FOUR TIMELY INSIGHTS ON AUTOMATIC CHORD ESTIMATION

First author
Affiliation1
author1@ismir.net

Second author
Retain these fake authors in submission to preserve the formatting

Third author
Affiliation3
author3@ismir.net

ABSTRACT

Automatic chord estimation (ACE) is now a hallmark research topic in content-based music informatics, but like many other tasks, system performance appears to be converging to yet another glass ceiling. Recently, two different large-vocabulary ACE systems were developed in the hopes that complex, data-driven models might significantly advance the state of the art. While arguably achieving some of the highest results to date, both approaches plateau at the same level, well short of having solved the problem. Therefore, this work explores the behavior of these two systems as a means of understanding obstacles and limitations in chord estimation, arriving at four critical observations: one, music recordings that invalidate tacit assumptions about harmony and tonality result in erroneous and even misleading performance; two, standard lexicons and comparison methods struggle to reflect the natural relationships between chords; three, conventional approaches conflate the competing goals of recognition and transcription to some undefined degree; and four, the perception of chords in real music can be highly subjective, making the very notion of "ground truth" annotations tenuous. Synthesizing these observations, this paper offers possible remedies going forward, and concludes with some perspectives on the future of ACE research.

1. INTRODUCTION

Among the various subtopics in content-based music informatics, automatic chord estimation (ACE) has matured into a classic MIR challenge, receiving healthy attention from the research community for the better part of two decades. Complementing our natural sense of academic intrigue, the general music learning public places a high demand on chord-based representations of popular music, as evidenced by large online communities surrounding websites like e-chords [1] or Ultimate Guitar [2]. Given the prerequisite skill necessary to manually identify chords from recorded audio, there is considerable motivation to develop automated systems capable of reliably performing this task.

The goal of ACE research is —or, at least, has been— to develop systems that produce "good" time-aligned sequence of chords from a given music signal. Supplemented by efforts in data curation [3], syntax standardization [7], and evaluation [11], the bulk of chord estimation research has concentrated on building better systems, mostly converging to a common architecture [4]: first, harmonic features, referred to as pitch class profiles (PCP) or *chroma*, are extracted from short-time observations of the audio signal [6]; these features may then be processed by any number of means, referred to in the literature as *pre-filtering*; next, *pattern matching* is performed independently over observations to measure the similarity between the signal and a set of pre-defined chord classes, yielding a time-varying likelihood; and finally, *post-filtering* is applied to this chord class posterior, resulting in a sequence of chord labels over time.

However, despite continued efforts to develop better features [9], more powerful classifiers [8], or advanced post-filtering methods [2], performance appears to be tapering off, as evidenced by recent years' results at MIReX [3]. Acknowledging this situation begs an obvious question: what is contributing to this trend, and what might be done about it? The remainder of this paper is an effort to address this question: first, Section 2 details the research methodology used here, covering the choice of automatic systems, objective evaluation, and the datasets considered; Section 3 introduces a large-vocabulary chord estimation experiment and the corresponding high-level results; then, to develop a deeper understanding of how these systems behave and the kinds of errors that they make, Section 4 identifies four critical insights, indicating that performance ceilings may be as much a function of traditional ACE methodology as the automatic systems being considered; finally, Section 5 offers perspectives to help frame future work.

2. RESEARCH METHODOLOGY

2.1 Automatic Systems

Given its long history, there are ample potential automatic chord estimation systems that could be considered in this inquiry. Here, though, we choose to focus our investigation on two recent, data-driven, large vocabulary systems

[1] http://www.e-chords.com
[2] http://www.ultimate-guitar.com
[3] http://www.music-ir.org/mirex/wiki/MIREX_HOME

for which we are able to obtain software implementations, providing control over training and choice of chord vocabulary. Additionally, these system architectures are quite different and should, as a result, yield different machine perspectives, a strategy that has proven useful in the analysis of beat tracking systems [13].

2.1.1 K-stream GMM-HMM with Multiband Chroma

The first system considered is a modern, high-performing GMM/HMM chord estimation system [4], referred to here as "kHMM." A multiband chroma representation is computed from beat-synchronous audio analysis, producing four parallel feature representations. Each is modeled by a separate multivariate Gaussian Mixture Model (GMM), whereby all chroma vectors and chord labels are rotated to a `C` root. During inference, four separate observation likelihoods over all chord classes are obtained by circularly rotating the feature vector the GMM, thereby empowering the model with transposition invariance. These four chord class posteriors are then decoded jointly using a k-stream HMM, resulting in a single beat-aligned chord sequence.

2.1.2 Deep Convolutional Neural Network

Acknowledging both the limited representational power of GMMs and the recent widespread success of deep learning methods, a deep convolutional network is also considered [1], referred to as "DNN." Time-frequency patches of local contrast normalized constant-Q spectra, on the order of one second, are transformed by a four-layer convolutional network. Finding inspiration in the root-invariance strategy of GMM training, explicit weight-tying is achieved at the classifier across roots such that all qualities develop the same internal representations, allowing the model to generalize to chords unseen during. Following the lead of deep network research in automatic speech recognition, likelihood scaling is performed after training to control class bias resulting from the severe imbalance in the distribution of chords. Finally, chord posteriors are decoded via the basic Viterbi algorithm.

2.2 Evaluation

Expressed formally, the modern approach to scoring an ACE system is a weighted measure of chord-symbol recall, R_W, between a reference, $\mathcal{R}$, and estimated, $\mathcal{E}$, chord sequence as a continuous integral over time, summed over a discrete collection of N annotation pairs:

$$R_W = \frac{1}{S} \sum_{n=0}^{N-1} \int_{t=0}^{T_n} C(\mathcal{R}_n(t), \mathcal{E}_n(t))\, dt \qquad (1)$$

Here, C is a chord *comparison* function, bounded on $[0, 1]$, t is time, n the index of the track in a collection, T_n the duration of the n^{th} track. This total is normalized by the *support*, S, corresponding to the cumulative amount of time over which the comparison rule is defined for $\mathcal{R}$, given by the indicator function in a similar integral:

$$S = \sum_{n=0}^{N-1} \int_{t=0}^{T_n} \mathbb{1}_{\mathcal{R}_n(t)}\, dt \qquad (2)$$

Defining the normalization term S separately is useful when comparing chord names, as it relaxes the assumption that the comparison function is defined everywhere. Furthermore, setting the comparison function as a free variable allows for flexible evaluation of a system's outputs, and thus the focus on vocabulary can largely focus on the choice of comparison function, C. The work presented here leverages `mir_eval`, an open source evaluation toolbox providing a set of seven chord comparison functions, characterizing different relationships between chords [12].

2.3 Reference Annotations

2.3.1 Ground Truth Data

The first major effort to curate reference chord annotations, now part of the larger Isophonics [4] dataset, covers the entire 180-song discography of *The Beatles*, as well as 20 songs from *Queen*, 14 from Carole King, and 18 from *Zweieck*; due to content access, only the 200 songs from *The Beatles* and *Queen* are used here. Two other large chord annotation datasets were publicly released in 2011, offering a more diverse musical palette. The McGill *Billboard* dataset consists of over 1000 annotations, of which more than 700 have been made public. This project employed a rigorous sampling and annotation process, selecting songs from Billboard magazine's "Hot 100" charts spanning more than three decades. The other, provided by the Music and Audio Research Lab (MARL) at NYU [5], consists of 295 chord annotations performed by undergraduate music students; 195 tracks are drawn from the USPop dataset [6], and 100 from the RWC-Pop collection [7], in the hopes that leveraging common MIR datasets might facilitate access within the community. In all three cases, chord annotations are provided as "ground truth," on the premise that the annotations represent the gold standard.

2.3.2 The Rock Corpus

Importantly, the reference chord annotations discussed previously offer a singular perspective, either as the output of one person or the result of a review process. The *Rock Corpus*, alternatively, is a set of 200 popular rock tracks with time-aligned chord and melody transcriptions performed by two expert musicians [5]: one, a pianist, and the other, a guitarist, referred to as DT and TdC, respectively. This collection of chord transcriptions has seen little use in the ACE literature, as its initial release lacked timing data for the transcriptions. A subsequent release resolved this issue, however, and doubled the size of the collection. While previous efforts have sought to better understand the role of

[4] `http://isophonics.net/content/reference-annotations`

[5] `https://github.com/tmc323/Chord-Annotations`

[6] `http://labrosa.ee.columbia.edu/projects/musicsim/uspop2002.html`

[7] `https://staff.aist.go.jp/m.goto/RWC-MDB/rwc-mdb-p.html`

	Ref–DNN	Ref–kHMM	kHMM–DNN
root	0.789	0.808	0.840
thirds	0.757	0.775	0.815
majmin	0.759	0.776	0.798
mirex	0.769	0.783	0.806
triads	0.705	0.721	0.783
sevenths	0.620	0.645	0.691
tetrads	0.567	0.588	0.678
v157	0.649	0.659	0.678

Table 1. Weighted recall across comparison rules between the ground truth references and both models, respectively, as well as against each other.

	DT–TdC	(DT|TdC)–DNN	(DT|TdC)–kHMM
root	0.932	0.792	0.835
thirds	0.903	0.750	0.785
majmin	0.905	0.723	0.766
mirex	0.902	0.737	0.776
triads	0.898	0.719	0.760
sevenths	0.842	0.542	0.595
tetrads	0.835	0.540	0.590
v157	0.838	0.539	0.590

Table 2. Weighted recall across comparison rules for the two human annotators, and the better match of each against the two automatic systems.

subjectivity in chord annotations [10], this dataset provides an opportunity to explore the behavior of ACE systems as a function of multiple reference transcriptions at a larger scale.

3. LARGE-VOCABULARY CHORD ESTIMATION

Here we investigate large-vocabulary chord estimation as a basis for experimentation. First and foremost, it presents a particularly challenging problem, and therefore offers a good deal of potential for subsequent analysis. Large chord vocabularies also avoid the inherent noise introduced by approximately mapping chords into the classic major-minor formulation, e.g. `A:sus2`→`A:maj` or `C:dim7`→`C:min`. Additionally, the large amount of available data should be sufficient for learning a large number of chord classes.

Before proceeding, the ground truth collections are merged for training and evaluation, totaling 1235 tracks. A total of 18 redundant songs are identified via the EchoNest Analyze API [8] and removed to avoid potential data contamination during cross validation. All but one is dropped for each collision, preferring content from Isophonics, Billboard, and MARL, respectively, resulting in a final count of 1217 unique tracks.

To ensure a fair comparison between algorithms, the ground truth data is partitioned into five distinct splits. Training is repeated five times for both systems addressed in Section 2.1 for cross validation, such that each split is used as a holdout test set once. Both models adopt the same chord vocabulary, comprised of the thirteen most frequent chord qualities in all twelve pitch classes, as well as a no-chord class, for a total of 157 chord classes, consistent with previous efforts [4]. Chords outside this strict vocabulary are ignored during training, rather than mapped to their nearest class approximation. The Rock Corpus data is not used for training, and saved exclusively for analysis.

3.1 Experimental Results

Weighted recall is averaged over the five test splits are for all reference chord labels according to the seven `mir_eval` comparison rules, shown in Table 1. At first glance, the overall statistics seem to indicate that the two systems are roughly equivalent, with "kHMM" outperforming "DNN" by a small margin. The automatic systems perform best at root-level recall, and performance drops as the comparison rules encompass more chords. Notably, as shown in the third column, these two systems do in indeed yield quite different estimations. Therefore, it will be valuable to not only investigate where the estimated chord sequences differ from the reference, but also how these estimated sequences differ from each other.

Similarly, weighted recall is also given for both systems over the Rock Corpus in Table 2. It is an open question as to how an estimated annotation might best be compared against more than one human reference. For the purposes of analysis, the best matching reference-estimation pair is chosen at the track level and used to compute the weighted average. Still, performance on the Rock Corpus is lower for both automatic algorithms. This is likely a result of a mismatch in chord vocabulary, as space of chords used in the Rock Corpus is a smaller subset than the 157 estimated by automatic systems. Additionally, it is curious to observe a non-negligible degree of disagreement between the two human perspectives, with more than a 15% discrepancy in the tetrads condition. That said, the human annotators do agree a deal more that is attained by either system, indicating that there is likely room for improvement.

3.2 Track-wise Visualizations

While weighted recall gives a good overall measure of system performance, we are particularly interested in developing a more nuanced understanding of how these systems behave. To this end, system performance is now examined at the track-level, as real music is often highly self-similar and the chords within a song with be strongly related. Errors and other kinds of noteworthy behavior should be well-localized as a result, making it easier to draw conclusions from the data.

Two track-wise scatter plots are given in Figure 1, for the ground truth and Rock Corpus datasets. The former compares the agreement between multiple *estimations*, along x, with the better matching estimation for the given reference, along y, where each quadrant characterizes a different behavior: (I), all annotations agree; (II), one estimation matches the reference better than the other; (III), all annotations disagree; and (IV), the estimations agree more with each other than the reference. Impor-

[8] `http://developer.echonest.com/docs/v4`

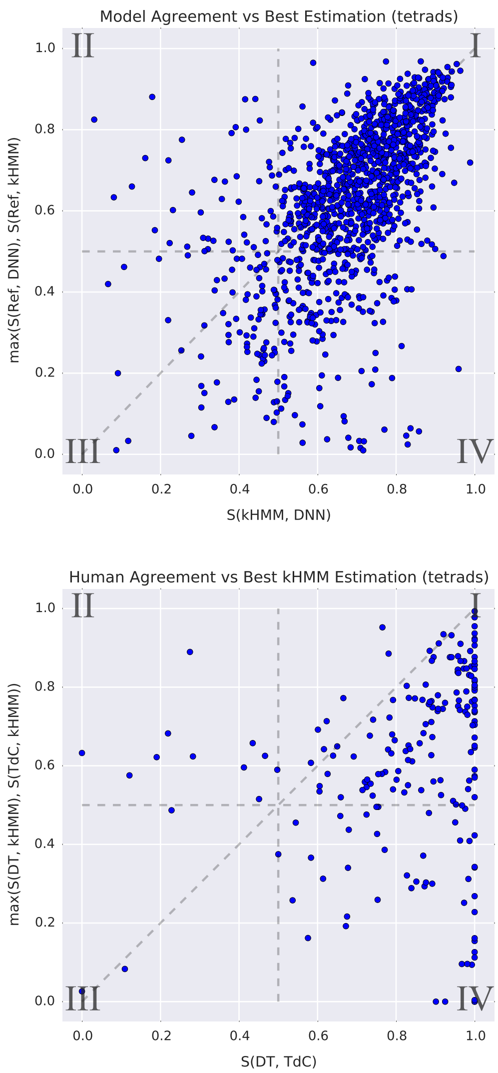

Figure 1. Trackwise recall for the "tetrads" in two conditions: (top) over the ground truth data, illustrating *model* agreement versus the better match between the reference and estimated annotations; (bottom) over the Rock Corpus data, illustrating annotator agreement versus the better match between the two reference and kHMM annotations.

tantly, this track-wise comparison makes it easier to identify datapoints that can help address the original research questions. Tracks for which only one algorithm performs well (II) likely indicate boundary chords. Alternatively, instances where both algorithms produce poor estimations, and yet *neither* agree (III), are curious and warrant further inspection. Finally, tracks that result in similarly incorrect estimations (IV) highlight some kind of greater challenge to automatic systems.

The second plot, conversely, compares the agreement between multiple *references*, along x, with the better matching reference for the given estimation, along y, and analogous characterizations by quadrant: (I), all annotations agree; (II), one reference matches the estimation better than the other; (III), all annotations disagree; and (IV), the references agree more with each other than the estimation. Here, annotator disagreement in the presence of a matching estimation (II) is indicative of subjectivity, while disagreement between all annotations (III) is suspicious and should be explored. Furthermore, tracks with an estimated annotation that fails to match either human perspective (III & IV) likely identify room for improvement.

4. QUALITATIVE ANALYSIS, IN FOUR PARTS

Using this suite of analysis tools described previously, a thorough exploration of the relationship between reference and estimated annotations is conducted, resulting in four significant insights. In the spirit of both reproducibility and open access, a companion IPython notebook is made available online [9] , providing additional visualizations complementary to the following discussion.

4.1 Invalid Harmonic Assumptions

An exploration of quadrant (IV) from Figure 1 reveals that a large source of error stems from musical content or reference chord annotations that violate basic assumptions about how chords are used. One common form of this behavior is due to issues of intonation, where a handful of recordings are not tuned to A440, with some varying by more than a quarter-tone: for example, "Stand By Me" by Jimmy Ruffin, "I'll Tumble 4 Ya" by *The Culture Club*, "Every Breath You Take" by *The Police*, or "Nowhere to Run" by Martha Reeves and *the Vandellas*. Understandably, as a result, the estimated annotations differ by a semitone from the reference, and perform poorly across all comparison rules.

The second observation finds that some tracks in the dataset do not truly make use of, and are thus not well described by, chords. While a few classic songs by *The Beatles* have been known to be of questionable relevance for their instrumentation and lack of standard chords, such as "Revolution 9," "Love You To," or "Within You, Without You", analysis here identifies several other tracks, spanning rap, hip hop, reggae, funk and disco, that behave similarly: for example, "Brass Monkey" by *The Beastie Boys*, "I, Me, & Myself" by *de la Soul*, "Don't Push" by *Sublime*, "Get Up (I Feel Like Being a Sex Machine)" by James Brown, or "I Wanna Take You Higher" by Tina and Ike Turner. This realization encourages the conclusion that chords may not be a valid way to describe all kinds of music, and that using such songs for evaluation may lead to erroneous or misleading results.

4.2 Limitations of Chord Comparisons

The second observation resulting from this analysis is the difficulty faced in the comparison of related chords. By and large, ACE systems are often forced to either map chords to a finite dictionary, or develop embedding rules for equivalence testing [12]. In either case, this quantization process assigns all observations to a one-of-K representation effectively making all errors equivalent. For the

[9] `http://github.com/username/demo.ipynb`

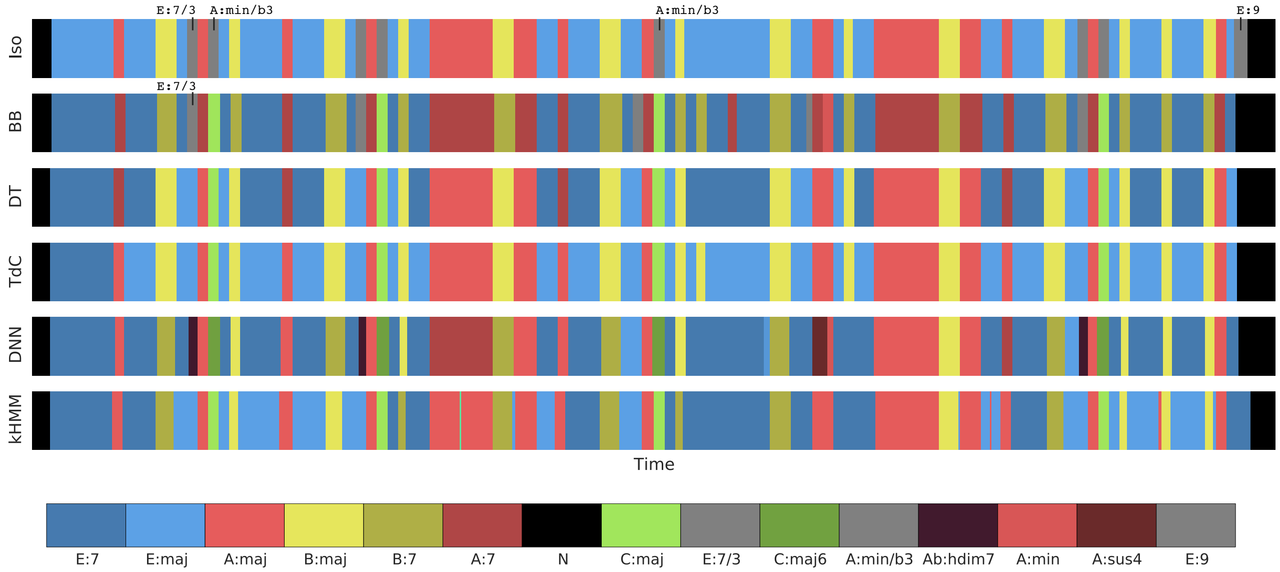

Figure 2. Six perspectives on "I Saw Her Standing There", by *The Beatles*, according to Isophonics (Iso), Billboard (BB), David Temperley (DT), Trevor deClercq (TdC), the Deep Neural Network (DNN), and the k-stream HMM (kHMM).

purposes of stable evaluation, this can have significantly negative consequences.

Chords are naturally related to each other hierarchically, and cannot always be treated as distinct classes. Flat classification problems —i.e. those in which different classes are disjoint— are built on the assumption of mutually exclusive relationships. In other words, assignment to one class precludes the valid assignment to any other class considered. In the space of chords, `C:dim7` and `C:maj` are perhaps mutually exclusive classes, but it is difficult to say the same of `C:maj7` and `C:maj`, as the former *contains* the latter. This conflict is a common source of disagreement between annotators of the Rock Corpus tracks, which are easily identified in or near the quadrant (II) of Figure **??**: for example, "Dancing In The Street" by Martha Reeves & The Vandellas, "All Apologies" by *Nirvana*, or "Papa's Got a Brand New Bag" by James Brown. In each case, the human perspectives each report related tetrads and triads, e.g. `E:7` and `E:maj`, causing low annotator agreement, while the machine estimation alternates between the two trying to represent both. These kinds of errors are not "confusions" in the classic sense, but a limitation of evaluation methods to reliably quantify this behavior, and of the model to represent this naturally structured output.

4.3 Conflicting Problem Definitions

Over the years, the automatic prediction of chord sequences from music audio has taken several names: estimation, recognition, identification, or transcription. The analysis here motivates the notion that this is not merely a matter of semantics, but actually a subtle distinction indicative of two slightly different problems being addressed. Chord *transcription* is an abstract task related to functional analysis, taking into consideration high-level concepts such as long term musical structure, repetition, segmentation or key. Chord *recognition*, on the other hand, is quite literal, and is closely related to polyphonic pitch detection. Both interpretations are easily found in the collection of reference annotations, however, conflating these two tasks to some unknown degree.

Furthermore, the goal in transcription is to assign chord labels to regions, and is closer in principle to segmentation than classic approaches to chord estimation. One illustrative instance, "All Apologies" by *Nirvana*, is identified in quadrant (II) of Figure 1. Here, the human annotators have disagreed on the harmonic spelling of the entire verse, with DT and TdC reporting `C#:maj` and `C#:7`, respectively. On closer inspection, it would appear that both annotators are in some sense correct; the majority of the verse is arguably `C#:maj`, but a cello sustains the flat-7^{th} of this key intermittently. The regions in which this occurs are clearly captured in the estimated annotations, corresponding to its `C#:7` predictions. This proves to be an interesting discrepancy, because one annotator (DT) is using long-term structural information about the song to apply a single chord to the entire verse.

4.4 Ground Truth *vs.* Subjectivity

While the role that subjectivity can play in chord estimation is becoming better understood [10], it is not handled gracefully in current ACE methodology. To this point, there are two examples worth analyzing here, which intersect multiple datasets and provide several expert perspectives.

The first, "I Saw Her Standing There" by *The Beatles*, is given in Figure 2, where the pitch class of the chord's root is mapped to color hue, and the darkness is a function of chord quality, e.g., all `E:*` chords are a shade of blue. No-chords are always black, and chords that do not fit into one of the 157 chord classes are shown in gray.

Ver.	Chord Sequence				Score	Ratings	Views
Billboard	D:maj	A:sus4(b7)	B:min7	G:maj9	—	—	—
MARL	D:maj	D:maj/5	D:maj6/6	D:maj(4)/4	—	—	—
DT	D:maj	A:maj	B:min	G:maj	—	—	—
TdC	D:maj	A:maj	B:min	G:maj	—	—	—
DNN	D:maj	A:sus4	B:min7	G:maj7	—	—	—
kHMM	D:maj	A:sus4	B:min	G:maj	—	—	—
1	D:maj	A:maj	B:min	G:maj	4/5	193	1,985,878
2	D:5	A:sus4	B:min7	G:maj	5/5	11	184,611
3*	D:maj	A:maj	B:min	G:maj	4/5	23	188,152
4*	D:maj	A:maj	B:min	G:maj7	4/5	14	84,825
5*	D:maj	A:maj	B:min	G:maj	5/5	248	338,222
6	D:5	A:5	D:5/B	G:5	5/5	5	16,208

Table 3. Various interpretations of the verse from "With or Without You" by *U2*, comparing the reference annotations and automatic estimations with six interpretations from a popular guitar tablature website; a raised asterisk indicates the transcription is given relative to a capo, and transposed to the actual key here.

Perhaps the most striking observation is the degree of variance between all annotations. Based on the tetrads comparison, no two annotations correspond to greater than a 65% agreement, with the DNN and kHMM scoring 28% and 52% against the ground truth Isophonics reference, shown at the top. Despite this low score, the DNN and kHMM estimations agree with at least one of the four human annotations 89.1% and 92.3% of the song, respectively. The two exceptions occur during the out-of-gamut chords, `E:7/3` and `A:min/b3`, which the DNN calls `Ab:hdim7` and `C:maj6`, respectively. While both estimated chords share three pitches with the Isophonics reference, the other human annotators mark the `A:min/b3` instead as a root position `C:maj`. Given how subjective it might be for human experts to agree on possible inversions, typical evaluation strategies may be placing too much emphasis on the correctness of an estimated chord's root.

A second example to consider in the larger discussion of subjectivity is the verse of "With or Without You" by *U2*. Musically, one finds reasonably ambiguous harmonic content, consisting of a vocal melody, a moving bass line, a guitar riff, and a string pad sustaining a high-pitched `D`. Complementing to the four expert perspectives provided here, an Internet search yields six additional guitar chord transcriptions from the website Ultimate Guitar [10]. All human perspectives and both machine interpretations are consolidated in Table 3, noting both the average and number of ratings, as well as the number of views the public chord annotation has received. Though the view count is not directly indicative of a transcription's accuracy, it does provide a weak signal indicating that a large number of users did *not* rate it negatively.

In considering this particular example, there are a handful of takeaways to note. First, nearly all perspectives are equivalent at the major-minor level, with the exception of the MARL annotation, differing only slightly. That said, the differences between the public guitar chords do not noticeably impact the average ratings. This is an important consideration when building user systems, whereby objective measures are valuable insofar as they correlate with subjective experience. In a related vein, this specific community of musicians seems to show, at least for this song, a strong preference for root position chords. Therefore, subjectivity will likely play a role not only in the process of collecting reference annotations, but in the end-user experience as well.

[10] `http://tabs.ultimate-guitar.com/u/u2/with_or_without_you_crd.htm`, accessed 19 April 2015.

5. CONCLUSIONS AND FUTURE PERSPECTIVES

In this work, qualitative analysis of system performance led to the identification of four key observations affecting current chord estimation methodology: one, not all music content is valid in the context of chord estimation; two, conventional comparison methods struggle to accurately characterize the complex relationships between chords; three, conventional methodology has mixed the somewhat conflicting goals of chord transcription and recognition to an undefined degree; and four, the subjective nature of chord perception may render objective ground truth and evaluation untenable.

Looking to the future of automatic chord estimation, a few opportunities stand out. First, given the parallels to segmentation, it seems chord transcription systems could benefit greatly from recent advances in music structure analysis. Additionally, subjectivity in the reference chord annotations should be embraced, rather than resolved. In the absence of objective truth, chord estimation is arguably more of a tagging problem than classification one, especially where larger vocabularies and inversions are concerned.

Finally, it would be advantageous to incorporate more subjective evaluation in standard chord estimation methodology. If nothing else, users studies can help identify objective measures that align well with subjective experience, and resolve issues of heuristically comparing chords. More practically though, this feedback loop could be used to solicit human input at a larger scale, for additional perspectives on both new as well as previously annotated content.

6. REFERENCES

[1] Anonymous. Currently in the process of submission. In *Proceedings of Anonymous*, pages 1–8, 2015.

[2] Nicolas Boulanger-Lewandowski, Yoshua Bengio, and Pascal Vincent. Audio chord recognition with recurrent neural networks. In *Proceedings of the 14th International Society of Music Information Retrieval Conference (ISMIR)*, pages 335–340, 2013.

[3] John Ashley Burgoyne, Jonathan Wild, and Ichiro Fujinaga. An expert ground truth set for audio chord recognition and music analysis. In *Proceedings of the 11th International Society of Music Information Retrieval Conference (ISMIR)*, pages 633–638, 2011.

[4] Taemin Cho. *Improved techniques for automatic chord recognition from music audio signals*. PhD thesis, New York University, 2014.

[5] Trevor De Clercq and David Temperley. A corpus analysis of rock harmony. *Popular Music*, 30(01):47–70, 2011.

[6] T. Fujishima. Realtime chord recognition of musical sound: a system using common lisp music. In *Proceedings of the International Computer Music Conference (ICMC)*, 1999.

[7] Christopher Harte, Mark B Sandler, Samer A Abdallah, and Emilia Gómez. Symbolic representation of musical chords: A proposed syntax for text annotations. In *Proceedings of the 6th International Society of Music Information Retrieval Conference (ISMIR)*, volume 5, pages 66–71, 2005.

[8] E. J. Humphrey and J. P. Bello. Rethinking automatic chord recognition with convolutional neural networks. In *Proceedings of the International Conference on Machine Learning and Applications*, 2012.

[9] Meinard Müller and Sebastian Ewert. Towards Timbre-Invariant Audio Features for Harmony-Based Music. *IEEE Transactions on Audio, Speech and Language Processing*, 18(3):649–662, March 2010.

[10] Yizhao Ni, Matt McVicar, Raul Santos-Rodriguez, and Tijl De Bie. Understanding effects of subjectivity in measuring chord estimation accuracy. *Transactions on Audio, Speech, and Language Processing*, 21(12):2607–2615, 2013.

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

[12] Colin Raffel, Brian McFee, Eric J Humphrey, Justin Salamon, Oriol Nieto, Dawen Liang, and Daniel PW Ellis. mir_eval: A transparent implementation of common mir metrics. In *Proceedings of the 15th International Society of Music Information Retrieval Conference*, 2014.

[13] José R Zapata, André Holzapfel, Matthew EP Davies, João Lobato Oliveira, and Fabien Gouyon. Assigning a confidence threshold on automatic beat annotation in large datasets. In *Proceedings of the 13th International Society of Music Information Retrieval Conference*, pages 157–162, 2012.