Difficulty* factor can explain both annotators' perceived difficulty and their agreement with the *Billboard* reference annotation; more difficult songs exhibit less agreement, and our chosen annotator-dependent measures are consistent with the common external measures of WCSR. While we find no evidence of a learning effect from annotation experience, we find song length has 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 can be explained by four lower level factors: *Triad Difficulty*, *Sevenths Difficulty*, *Inversions Difficulty* and *Annotation Difficulty*. On average, all annotators found the songs equally difficult with respect to these factors, but the variance differs. Finally, even after taking into account the difficulty factors, some annotators are 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. Low variance suggests a lack of sensitivity to a particular type of difficulty; 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 transcription ability *per se*, but in the relative importance of triads, sevenths, inversions and other musical factors for each annotator.

Looking at a Schmid–Leiman factorisation of the final model, which separates the loading for each measure into a portion arising exclusively from *General Difficulty* and the portions arising from the residual variance of the other factors (Schmid & Leiman, 1957), several characteristics of the annotators 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 difficulty ratings, annotation time and number of unique chords are also influenced to a moderate degree by specific *Annotation Difficulty* above and beyond the high-level *General 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 analyse this finding more deeply. At the lower level, we see that Annotator 2 is highly sensitive to *Sevenths Difficulty*, and that Annotator 4 is quite sensitive to *Inversions Difficulty*. Consistent with the earlier findings, the performance of Annotator 2 is more idiosyncratic with respect to the model than the other three annotators.

In short, each annotator in our sample is unique, exhibiting a distinct pattern of sensitivity to particular types of difficulty in the song sample. Inevitably, these differing sensitivities lead to differing transcriptions.

7. Conclusion and discussion

In this paper, we presented a new harmonic annotator subjectivity dataset of 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. The intersection of the four annotators' chord vocabularies was less than 20 percent of the union of the four vocabularies.

Furthermore, in a pairwise analysis of the annotations using the commonly used MIREX evaluation measures, we find that annotators agree on average on only 76 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. Hence, our results are in line with a comparable experiment by Ni et al. (2013). Using annotations from formally trained musicians, Ni et al. (2013) reported annotator subjectivity of around 10% among the annotators when compared to their consensus. Comparable but slightly 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 average agree just as much with each other as with the *Billboard* annotations. This suggests that a consensus annotation from the *Billboard* dataset can be regarded as equally valid as the annotations 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 personalised 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 personalisation. In fact, a first approach to such a system has already been introduced by Koops, de Haas, Bransen, and Volk (2017), showing that chord labels can be tuned to an annotator's specific vocabulary from a representation shared by multiple annotators.

It should be noted that the chord-label agreement measures used in this study are based solely on pitch-class level comparisons. These agreement measures are commonly used in MIR computational harmony research, for example to evaluate the performance of automatic chord estimation systems. However, these measures are agnostic towards the functional properties of a chord in its larger tonal context. It would be interesting to investigate the agreement of the annotators in the CASD on a more functional level, for example using Riemannian analysis. This could reveal that although annotators disagree on pitch-class content of the chord labels, they might agree on the function that chord has in the context of the key of the song.

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 to 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.

Acknowledgments

The authors would like to thank Matt McVicar, Arthur Flexer, Alan Marsden and anonymous reviewers for their feedback on an earlier draft of this paper.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Hendrik Vincent Koops http://orcid.org/0000-0002-6980-7027

References

AKA. (2001). *George Harrison Yahoo! Chat Transcript - 02/15/2001*.

Balke, S., Driedger, J., Abeßer, J., Dittmar, C., & Müller, M. (2016). Towards evaluating multiple predominant melody annotations in Jazz recordings. In *Proceedings of the 17th international society for music information retrieval conference* (pp. 246–252). New York.

Bennett, J. (2001). *Guitar on Tap!: Chords, scales, tunings, riffs*. Sydney: Wise.

Bosch, J. J., & Gómez, E. (2014). Melody extraction in symphonic classical music: A comparative study of mutual agreement between humans and algorithms. In *Proceedings of the 9th conference on interdisciplinary musicology*, Berlin.

Bradlow, E. T., & Fader, P. S. (2001). A Bayesian lifetime model for the 'hot 100' billboard songs. *Journal of the American Statistical Association*, *96*(454), 368–381.

Brown, T. A. (2015). *Confirmatory factory analysis for applied research* (2nd edn). New York: Guilford.

Burgoyne, J. A., Wild, J., & Fujinaga, I. (2011). An expert ground-truth set for audio chord recognition and music analysis. In *Proceedings of the 12th international society for music information retrieval conference* (pp. 633–638). Miami.

Burgoyne, J. A., Wild, J., & Fujinaga, I. (2013). Compositional data analysis of harmonic structures in popular music. In *Lecture notes in computer science: Lecture notes in artificial intelligence: 0302-9743* (vol. 7937 LNAI, pp. 52–63). Heidelberg: Springer.

Chuan, C.-H., & Chew, E. (2007). A hybrid system for automatic generation of style-specific accompaniment. In *Proceedings of the 4th international joint workshop on computational creativity* (pp. 57–64). London.

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

De Clercq, T., & Temperley, D. (2011). A corpus analysis of rock harmony. *Popular Music*, *30*(1), 47–70.

Doll, C. (2017). *Hearing Harmony: Toward a tonal theory for the rock era*. Ann Arbor: University of Michigan Press.

Fleiss, J. L. (1975). Measuring agreement between two judges on the presence or absence of a trait. *Biometrics*, *31*(3), 651.

Flexer, A. (2014). On inter-rater agreement in audio music similarity. In *Proceedings of the 15th international society for music information retrieval conference* (pp. 245–250). Taipei.

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

Fujita, T., Hagino, Y., Kubo, H., & Sato, G. (1993). *The Beatles: Complete scores*. Wisconsin: Hal Leonard Publishing.

Gauvin, H. L. (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.

Gould, J. (2014). *Can't buy me love: The Beatles, Britain, and America*. UK: Hachette.

Gwet, K. L. (2010). *Handbook of Inter-Rater Reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC.

Haas, W. B. de, Volk, A., & Wiering, F. (2013). Structural segmentation of music based on repeated harmonies. In *IEEE international symposium on multimedia* (pp. 255–258).

Harte, C., Sandler, M., Abdallah, S. A., & Gómez, E. (2005). Symbolic representation of musical chords: A proposed syntax for text annotations. In *Proceedings of the 6th international society for music information retrieval conference* (pp. 66–71), London.

Hickey, A. (2010). *The Beatles In Mono*. Morrisville: lulu.com.

Hu, X., & 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. J., & Bello, J. P. (2015). Four timely insights on automatic chord estimation. In *Proceedings of the 16th international society for music information retrieval conference* (pp. 673–679). Spain: Málaga.

Humphrey, E. J., Salamon, J., Nieto, O., Forsyth, J., Bittner, R. M., & Bello, J. P. (2014). JAMS: A JSON annotated music specification for reproducible MIR research. In *Proceedings of the 15th international society for music information retrieval conference* (pp. 591–596).

Humphreys, L. G., & Montanelli, R. G. (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., & Wu, Y. (2017). Extended abstract for MIREX 2017 submission: Chord recognition using random forest model. In *MIREX evaluation results*, Suzhou.

Jones, M. C., Downie, J. S., & Ehmann, A. F. (2007). Human similarity judgments: Implications for the design of formal evaluations. In *Proceedings of the 8th international society for music information retrieval conference* (pp. 539–542), Vienna.

Kaliakatsos-Papakostas, M., Cambouropoulos, E., Kühnberger, K.-U., Kutz, O., & Smaill, A. (2014). Concept invention and music: Creating novel harmonies via conceptual blending. In *Proceedings of the 9th conference on interdisciplinary musicology (cim2014)*, Berlin.

Koops, H. V. (2019). *Computational modelling of variance in musical harmony* (PhD dissertation). Utrecht University, Utrecht, the Netherlands.

Koops, H. V., de Haas, W. B., Bountouridis, D., & Volk, A. (2016). Integration and quality assessment of heterogeneous chord sequences using data fusion. In *Proceedings of the 17th international society for music information retrieval conference* (pp. 178–184). New York City.

Koops, H. V., de Haas, W. B., Bransen, J., & 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* (pp. 19–25). Anchorage.

Korzeniowski, F., Böck, S., Krebs, F., & Widmer, G. (2017). MIREX submissions for chord recognition and key estimation 2017. In *MIREX evaluation results*, Suzhou.

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: Some common misconceptions and recommendations. *Human Communication Research*, *30*(3), 411–433.

Lippens, S., Martens, J.-P., & De Mulder, T. (2004). A comparison of human and automatic musical genre classification. In *2004 IEEE international conference on acoustics, speech, and signal processing* (vol. 4, pp. iv–233–iv–236), Montreal.

Macrae, R., & Dixon, S. (2011). Guitar tab mining, analysis and ranking. In *Proceedings of the 12th international society for music information retrieval conference* (pp. 453–458).

Martin, G., & Hornsby, J. (1994). *All you need is ears: The inside personal story of the genius who created The Beatles*. London: Macmillan.

Mauch, M., Cannam, C., Davies, M., Dixon, S., Harte, C., Kolozali, S., ... Sandler, M. (2009). OMRAS2 Metadata Project 2009. In *Late breaking session of the 10th international society of music information retrieval*.

Mauch, M., & Dixon, S. (2010). Approximate note transcription for the improved identification of difficult chords. In *Proceedings of the 11th international society for music information retrieval conference* (pp. 135–140). Utrecht.

Mauch, M., MacCallum, R. M., Levy, M., & Leroi, A. M. (2015). The evolution of popular music: USA 1960–2010. *Royal Society Open Science*, *2*, 150081–150081.

McKinney, M. F., Moelants, D., Davies, M. E. P., & Klapuri, A. (2007). Evaluation of audio beat tracking and music tempo extraction algorithms. *Journal of New Music Research*, *36*(1), 1–16.

McVicar, M., Sach, B., Mesnage, C., Lijffijt, J., Spyropoulou, E., & De Bie, T. (2016). SuMoTED: An intuitive edit distance between rooted unordered uniquely-labelled trees. *Pattern Recognition Letters*, *79*, 52–59.

McVicar, M., Santos-Rodríguez, R., Ni, Y., & 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*, *22*(2), 556–575.

Moore, A. F. (2012). *Song means: Analysing and interpreting recorded popular song*. Abingdon: Routledge.

Nattiez, J.-J. (1990). *Music and discourse: Toward a semiology of music*. Princeton: Princeton University Press.

Ni, Y., McVicar, M., Santos-Rodriguez, R., & De Bie, T. (2013). Understanding effects of subjectivity in measuring chord estimation accuracy. *IEEE Transactions on Audio, Speech, and Language Processing*, *21*(12), 2607–2615.

Nieto, O., Farbood, M. M., Jehan, T., & 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* (pp. 265–270).

Paulus, J., & 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., & Peeters, G. (2013). Evaluating automatically estimated chord sequences. In *2013 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 749–753).

Pedler, D. (2010). *The songwriting secrets of the beatles*. London: Omnibus Press.

Raffel, C., Mcfee, B., Humphrey, E. J., Salamon, J., Nieto, O., Liang, D., ... Raffel, C. (2014). mir_eval: A transparent implementation of common MIR metrics. In *Proceedings of the 15th international society for music information retrieval conference* (pp. 367–372).

Revelle, W. (2018). *psych: Procedures for psychological, psychometric, and personality research*. Evanston: Northwestern University.

Roessner, J. (2009). *Creative guitar writing and playing rock songs with originality*. Fenton: Mel Bay Publications.

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

Salamon, J., & Urbano, J. (2012). Current challenges in the evaluation of predominant melody extraction algorithms. In *Proceedings of the 13th international society for music information retrieval conference* (pp. 5–10). Porto.

Schedl, M., Eghbal-Zadeh, H., Gomez, E., & Tkalčič, M. (2016). An analysis of agreement in classical music perception and its relationship to listener characteristics. In *Proceedings of the 17th international society for music information retrieval conference* (pp. 578–583).

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

Seyerlehner, K., Widmer, G., & Knees, P. (2011). A comparison of human, automatic and collaborative music genre classification and user centric evaluation of genre classification systems. In M. Detyniecki, P. Knees, A. Nürnberger, M. Schedl, and S. Stober (Eds.), *Adaptive multimedia retrieval. Context, exploration, and fusion* (pp. 118–131). Berlin, Heidelberg: Springer.

Smith, J., Burgoyne, J. A., Fujinaga, I., De Roure, D., & Downie, J. S. (2011). Design and creation of a large-scale database of structural annotations. In *Proceedings of the 12th international society for music information retrieval conference* (pp. 555–560). Miami.

Spitz, B. (2005). *The Beatles: The biography*. Boston: Little, Brown.

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

Van Balen, J., Burgoyne, J. A., Bountouridis, D., Müllensiefen, D., & Veltkamp, R. C. (2015). Corpus analysis tools for computational hook discovery. In *Proceedings of the 16th international society for music information retrieval conference* (pp. 227–233). Malaga.

Van Balen, J., Burgoyne, J. A., Wiering, F., & Veltkamp, R. C. (2013). An analysis of chorus features in popular song. In *Proceedings of the 14th international society for music information retrieval conference* (pp. 107–112). Curitiba.

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

Womack, K. (2017). *Maximum volume: The life of beatles producer George Martin, the early years, 1926–1966*. Chicago: Chicago Review Press.

Wu, Y., Feng, X., & Li, W. (2017). MIREX 2017 Submission: Automatic audio chord recognition with miditrained deep feature and BLSTM-CRF sequence decoding model. In *MIREX evaluation results*, Suzhou.